Immunity to Error through Misidentification and Thought Insertion

Max Seeger

Table of Contents

Summary

The question of this book is whether the so called Immunity Thesis holds, i.e. the thesis that introspection-based self-ascriptions of mental states are immune to error through misidentification. Very roughly, Immunity states that in self-ascribing a mental state one cannot be wrong about whose mental state it is. For instance, my belief that I am hungry may be mistaken about the state I am in (I may have mistaken my tiredness for hunger), but it couldn't be mistaken about it being me that is (or seems to be) hungry.

In particular, I ask whether thought insertion and similar pathologies of alienation present counterexamples to Immunity. Thought insertion, anarchic hand syndrome, made impulses, and somatoparaphrenia seem to undermine Immunity since, in these phenomena, subjects misidentify whose thoughts, actions, or sensations they are aware of. However, I argue in a first step that, in their most simple form, the counterexamples miss their target. The judgments involved in alienation are either not introspection-based self-ascriptions, and therefore do not fall within the scope of Immunity, or are not in error through misidentification.

In a second step, I offer a more refined critique from alienation which essentially targets the idea that introspection is identification-free. In response to that critique, I distinguish between two approaches on what it means to be in error through misidentification. The epistemic approach understands error through misidentification as a judgment's being epistemically based on a false identification and consequently equates immunity with epistemic identification-freedom; the ontological approach understands error through misidentification as the divergence of source and target (i.e. the object from which the predication information derives is distinct from the object to which the property is ascribed). I argue that the refined critique undermines Epistemic Immunity, but does not challenge Ontological Immunity.

My defense of Ontological Immunity offers an explanation of Immunity that does without the assumption of identification-freedom, but is based on the idea that introspection implies ownership of a mental state. This explanation dovetails perfectly with the idea that different kinds of self-ascriptions enjoy varying degrees of immunity. Hence, my result is that, depending on what kind of mental state is self-ascribed, introspection-based self-ascriptions are immune to error through misidentification in the sense that they cannot (with varying degrees of modal force) be wrong about it being one's own mental state.

Preface

I want to thank my supervisor Gottfried Vosgerau for providing me with the perfect conditions to write this thesis. Thanks for many stimulating discussions of my work and for always being available providing guidance and support.

This research was generously funded by the Volkswagen Stiftung as part of the research project *Who is thinking?*. Thanks to all other project members, in particular to Martin Voss, and thanks to my colleagues in Düsseldorf, in particular to Tim Seuchter and Arne Weber.

The main ideas of chapter 3 and 6 were developed during a stimulating and fruitful research visit at the Institut Jean Nicod from March to July 2013 and parts of these also appear in my paper "Immunity and Self-Awareness" in *Philosophers' Imprint*. I am very grateful to François Recanati for the invitation and to the DAAD for providing the funding for this visit (grant D/12/45380). For helpful discussion of my ideas and for extensive comments on the material now presented in chapters 3 and 6 I want to thank in particular François Recanati, Alexandre Billon, and Jérémie Lafraire.

Substantial parts of chapter 4, Authorship in Thought Insertion, appear also in my paper "Authorship of thoughts in thought insertion: What is it for a thought to be one's own?" in *Philosophical Psychology*. The ideas of this chapter were presented at the ESPP 2012 and ASSC16. For comments and discussions on these ideas I want to thank my colleagues in Düsseldorf, Patrice Soom, Alex Tillas, Alex auf der Straße, Arne Weber, and Tim Seuchter, as well as Colin Klein, Peter Langland-Hassan, Michael Sollberger, and Ansgar Beckermann.

One of the central ideas of chapter 5 and the main motivation for this chapter are due to a comment by Daniel Stoljar at the ANU philsoc colloquium, for which I am very thankful. Many thanks to Wolfgang Schwarz and Al Hájek for generously inviting me to a research visit at the ANU Canberra from October to December 2013.

In writing this thesis, I have strongly benefitted from an exceptionally good philosophical upbringing at the University of Bielefeld. In particular, I want to thank my teachers Eike von Savigny for his very early support and encouragement, and Ansgar Beckermann whose way of doing philosophy

strongly shaped my thinking and continues to inspire me. Not only did I have excellent teachers, but also excellent fellow students. For inspiring discussions, reading groups, and a wonderful time I want to thank Guido Erhardt, Sebastian Köhler, Romy Jaster, Peter Schulte, and Lars Dänzer.

Without my parents' endless support I would have never gotten this far. They encouraged me in the pursuit of many different goals and gave me the freedom to find my way.

Most of all, I want to thank Romy Jaster for countless discussions in which she helped me develop my ideas, for emotional support, and for her love.

This work is dedicated to Clara and Lili.

1. Introduction

In thought insertion, a pathological phenomenon found in schizophrenia, patients experience thoughts of which they claim not to be the thinker. In anarchic hand syndrome, a condition in which a patient's hand seems to acquire a will of its own, the anarchic hand performs goal-directed movements which the patient is aware of, but claims not to be doing. In a particular case of somatoparaphrenia, a delusion in which subjects disown parts of their bodies, one subject is aware of touches delivered to her disowned hand yet claims that these sensations are not experienced by her, but by her niece. These cases have been alleged to refute the philosophical dictum that one cannot doubt who is the thinker of the thoughts one is aware of, who is the agent of the actions one is aware of, and who is the subject of experiences that one is aware of. In particular, these cases have been claimed to refute the thesis that self-ascriptions of mental states based on introspection are immune to error through misidentification (cf. Campbell 1999b, Marcel 2003, Lane & Liang 2011).

What does it mean to be in error through misidentification? We can intuitively distinguish two kinds of mistakes that could afflict a judgment of the form '*a* is F'. One error pertains to the object, *a*, the other error pertains to the property, F. Consider the following case (FORREST): Suppose, that upon seeing a man running, whom I take to be Forrest, I judge that Forrest is in a hurry. If the man I've seen is not Forrest, my judgment is in error through misidentification. If the man is not in a hurry, my judgment is in error through mispredication.

Bracketing the possibility that both errors occur at the same time, we can say that a belief is in error through misidentification when the correct property is attributed to the wrong person or object. In what follows, when I use this characterization (being in error through misidentification means getting the property right but the subject wrong) it should always be borne in mind that it is really an independent question whether the correct property is being attributed.

Note also that the correct property being attributed to the wrong person does not imply that the belief is false. For instance, my belief 'Forrest is running', when based on seeing Tom running, would be in error through misidentification even if, by chance, Forrest actually were running. That is, even if my belief happens to be true, I would attribute the property to the wrong person, not in the sense that Forrest isn't running, but in the sense that the person whom I've seen running (i.e. the person who I know to be running) isn't Forrest.[1] The point is that the judgment is in error through misidentification relative to a certain basis, for instance relative to its grounds or its source of information; in this case: relative to my seeing Tom running.

The classical Immunity Thesis is the claim that error through misidentification is not possible in present-tense introspection-based self-ascriptions of mental states. Very roughly, this means that if one self-ascribes a mental state of which one is introspectively aware then one cannot go wrong in ascribing the state to the wrong subject. Witness John Campbell:

> [F]irst-person present-tense psychological judgments are characteristically immune to error through misidentification. If you think "I hear trumpets," you might be making a mistake about whether it is trumpets you are hearing. But you cannot have any ground for doubt about whether it is you who is hearing trumpets that is not also a ground for doubt about your evidence that trumpets are being heard. This immunity to error through misidentification is a datum. (Campbell 1999a: 91)

Now, calling immunity a datum is imposture; it's not a datum, it's a debatable claim (a claim that Campbell himself aims to challenge with the phenomenon of thought insertion). To better see what this claim amounts to, contrast introspection-based self-ascriptions with a typical exteroception-based self-ascription. Suppose I judge that my coat is dirty upon seeing

[1] The view that accidentally true judgments can still be in error through misidentification is explicitly adopted by Pryor (1999: 298, fn. 9) and is implied by Shoemaker's definition of error through misidentification (1968: 557).

a reflection in a shop window. This judgment is vulnerable to error through misidentification as it may happen that I mistake somebody else's reflection for mine. In that case, I would attribute the property which I correctly perceived (wearing a dirty coat) to the wrong subject. Again, I am the wrong subject of attribution not in the sense that my coat isn't dirty (for even if my coat happened to also be dirty, the judgment would still involve a misidentification), but in the sense that the dirty coat I've actually seen is not mine, but somebody else's.

When self-ascriptions are based on introspection, the Immunity Thesis states, this kind of error is not possible. A hallmark of judgments that are immune to error through misidentification is that it does not make sense to wonder, concerning these judgments, whether one is getting the subject right: 'Surely, somebody is F, but is it *a* that is F?' Consider Campbell's example: When I believe to be hearing trumpets, while it makes sense to wonder 'Surely, I am hearing something, but is it trumpets that I am hearing?' it does not make sense to wonder 'Surely, somebody is hearing trumpets, but is it *me* who is hearing trumpets?'

Many writers hold that such self-ascriptions, self-ascriptions which are immune to error through misidentification, are fundamental to self-consciousness (see e.g. Bermúdez 2003a). By self-consciousness I mean consciousness of oneself *as oneself*. Ernst Mach famously illustrates the idea with this story: upon entering a bus, he saw himself in a mirror. Not knowing that he was seeing his own reflection, he thought 'what a shabby pedagogue' (cf. Mach 1922: 3, fn. 1). While this judgment happened to be about himself, it wasn't about himself *as himself*; hence, it wasn't a self-conscious judgment. Only after Mach had recognized that it was he himself whom he saw in the mirror, did he entertain a self-conscious judgment about his appearance, a judgment about himself *as himself*. Self-conscious self-ascriptions, that is ascriptions to oneself *as oneself*, are simply called *I-thoughts*, as their natural linguistic expression involves the first-person pronoun.

Obviously, the I-thought 'I look like a shabby pedagogue' is based on external perception and involves a fallible identification. So, while it is a self-conscious judgment, it is not immune to error through misidentification. Introspective awareness of one's mental states, in contrast, gives rise to self-conscious judgments that are immune to error through misidentification. Many writers have claimed that immune I-thoughts lie at the core of self-consciousness (cf. Smith 2006, Hamilton 2007, Chen 2009, Lane & Liang 2011: 81f., Musholt 2011).

The most common argument for the centrality of immune I-thoughts is the regress argument.[2] It holds that every I-thought must be based fundamentally on an immune I-thought. Roughly, it goes as follows. Every I-thought is either immune to error through misidentification or not. If it is immune, it is a fundamental I-thought already. For instance, the thought 'my head hurts' (when based on awareness of that pain) is of this kind. If an I-thought is not immune to error through misidentification, it must be based on another I-thought. The thought 'I look like a shabby pedagogue', for instance, is not immune, but is based on the further I-thought 'I = the person in the mirror'. Now, that underlying I-thought must itself either be immune and thereby basic, or not immune and therefore based on yet another I-thought. If it is based on another I-thought, that underlying thought, again, is itself either immune or based on another I-thought, and so on. Since this inference cannot go on infinitely, every I-thought must be anchored fundamentally in an immune I-thought.

Immunity to error through misidentification also plays a role in discussions of Self-Knowledge, by which I mean a person's knowledge of her own mind (or, more precisely, the discussion concerns very generally a person's epistemic relation to her own mind). Subjects are often said to have privileged access to their own minds, but it is highly controversial what that exactly means, i.e. in which way the epistemic access is privileged. For instance, the Cartesian view that subjects are infallible and omniscient with respect to their mental life is no longer held these days. The assumption

[2] The *locus classicus* is Shoemaker 1968: 561. See also Bermúdez 2000: 7.

that self-ascriptions of mental states are immune to error through misidentification offers a way to spell out the privileged access a person may enjoy vis-à-vis her mental life. It can be construed, so to speak, as an attenuated infallibility claim. A subject may be wrong about the content and nature of her mental states, but she cannot be wrong with respect to the question whose mental states they are.

Various forms of critique have been raised against the Immunity Thesis (henceforth: *Immunity*), both theoretically and empirically motivated. In this thesis, I discuss whether certain pathological cases of alienation, in particular thought insertion, refute Immunity. To properly assess whether an alleged counterexample actually undermines the Immunity Thesis, we have to get clear on what exactly the thesis states. In particular, two questions have to be answered.

- Definition of misidentification: What is an error through misidentification?
- Scope of the thesis: Which judgments are claimed to be immune to such error?

Unfortunately, both questions have been neglected by those who claim to have refuted the thesis. This shortfall shall be corrected in this thesis.

Here is a brief overview. I begin by introducing the idea of immunity to error through misidentification (§ 2). I then argue, in a first step, that the pathological counterexamples do not threaten the Immunity Thesis because they either do not fall within the scope of the thesis or do not involve an error through misidentification (§ 3). After taking a closer look at thought insertion, particularly at the question how to understand the claim that a thought is not one's own (§ 4), I offer, in a second step, a refined critique of Immunity, building on the pathological cases (§ 5). To defend Immunity against this critique, or, more precisely to defend a version of Immunity, I appeal to a distinction between two different versions of the thesis, Epistemic Immunity and Ontological Immunity. I argue that only Epistemic Immunity, which ties the immunity of a judgment to its identification-freedom, is subject to the refined critique and that Ontological Immunity

remains unchallenged by the pathological cases (§ 6). Finally, I present several corollaries pertaining to the distinction between the epistemic and the ontological approach which underwrite my defense of Immunity (§ 7).

Finally, a brief remark on notation. Throughout this thesis, double quotes are used solely to mark quotations; single quotes are used to mark words or sentences and concepts or thoughts. All emphases in quotations are by the cited author(s) unless otherwise noted.

2. Immunity to Error through Misidentification and the First Person

In this chapter, I provide an overview of the most important topics in the debate of immunity to error through misidentification. Next to a few general points of clarification, I explain in more detail which types of judgments are claimed to be immune to error through misidentification and I present different views on what it means for a judgment to be in error through misidentification.

2.1 Linguistic Meaning and the First-Person Pronoun

Let me begin to clarify the phenomenon under discussion by explaining what it is not, i.e. by putting aside certain misconceptions. In particular, I want to quickly discard the idea that immunity is a linguistic phenomenon which has to do with the meaning of the first-person pronoun.

Wittgenstein: Two uses of 'I'

The whole idea of immunity to error through misidentification goes back to Ludwig Wittgenstein's distinction between two uses of the first-person pronoun, the use as subject and the use as object. Since it is still very influential in the debate, it is worth quoting the passage at full length:

> There are two different cases in the use of the word 'I' (or 'my') which I might call 'the use as object' and 'the use as subject'. Examples of the first kind of use are these: 'My arm is broken', 'I have grown six inches', 'I have a bump on my forehead', 'The wind blows my hair about'. Examples of the second kind are '*I* see so-and-so', '*I* hear so-and-so', '*I* try to lift my arm', '*I* think it will rain', '*I* have toothache'. One can point to the difference between these two categories by saying: The cases of the first category involve the recognition of a particular person, and there is in these cases the possibility of an error, or as I should rather put it: The possibility of an error has been provided for. [...] It is possible that, say in an accident, I should feel a pain in my arm, see a broken arm at my side, and think it is mine, when really it is my neighbour's. And I could, looking into a mirror,

> mistake a bump on his forehead for one on mine. On the other hand, there is no question of recognizing a person when I say I have toothache. To ask "are you sure that it's *you* who have pains?" would be nonsensical. (Wittgenstein 1958: 66f.)

This passage is remarkable in having sparked a long and ongoing debate about how to properly understand the phenomenon of immunity to error through misidentification. More than that, two ideas which still play a central role in the discussion of immunity today are already contained in this passage: First, Wittgenstein remarks that the use as object involves "the recognition of a particular person", thereby implying that the use as subject does not. This is the idea that immune self-ascriptions are identification-free, a point that will play an important role later. Secondly, Wittgenstein remarks that sentences involving the use as subject do not seem to allow for what I will call *the who-question*: 'are you sure it is you who is *F*?' Many writers take this to be a mark of immunity: if it does not make sense to ask the who-question, the judgment is immune to error through misidentification.

There is one aspect in Wittgenstein's writing on the subject-object distinction that I want to briefly mention just to put it aside. Wittgenstein is often interpreted as having endorsed an expressivist explanation of the phenomenon. In a nutshell, expressivism about the first person pronoun is the idea that sentences involving 'I' as subject are not self-ascriptions in any good sense, but are rather *expressions* of the mental states in questions. The passage quoted above continues thus:

> [I]t is as impossible that in making the statement "I have toothache" I should have mistaken another person for myself, as it is to moan with pain by mistake, having mistaken someone else for me. To say, "I have pain" is no more a statement *about* a particular person than moaning is. (ibid.: 67)

So, according to Wittgenstein, my assertion 'I have pain' does not express my *belief* that I am in pain, but rather expresses my pain, just like moaning does. Following this idea to the extreme, Elizabeth Anscombe (1975)

famously argues that the first-person pronoun, when used as subject, is not a referring term at all.

To be sure, expressivism does offer an explanation of the fact that uses of 'I' as subject are immune to error through misidentification. For, if there is no reference, there can be no reference to the wrong subject. The no-reference view may seem absurd at first, but it can be made more plausible if you consider the following analogy: uses of 'I' as subject are taken to not refer in the same sense in which uses of 'it' do not refer in sentences such as 'it is raining'. However, expressivism about the first person faces many severe objections. Let me just mention the strongest: Expressivism implies that the two sentences 'I am in pain' and 'MS is in pain' do not involve the same predicate '_ is in pain'. Or, putting virtually the same point slightly differently, Expressivism entails that from the judgment 'I am in pain' one cannot deduce the judgment 'someone is in pain'. I take this to be a conclusive objection and will not further discuss expressivism.

In what follows, when I discuss Wittgenstein's remarks on immunity, I deliberately ignore his expressivist backdrop. Not only did this idea never gain many followers, it is also a decidedly linguistic thesis. Indeed, Wittgenstein discusses Immunity as a linguistic phenomenon. But today it is generally agreed that Immunity is an epistemic phenomenon, a property not of utterances but of thoughts, in particular of judgments or beliefs.

Referential Mistakes

Sydney Shoemaker, in his seminal "Self-Reference and Self-Awareness" (1968), is the first to pick up and elaborate on Wittgenstein's distinction. It is Shoemaker who coins the term 'immunity to error through misidentification'. He offers the following definition:

> [T]o say that a statement "*a* is *φ*" is subject to error through misidentification relative to the term '*a*' means that the following is possible: the speaker knows some particular thing to be *φ*, but makes the mistake of asserting "*a* is *φ*" because, and only because, he mistakenly thinks that the thing he knows to be *φ* is what '*a*' refers to. (1968: 557)

The definition reveals that Shoemaker, following Wittgenstein, still discusses immunity to error through misidentification in predominantly linguistic terms (even if Shoemaker's discussion of immunity leads the way to understanding the matter as a phenomenon of self-knowledge rather than semantics). As a result, his definition has mislead several authors to apply the term 'error through misidentification' to a certain kind of referential mistake, viz. a divergence between intended reference and linguistic reference. Consider the following case by James Pryor:

> [CARNAP] Without turning and looking, I point to a place on the wall behind me, which has long been occupied by a picture of Rudolf Carnap, and I say, "That is a picture of a great philosopher." Unbeknownst to me, someone has replaced my picture of Carnap with a picture of Spiro Agnew. (1999: 277)

On some views, what I am claiming in this case is that Agnew is a great philosopher. But, as Pryor argues, that is not a case of error through misidentification, but a case of what he calls an error through 'badly aimed reference' (cf. ibid.: 276ff.). For obviously, my intention was not to refer to Agnew, but to Carnap, and what matters for error through misidentification is what Pryor calls the 'basic referential attempt' rather than the actual linguistic reference. However, note that the case can be construed to fit Shoemaker's definition: I know a particular person to be a great philosopher and I make the mistake of asserting 'this person (I am pointing at) is a great philosopher' because, and only because, I think that the person I know to be a great philosopher is what 'this person' refers to. Even though the case fits Shoemaker's definition, I think that also Shoemaker would not have considered this a case of error through misidentification. His definition, it seems, is a bit too wide.

In fact, Shoemaker's definition, considered in the light of Wittgenstein's claims regarding the first-person pronoun, has misled for instance Rovane (1987) and Christofidou (1995) to claim that virtually all uses of the first-person pronoun are immune to error through misidentification. Their claim is based, first, on the misunderstanding that error through misidentification is a divergence of intended and actual reference, and, second, on the

assumption that such a divergence cannot afflict the first-person pronoun. Intuitively, one can hardly intend to refer to anyone other than oneself with the first-person pronoun, and the actual reference of the first-person pronoun also cannot be anyone other than oneself. But the guaranteed reference of the first-person and the guaranteed knowledge of the reference is not what immunity to error through misidentification is about.[3] Apart from the fact that, even on the linguistic level, this is arguably not the notion of error through misidentification that Shoemaker has in mind, it is also a notion that does not have an equivalence on the level of thought. We do not find any corresponding split between intended and actual reference in thought.[4] Being interested in the immunity of I-thoughts, I will assume that badly aimed reference is different from error through misidentification.

I have said that Wittgenstein and (for the most part) Shoemaker discuss Immunity on the linguistic level: they are concerned with utterances and sentences. Yet, I will discuss Immunity on the mental level, i.e. as a thesis about thoughts. However, as far as the description of the phenomenon is concerned, the Wittgenstein-Shoemakerian ideas can easily be transferred to the discussion of mental Immunity and I will henceforth just treat them as if they were directed at mental Immunity.

2.2 The Immunity Thesis

Having established that immunity to error through misidentification is an epistemic rather than linguistic phenomenon, I want to say, in this section, a bit more on what the Immunity Thesis exactly claims. In particular, I will address the fact that immunity is exhibited not only in first-person thought,

[3] See Coliva (2003) for criticism and a lucid correction of Rovane's and Christofidou's misunderstanding.

[4] Note how Pryor, who discusses immunity on the level of thoughts, suddenly talks of assertions and utterances in distinguishing error through misidentification from badly aimed reference. I assume this is so because there simply is no such thing as badly aimed reference in thought.

but also in demonstrative thought, and I will stress how immunity is not a property of beliefs as such, but of beliefs founded on a particular basis.

Demonstrative Immunity

Immunity is taken by most to be exhibited not only in first-person thought, but also, for instance, in demonstrative ways of thinking about objects, time, or space. In this thesis, I am concerned solely with the immunity of first-personal thought. However, it will be helpful for the understanding of the phenomenon to briefly present how it is exhibited in demonstrative ways of thinking.

Shoemaker himself suggests that there are cases of demonstrative reference that exhibit immunity. Shoemaker's example (NECKTIE-1) is that in seeing a red necktie and demonstratively saying of it "This [necktie] is red" one cannot be in error through misidentification with respect to the question which object is red (1968: 558). However, two important points are noted by Shoemaker. First, not all demonstrative thoughts are immune to error through misidentification. For, a demonstrative judgment can also be based on an identity assumption. Here is Shoemaker's variation on the necktie example:

> [NECKTIE-2] Suppose that I am selling neckties, that a customer wants a red necktie, and that I believe I have put a particular red silk necktie on a shelf of the showcase that is visible to the customer but not to me. Putting my hand on a necktie on that shelf, and feeling it to be silk, I might say "This one is red." (1968: 558)

In this case there is room for error through misidentification because my judgment 'This one is red' is based on my identity assumption that the tie I am touching is the tie I know to be red.

Secondly, Shoemaker says that the explanation of demonstrative immunity cannot be applied to immunity of the first-person. His explanation of demonstrative immunity is that "the speaker's intention determines what the reference of his demonstrative pronoun is and that reference cannot be other than what he intends it to be." (1968: 558) For instance, in NECKTIE-1, my demonstrative 'this' cannot refer to anything other than the object I

intend to refer to, viz. the necktie that I see. Shoemaker claims that in cases where the intention fully determines reference demonstrative judgments are identification-free (and hence immune):

> [It is possible for] the reference of 'this' on a particular occasion to be fixed by the speaker's intention to say of a particular thing that it is red, i.e., fixed in such a way that it can refer to nothing other than that thing and, consequently, in such a way that his statement "This is red" does not involve an identification. (ibid.: 558f.)

He goes on to argue that this explanation of immunity (intention determines reference), cannot be applied to the case of the first-person. For, the speaker's intention does not play any role in determining the reference of the first-person pronoun.

Let me briefly discuss Shoemaker's explanation of demonstrative immunity as this will underwrite the importance of treating immunity to error through misidentification as an epistemic rather than linguistic phenomenon. I do agree with Shoemaker that the intention's determining the reference cannot explain first-person immunity. But neither, I submit, does it explain demonstrative immunity. Consider the difference between NECKTIE-1 and NECKTIE-2. Again, it is important to note that Shoemaker discusses both cases on the linguistic level. He construes NECKTIE-2 as a case in which the linguistic reference of the demonstrative is not fully determined by the intention, or more precisely, as a case in which there are two intended references and one of the intended references diverges from the actual (linguistic) reference:

> I intend to refer to a certain red necktie I believe to be on the shelf, but there is also a sense in which I intend to refer, and do refer, to the necktie actually on the shelf, and there is a possibility of a disparity between my intended reference and my actual reference. (ibid.)

My critique is that the coming apart of intended and actual reference is a purely linguistic phenomenon for which there is no counterpart on the level of thought. The difference between NECKTIE-1 and NECKTIE-2, on the level of thought, is not that intention fully determines reference in the first and fails to fully determine reference in the latter. Rather, the reason why the

latter judgment is liable to error through misidentification is simply due to the identity assumption that the felt tie is the same as the previously seen tie.

To see that intention's determining the linguistic reference does not guarantee immunity, consider the following case. Suppose that, looking at a tie I say of it 'This tie was used by Fink & Mao to determine the aesthetic value of different tie knots' (cf. Fink & Mao 1999). In this case, my demonstrative 'this' is determined to refer to the tie I am seeing, just like in NECKTIE-1 my demonstrative 'this' is determined to refer to the tie I am seeing in my judgment 'This is red'. But of course, my claim that this is the tie used by Fink & Mao is liable to error through misidentification. Hence, the fact that my intention determines the linguistic reference of my demonstrative does not imply identification-freedom.

I believe the correct explanation of demonstrative immunity lies in the fact that reference and predication are based on the same source of information. Let's focus on visual demonstratives. In NECKTIE-1, my reference is solely determined by my intention to refer to the object I see and my predication is equally based on my perception. In NECKTIE-2, in contrast, my predication is based on my memory of knowing something to be red whereas my reference is determined (at least partly) by my haptic perception.

A view along these lines, although a lot more fine grained, has been developed by Campbell (1999a: §§ 3–4). To be precise, Campbell has some reservations about the immunity of demonstrative color ascriptions, as in NECKTIE-1. He offers the following counterexample:

> [YELLOW] [I]f you judge, 'that chair is yellow', it may be that you thereby know of something that it is yellow, but that thing is not the chair, if, for instance, the chair is transparent and set against a yellow background. (1997: 70)

He takes this to show that demonstrative ascriptions of colors are not immune to error through misidentification. However, Campbell fully agrees that demonstrative ascriptions of an object's location are immune to error through misidentification. Even if, due to refractions or mirrors, the object

is not actually at the place at which it appears to be to the subject, her judgment 'this chair is two meters in front of me' cannot be in error through misidentification. In that case, her judgment would be a mispredication, rather than a misidentification, for she wouldn't know of some other thing that it is two meters in front of her, but she would just be wrong about where the chair is. As Campbell puts it, in demonstrative judgments that ascribe a location to an object, "you are using the perceived location of the [object] to single out which [object] you have in mind." (1999a: 96; see also 1997: § 4) Therefore, the object that one judges to be at a certain location, cannot be any other than the object that one perceives to be at a certain location. And that means that one cannot be wrong about which object it is that one perceives to be at a certain location.

I will not further discuss the immunity of demonstrative judgments. One important lesson is that in demonstrative thoughts, just like in first-person thoughts, immunity cannot be explained as a linguistic phenomenon or by reference to the referential rules for demonstratives.

Relativity to Grounds

The idea that immunity to error through misidentification is an epistemic rather than linguistic phenomenon is reflected in the fact that all writers, apart perhaps from Wittgenstein, agree that it is not utterances or beliefs *per se* which are immune, but utterances or beliefs relative to the particular grounds or information they are based on.[5] To stress the role of the basis, I like to say that certain beliefs are immune to error through misidentification *in virtue of* their bases, or that the bases *confer* immunity on the resulting beliefs. In the present debate immunity is mostly attributed to judgments (rather than beliefs), and some authors imply the relativity to bases already in the use of that terminology. Prosser, for instance, holds that "IEM is a property of *judgments* (where a judgment is the formation of

[5] See e.g. Evans 1982: § 7, Bermúdez 2000: § 1, 2003b: 216, Smith 2006: 274f., Prosser 2012: 160.

a belief or knowledge based on specific reasons and/or evidence); it is not a property of a specific belief *per se*" (2012: 160).

Gareth Evans was the first to make this relativity to bases explicit (cf. 1982: 219). To take his example, one and the same belief, that my legs are crossed, is immune when based on proprioception, but is not immune when based on visual perception of one's legs. As the example suggests, in being a self-ascription of a bodily rather than a mental state, Evans's point about the relativity to bases is closely connected to another point. Shoemaker put a lot of emphasis on the question what kind of predicate is self-ascribed, his guiding idea being that self-ascriptions of *mental predicates* are immune to error through misidentification in a particularly fundamental way. Evans, in contrast, focuses on the bases of self-ascriptions, his guiding idea being that it is the first-personal nature of the information channel that confers immunity on a judgment. As a result, he was open to the idea that bodily self-ascriptions, when based on proprioception, are immune to error through misidentification as well.

All that being said, it should be noted that the idea of the relativity to grounds is already present, even if not very explicit, in Shoemaker (1968). The most obvious witness is Shoemaker's distinction between *absolute* and *circumstantial* immunity (cf. 1968: 557). Paradoxically, a common misunderstanding of this distinction has led many to claim the exact opposite, viz. that Shoemaker denies the role of the grounds for certain self-ascriptions. Let me elaborate. For a statement to be circumstantially immune means that the statement can be immune when made on a certain basis but not immune when made on a different basis. As an example, Shoemaker considers the statement 'There is a table in front of me'. It is circumstantially immune to error through misidentification when based on visual perception as of a table in front of oneself. However, if the same judgment were made based on seeing oneself in a mirror as if standing in front of a table, the judgment is not immune – it could happen that one mistakes somebody else's reflection for one's own (cf. ibid.: 557). According to Shoemaker, things are different in self-ascriptions of mental states. These are *absolutely* immune. That is to say that they are always immune,

always—one is tempted to add—no matter what grounds they are based on. This would mean, that in cases of absolute immunity, the grounds do not matter after all. But that would promote a misunderstanding of Shoemaker's notion of absolute immunity. For, Shoemaker does not think the basis did not matter. Rather, Shoemaker thinks that when a mental predicate is self-ascribed, it is necessarily self-ascribed on the basis of introspection.

I want to stress this point because the notion has often been misunderstood as implying that the basis does not matter. For instance, Chen says being absolutely immune to error through misidentification means that "there are no possible grounds with respect to which [judgments] are subject to error through misidentification" (2009: 28f.), Howell claims that "judgments have absolute IEM iff they are IEM when made upon *any* ground, and not just upon some ground or other" (2007: 293), and Pryor, contrasting immunity relative to certain grounds with absolute immunity, says that "a proposition is absolutely immune to *de re* misidentification just in case it is immune to *de re* misidentification when justified by *every* possible ground for believing it." (1999: 279; Pryor's italics, bold emphasis omitted). All these claims are false. Shoemaker holds that mental self-ascriptions are immune when they are based on introspection. His claim that mental self-ascriptions are absolutely immune to error through misidentification can be understood as a claim about the semantics of mental predicates: Knowing the predicate to apply to oneself implies that this is known through introspection. If a predicate is self-ascribed on solely third-personal grounds, Shoemaker would either deny that the subject *knows* the predicate to be instantiated at all or he would deny that the predicate in question belongs to the class of core mental predicates which he labels *P**-predicates (the self-ascription of which he takes to be absolutely immune to error through misidentification). In other words, the reason why Shoemaker takes the self-ascription of *P**-predicates to be absolutely immune to error through misidentification is that, with conceptual necessity, if they are known to apply to oneself they are known to apply in a first-personal way.[6]

[6] However, Shoemaker does not claim that they can *only* be known to apply to oneself in a first-personal way. It is possible that one knows a *P**-predicate to apply

This is why, contrary to appearances, the grounds do play a decisive role also in Shoemaker's notion of absolute immunity.[7]

2.3 The Scope of Immunity

The classical Immunity Thesis holds that present-tense introspection-based *de se* self-ascriptions of mental states are immune to error through misidentification. Wittgenstein and Shoemaker illustrate the thesis with self-ascriptions of phenomenal states, beliefs, actions or intentions to act, and perceptions. Although they do not always explicitly say so, the self-ascriptions in these examples are to be understood as based on first-personal awareness of these states. Today, many other types of self-ascriptions have been claimed to be immune to error through misidentification (see below). However, the counterexamples I discuss in this thesis are clearly directed against the classical version and I will restrict my discussion accordingly.[8]

Of the many criteria that delineate the scope of the thesis (present-tense, self-ascription, *de se* mode, mental state, introspection-based), I suggest that two characteristics are of particular importance and capture the essence of the Immunity Thesis: First, the restriction to self-ascriptions, and second, the restriction to beliefs based solely on introspection. The first restriction specifies the content of the judgments that fall within the scope of Immunity: only self-ascriptions are claimed to be immune. I call this the *self-ascription constraint*. The second restriction specifies the grounds of the judgments that fall within the scope of Immunity (or, as I like to say, it specifies the kinds of grounds that yield immune judgments): only judg-

---

to oneself in both a first-personal way and additionally in a third-personal way. However, it is not possible, according to Shoemaker, to know solely in a third-personal way that a *P**-predicate applies to oneself.

[7] For an astute discussion and critique of absolute vs. circumstantial immunity see also Coliva 2006: 422, fn. 32.

[8] Yet, many of my arguments apply equally to the discussion of, for instance, bodily immunity (see e.g. Mizumoto & Ishikawa 2005, Lane & Liang 2011).

ments that are based on first-personal grounds are claimed to be immune. I call this the *introspection constraint*. These two criteria will receive a lot more discussion in what follows. For now, let me briefly explain the other criteria.

The *de se constraint* restricts the scope of Immunity to so called *I-thoughts*, i.e. to those thoughts the natural linguistic expression of which involves the first-person pronoun. The *de se* constraint can be understood as a refinement of the self-ascription constraint in the sense that not all self-ascriptions are in the *de se* mode, but (arguably) all ascriptions made in a *de se* mode are self-ascriptions. Self-ascriptions that are made in a mode that is not *de se* can be either *intentional* (non *de se*) self-ascriptions or, as I will say, *accidental* (non *de se*) self-ascriptions. For an accidental self-ascription, consider Perry's famous case in which he believes that the shopper with the torn sugar bag is making a mess (1979). As Perry puts it, he ascribes a property to "the person he happens to be" (cf. 1998). For an intentional self-ascription, imagine that I want to express my wish for a coffee break in a funny way by saying that the most tired person in the office wants to have a coffee. In this case, I intend to refer to myself, but I do so using the description 'the most tired person in the office'.

When proponents of Immunity claim that certain self-ascriptions are immune to error through misidentification, they typically do not mean to include self-ascriptions that involve a description, demonstrative, or name; neither if the self-ascription is accidental, nor if the self-ascription is intended.[9] Rather, the Immunity Thesis is traditionally taken to hold only for those self-ascriptions that are made in a *de se* mode. I will later discuss the *de se* constraint in more detail (§ 7.3). For now, note that the *de se* constraint, construed as a refinement of the self-ascription constraint, does not play any role in the cases under discussion. In what follows, unless noted otherwise, assume that the self-ascription constraint implies the *de se* constraint. I.e., to satisfy the self-ascription constraint a judgment has to be

[9] For a curious exception, see Shoemaker 1970: 270, fn. 5. I will discuss Shoemaker's idea in more detail later (see § 7.3).

in the *de se* mode. Accordingly (unless noted otherwise), when I speak of self-ascriptions of mental states I mean only *de se* self-ascriptions, not accidental or intentional non *de se* self-ascriptions.

Similarly, the restriction to mental state ascriptions, although a typical restriction in the debate, will not play any role in this work. First, the judgments under discussion plausibly involve the ascription of a mental state so that there is not much need to discuss this restriction. Second, I take this restriction to be somewhat superfluous in general: a restriction to mental states is already in place in virtue of the introspection-criterion. That is to say, the kinds of judgments that the mental state restriction would exclude from the scope are already excluded by the introspection criterion: we simply do not have introspective awareness of, say, our date of birth, our weight, or our haircut. Yet the other way around, we can make judgments about mental states that are not based on introspection. I take the introspection constraint to be the more fundamental and essential one (see also my discussion of Shoemaker's notion of absolute vs. circumstantial immunity in § 2.2).

Finally, there is the restriction to present-tense ascriptions. As we will see in the discussion of cross-wiring cases (§ 2.5), there is a long and complicated debate on the question whether memory based judgments are immune to error through misidentification. More precisely, the question is whether memory preserves the immunity of judgments the present-tense versions of which are immune. However, this question does not have any direct bearing on the pathological cases. Admittedly, reports of inserted thoughts and the like are typically memory-based reports about a past experience. But the question is not whether subjects correctly remember who was the subject of the experience. Rather, we simply assume that reports are a past-tense version of what subjects would have believed at the time of the experience. So, since the judgments we find in pathological alienation are simply treated as present-tense claims, the restriction to present-tense ascriptions is satisfied by these cases and need not be discussed further.

Different Kinds of Immune Self-Ascriptions

I have said that the classical Immunity Thesis is a thesis about introspection-based self-ascriptions of mental states. These are the kinds of self-ascriptions that Shoemaker has assumed to be logically immune to error through misidentification (see § 2.5). When I speak of introspection, I simply mean to denote the first-personal direct way in which we are typically aware of our occurrent conscious mental states. Take one of Wittgenstein's examples: my self-ascription that I am in pain, based on my experiencing the pain, is immune to error through misidentification in the strongest possible sense. While these cases certainly are at the core of the Immunity Thesis, they are not all there is. A number of different types of self-ascription have been claimed to also be immune. It is a matter of some controversy whether the thesis holds for these self-ascriptions as well, or with which modal force it holds. In this section, I introduce these other types of self-ascriptions without further discussing the question whether they truly are immune.

To start with an obvious candidate, since Evans (1982) the idea of Immunity is no longer restricted to self-ascriptions of *mental* states, but has been broadened to include also self-ascriptions of *bodily* states. Evans introduces two ways in which self-ascriptions of bodily states (or, more broadly, self-ascriptions of non-mental states) can be immune. On the one hand, self-ascriptions regarding one's bodily states are immune when based on proprioception. The term 'proprioception' is sometimes used in a narrow sense, denoting a specific inner sense modality that can be distinguished from other inner senses such as interoception and kinesthesia. Here, I want 'proprioception' to be understood in a broad sense as denoting all specifically first-personal ways of being aware of body states. Witness Evans's oft-quoted insight:

> we have what might be described as a general capacity to perceive our own bodies, although this can be broken down into several distinguishable capacities: our proprioceptive sense, our sense of balance, of heat and cold, and of pressure. Each of these modes of perception appears to give rise to judgements which are immune to er-

> ror through misidentification. None of the following utterances appears to make sense when the first component expresses knowledge gained in the appropriate way: 'Someone's legs are crossed, but is it my legs that are crossed?'; 'Someone is hot and sticky, but is it I who am hot and sticky?'; 'Someone is being pushed, but is it I who am being pushed?' (1982: 220f.)

On the other hand, Evans claims that external perceptions can ground immune judgments about our position, orientation and relation to external objects. As examples he suggests:

> knowing that one is in one's own bedroom by perceiving and recognizing the room and its contents; knowing that one is moving in a train by seeing the world slide by; knowing that there is a tree in front of one, or to the right or left, by seeing it; and so on. Once again, none of the following utterances appears to make sense when the first component expresses knowledge gained in this way: 'Someone is in my bedroom, but is it I?'; 'Someone is moving, but is it I?'; 'Someone is standing in front of a tree, but is it I?' (1982: 222)

Since these who-questions do not make sense, Evans argues, beliefs such as 'I am standing in front of a tree', based on visual perception as of a tree in front of oneself, are immune.

Do not confuse this type of judgment with another perception-related judgment that is often claimed to be immune. Most perception-predicates are somewhat ambiguous in that they can denote a factive perceptual state and a conscious experience. Let us focus on visual perception. In claiming that I see a canary, I could self-ascribe a perceptual state (standing in a perceptual relation to an external object in the sense of: light coming from that object falls into my eyes etc.) or I could self-ascribe a visual experience as of seeing a canary. Let's say, accordingly, that my judgment 'I see a canary' can be construed as a *perceptual* or as a *phenomenal* claim. Suppose, now, that my judgment is actually based on having a hallucination as of a canary. In this case, I suggest, my self-ascription is false if interpreted as a perceptual claim, but it is correct if interpreted as a phenomenal claim. Both types of self-ascriptions have been claimed to be immune to

error through misidentification. The phenomenal self-ascription falls within the core of Immunity, as it is the self-ascription of an occurrent conscious state based on the experience of that state. It is hard to see how one could be wrong about the question who is the subject of a visual experience that one is aware of.[10] The self-ascription of the perceptual state is less clearly immune. While Shoemaker and many others have taken these self-ascriptions to be immune, they seem to be open to more potential counter-examples than the self-ascription of the phenomenal experience.[11]

External perception features in yet another type of judgment which could be claimed to be immune.[12] Evans famously argues for the transparency thesis regarding self-knowledge which is the claim that knowledge of one's own beliefs is (often) not based on introspection, but on external perception.

> [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?' (1982: 225)

If the transparency thesis is correct, there is another candidate type of judgment that we should expect to be immune, viz. self-ascriptions of beliefs, based on external perception rather than introspection.

Finally, there is the controversial case of memory judgments. More precisely, what is under discussion is the question whether self-ascriptions of past experiences are immune when based on episodic memory. The idea is quite intuitive: when I believe that I once had a really good coffee at

---

[10] But see Langland-Hassan (forthcoming) for a putative counterexample.

[11] Cf. Shoemaker (1968: 557) for the claim and (Smith 2006: 278) for a putative counterexample.

[12] Thanks to an anonymous reviewer for *Philosophers' Imprint* for pointing this out.

Bonanza Coffee Heroes, based on my episodic memory of what it was like to have a coffee there, I cannot be wrong about who it was that had a coffee at Bonanza. Importantly, not any self-ascription of a past event that is based on episodic memory is claimed to be immune. When I judge that Jones called me a fool, based on my episodic memory of Jones calling me a fool, then, without fault of my memory I can be wrong about who it was that Jones called a fool (cf. Shoemaker 1970: 270 fn. 4). Only those past-tense self-ascriptions are claimed to be immune for which the corresponding present-tense self-ascription would have been immune, had the subject self-ascribed the property in question at the time that she underwent the experience which she is episodically remembering. Here is how Shoemaker puts it:

> [If] I could not have been [in error through misidentification] in the past in asserting what I then knew by saying "I *am Φ*," then my subsequent memory claim "I *was Φ*" will be immune to error through misidentification relative to 'I'; that is, it is impossible in such cases that I should accurately remember someone being *Φ* but mistakenly take that person to be myself. (1970: 270)

What this really means is that episodic memory is claimed to *preserve* immunity, rather than *confer* immunity.

The question whether memory really does preserve immunity is highly controversial (see § 2.5 and § 7.1). Not at all controversial I take to be François Recanati's idea that certain judgments enjoy what Recanati calls derivative immunity, when they are inferentially based on another judgment that is immune, where the inference in question does not affect the subject of the judgments. To take Recanati's example, the judgment 'This man is in a hurry' is derivatively immune, when based on the judgment 'This man is running'. Immunity is preserved in this inference, because the hurry-judgment, so to speak, anaphorically inherits its subject ('this man') from the running-judgment (cf. 2012a: § 1.2). When it is claimed that immune judgments are not based on inference, this should be taken to mean that they are not based on an inference which affects the subject of the judgment.

Summing up, the most general way of putting the Immunity Thesis is this: self-ascriptions that are based on a first-personal way of knowing are immune to error through misidentification. Here, the notion of a first-personal way of knowing is meant to include not just introspection and proprioception, but also ways of knowing such as knowing through external perception about one's relation to other objects or knowing through external perception about one's beliefs (insofar as the transparency thesis is correct). However, in what follows I will only be concerned with the core of Immunity, that is with introspection-based self-ascriptions of mental states.

2.4 Error through Misidentification

Now, that we have a better idea of which judgments are claimed to be immune, it is time to take another look at the question what it means exactly for a judgment to be in error through misidentification, or, as these questions are intimately connected, what it means for a judgment to be immune to such error.

Immunity to Error through Misidentification and Identification-Freedom

The most dominant approach traces back to a remark by Wittgenstein, viz. that the use of 'I' as subject does not involve the recognition of a particular person. Very roughly, I will say that a judgment is *identification-free* if and only if it is not based on the recognition or identification of a person, and that it is *identification-dependent* otherwise. The most dominant view holds that a judgment is immune iff it is identification-free, and that a judgment is in error through misidentification iff it is based on a false identification component.

The intuitive idea of linking immunity to identification-freedom is this: if a judgment does not involve the identification of a particular person, there is no identification that could go wrong and hence there can be no error through misidentification; but if a judgment does involve the identification of a particular person, there is room for misidentification. Crucially, the notion of being based on an identification or not being so based is tradi-

tionally construed in epistemic terms. Here is how Evans defines his notion of identification-freedom:

> When knowledge of the truth of a singular proposition, ⌜*a* is *F*⌝, can be seen as the result of knowledge of the truth of a pair of propositions, ⌜*b* is *F*⌝ (for some distinct Idea, *b*) and ⌜*a* = *b*⌝, I shall say that the knowledge is *identification-dependent*: it depends (in part) on the second basis proposition, which I shall call the *identification component*. We might say that knowledge of the truth of a singular proposition is *identification-free* if it is not identification-dependent. (Evans 1982: 180)

Roughly, a judgment is identification-free if it is not based on an identity judgment.[13] To illustrate, my judgment that Forrest is running can be seen as based on the judgment 'this person (which I see) is running' and on the identification-component 'this person = Forrest'. Hence, the judgment is identification-dependent. In contrast, my judgment 'I see a canary' does not seem to be based on an identification-component that identifies me as the person who is seeing something. Hence, it is identification-free with respect to 'I'.

Given this notion of identification-freedom, Evans equates identification-freedom with immunity:

> Clearly, judgements of the first kind [identification-free] are immune to a kind of error to which judgements of the second kind [identification-dependent] are liable. Since they do not rest upon an identification, they are immune to error through misidentification. (ibid.: 182)

Note that, in this passage, Evans describes identification-freedom as both a necessary and sufficient condition for immunity. That is, he more or less equates the two notions (more cautiously: he takes them to be co-extensional) (see also ibid.: 188f.). For he claims that identification-free

[13] To be precise, Evans restricts his notion of identification-freedom to judgments that are "based on a way of gaining information from objects" (1982: 180f.). Without this restriction, judgments that are completely groundless would count as identification-free.

judgments are immune whereas identification-dependent judgments are not.[14]

Many authors follow Evans in assuming a biconditional relation between error through misidentification and a false identification component on the one hand, and between immunity and identification-freedom on the other. Annalisa Coliva, for instance, defines error through misidentification and immunity to that error as follows: "a judgment of the form '*a* is F' is affected by error through misidentification if and only if the subject's [justification] for that judgment contain[s] a mistaken identification component" and "a judgment of the form '*a* is F' is immune to error through misidentification if and only if the subject's [justification] for that judgment do[es] not contain any identification component" (2006: 420).[15] Similarly, Recanati holds that "a singular judgement '*a* is F' has the property of IEM just in case its immediate grounds do not involve an identity '*a* = *b*'" (2012a: 183).[16]

Shoemaker also assumes a close connection between immunity and identification-freedom. However, as far as explicit claims go, he does not claim a biconditional, but only an implication in one direction: According to Shoemaker, "identification necessarily goes together with the possibility of misidentification" (1968: 562). That is to say, identification-dependence implies liability to error through misidentification, or, contrapositively, immunity implies identification-freedom. However, we can assume that Shoemaker would also subscribe to the other conditional (identification-freedom implies immunity). After all, he assumes that the identification-

[14] Other writers actually equate the two notions: "To say that '*a* is F' is IEM is to say that it does not depend on the identification of *a* with something that one knows to be F." (Chen 2009: 27)

[15] Where I inserted 'justification', Coliva distinguishes between two kinds of justification: a subject's own rational grounds and the background presuppositions of the judgment. This distinction is not relevant to the matter under discussion.

[16] However, as we will see in § 2.6, Recanati has a very different view from Coliva about what it means for grounds to involve an identification component.

freedom of mental self-ascriptions explains their immunity.[17] In any case, I know of only one writer, James Pryor, who explicitly rejects this conditional (see 1999: 292, for a discussion of his argument see § 7.1).

Two Kinds of Misidentification

Speaking of identification-freedom, it is time to introduce Pryor's (1999) distinction between two different types of error through misidentification. Pryor distinguishes what he calls *de re misidentification* from *which-object misidentification*. Very roughly, a judgment involves a *de re* misidentification iff it is based on a false identity belief of the form '*a* = *b*' where '*a*' and '*b*' are singular *de re* terms.[18] This is, so to speak, the familiar case. It corresponds to the Evansian notion of identification-dependence. To illustrate, my judgment that Forrest is running is based on the identity belief 'this man = Forrest' which is a *de re* identity belief since the concepts 'this man' and 'Forrest' are singular *de re* concepts.[19]

It is Pryor's notion of which-object misidentification which brings a new type of case into play. Pryor argues that judgments can be in error through misidentification without being based on a *de re* identity belief. This is so

[17] See e.g. Evans: "Certainly Shoemaker argues from the fact that a judgement is not immune to error through misidentification to the conclusion that it is identification-dependent" (1982: 189). See also Smith 2006: 275.

Rosenthal attributes the biconditional to Shoemaker, but does so as a result of an obvious *non sequitur*: "since 'identification necessarily goes together with the possibility of misidentification' [Shoemaker 1968], when no identifying figures in first-person reference to oneself, no misidentification is possible either." (Rosenthal 2012: 37)

[18] More precisely, a judgment is in error through *de re* misidentification iff *justification* of a judgment is based on *justification* for an identity claim of the form '*a* = *b*', where '*a*' and '*b*' are singular *de re* concepts.

[19] The distinction between *de re* and *de dicto* concepts is typically presupposed in the literature without further explanation. It may be asked, but I will not discuss this here, whether the originally linguistic distinction between *de re* and *de dicto* statements can really be applied to thoughts and concepts.

when a property is known to be instantiated by someone or other (not by any particular one), and the subject goes wrong in singling out who is the witness of this particular instantiation of the property. The identification that is involved in these cases is not an identity belief involving two *de re* singular terms. Rather, it is a belief that this thing or person is the one who is the bearer of that property, which is known to be instantiated (at least partially) independently of that belief. Pryor's famous example is this:

> [SKUNK] I smell a skunky odor, and see several animals rummaging around in my garden. None of them has the characteristic white stripes of a skunk, but I believe that some skunks lack these stripes. Approaching closer and sniffing, I form the belief, of the smallest of these animals, that it is a skunk in my garden. This belief is mistaken. There are several skunks in my garden, but none of them is the small animal I see. (1999: 281)

Pryor's main assumption behind this example is that the smell itself does not provide the subject with grounds for a singular *de re* thought about the animal that is causing the smell (cf. ibid.: 282). Therefore, we cannot construe the judgment in question as being based on the *de re* identity belief 'this (seen) animal = this (smelled) animal'.

Pryor further claims that the difference between *de re* and which-object misidentification concerns the structure of the judgments' grounds. In cases of *de re* misidentification, the justificational architecture involves, so to speak, the move form a singular judgment '*b* is F' via an identity judgment '*a* = *b*' to the judgment '*a* is F'. In which-object misidentification, in contrast, the justificational architecture involves the move from an existential claim 'something or other is F' to the singular judgment 'it is *a* that is F'.

There is some debate about the question whether which-object misidentification truly amounts to a genuinely distinct phenomenon, and if so, where exactly the difference lies. For instance, Coliva has attacked Pryor's assumption that there is a difference in the justificational architecture. Cases of which-object misidentification, she argues, can be construed analogously to cases of *de re* misidentification. In SKUNK, for instance, the judgment 'this animal is a skunk in my garden' can be construed as being

based on the predication component 'that animal (which I am smelling) is a skunk in my garden' and the identification component 'that animal (which I am smelling) = this animal (I am seeing)'. The difference between the two kinds of cases, argues Coliva, pertains to the nature of the concepts involved, rather than to the justificational architecture. That is to say, the concepts that figure in the identity beliefs in cases of which-object misidentification may not all be singular *de re* concepts, but the final judgment can still be construed as being based on an identity belief (cf. Coliva 2006).

Wright (2012) defends which-object misidentification as a genuine phenomenon. He does so by appealing to a type of case in which the predication component is based on purely general grounds.

> [JACKPOT] Suppose I know that, one way or another, someone this week has to win the roll-over jackpot. I consult a palmist, who persuades me that something extraordinarily fortunate is going to happen to me this week, and jump to the conclusion that I am going to win the roll-over jackpot. (Wright 2012: 258f.)[20]

In this type of case, Wright argues, the final singular judgment (here: 'I am going to win the jackpot') cannot plausibly be construed as being based on a singular predication and an identification. In this case, the final judgment 'I am going to win the jackpot' would have to be based on the tautological singular predication 'The person who is going to win the jackpot will win the jackpot' and the identification component 'The person who is going to win the jackpot = me'. Rather, Wright argues, the final judgment is based on an existential claim 'someone is going to win the jackpot' and the singling-out identification 'that person is me'.

I won't try to settle the dispute on whether there is a substantial difference between cases such as FORREST and cases such as JACKPOT and, if so, what that difference consists in. Rather, let me just note that both kinds of cases are agreed by almost everyone in the debate to involve error through

[20] Similar cases have been put forward by both Recanati (2012a) and Prosser (2012) in the same volume. Although the lottery case was probably introduced by Recanati or Prosser, I quote Wright's rendition since it is the most compelling.

misidentification. I say 'almost everyone', because, interestingly, Shoemaker does not agree. In a footnote that seems to have escaped everyone's notice he describes a case that fits Pryor's definition of which-object misidentification as a case which he "would not count as a case of error through misidentification" (1970: 270, fn. 4):

> [Fool] Suppose that Jones says "You are a fool," and I mistakenly think that he is speaking to me. [...] While this is a case of knowing *that* jones called someone (someone or other) a fool and mistakenly thinking that he was calling me a fool, it is not a case of knowing *of* some particular person that Jones called him a fool but mistakenly identifying that person as oneself. (ibid.)

Shoemaker's idea seems to be that for the judgment 'Jones called me a fool' to be in error through misidentification, the subject would have to have *de re* knowledge of somebody's being called a fool and go wrong in taking himself to be that person. Having solely existential knowledge that someone or other has been called a fool and going wrong in taking oneself to be that person does not amount to error through misidentification on Shoemaker's view. This point is also reflected in his definition of immunity to error through misidentification. To rehearse:

> [T]o say that a statement "*a* is *φ*" is subject to error through misidentification relative to the term '*a*' means that the following is possible: *the speaker knows some particular thing to be* φ, but makes the mistake of asserting "*a* is *φ*" because, and only because, he mistakenly thinks that the thing he knows to be *φ* is what '*a*' refers to. (1968: 557; my emphasis)

For a judgment to be in error through misidentification, according to this definition, the subject needs to have knowledge of a *particular* person or object to be F. Cases in which a subject merely has knowledge that something or other is F, but goes wrong in singling out the witness of that property, do not satisfy this definition.

This point is interesting for two reasons. First, the insight that there are two kinds of misidentification is unanimously attributed to Pryor (1999). But as we have seen, Shoemaker has envisaged a case of which-object misidentifi-

cation long before Pryor. Yet, in contrast to Pryor, Shoemaker didn't think this was a case of error through misidentification at all. Secondly, in assessing a dispute between Shoemaker and Evans, Pryor writes that he does not know which of the two notions of error through misidentification Shoemaker has had in mind. As I have shown, Shoemaker quite explicitly has *de re* misidentification in mind. Yet, Pryor attributes the notion of which-object misidentification to Shoemaker, as this allows a better defense of Shoemaker's claim that memory-based self-ascriptions of past experiences are not logically immune to error through misidentification (cf. Pryor 1999: 288).

Be that is it may: Apart from Shoemaker, everyone accepts that cases like SKUNK, JACKPOT, and FOOL involve error through misidentification. So do I. While the debate focuses on distinguishing between two types of *error*, Lafraire (2013) extends the idea to distinguish between two types of *identification*. Lafraire labels them *identification-1*, which is the singling-out identification that is involved in cases of which-object misidentification, and *identification-2*, which is an identity assumption involving two singular (de re) concepts and which is involved in cases of *de re* misidentification. This move allows us to hold on to the idea that an error through misidentification occurs iff the judgment is based on a false identification. We can accommodate cases of which-object misidentification in this definition by construing the notion of being based on a false identification to encompass both kinds of identification. Henceforth, when I speak of an identification-component or a judgment being based on an identification I will mean by this both types of identification. Immunity to error through misidentification can then be construed in terms of identification-freedom while doing justice to both kinds of misidentification. However, there is another way of defining immunity which may be thought to capture cases of which-object misidentification more naturally. It is this definition that I now turn to.

The Impossibility of Retreat to Existential Generalization

The, so to speak, classical definition of immunity to error through misidentification in terms of identification-freedom has been rivaled in the more

recent discussion by the definition in terms of the impossibility of retreat to an existential claim. Here is what that is supposed to mean. A judgment of the form '*a* is F' is immune to error through misidentification iff it is not possible to defeat the grounds of the judgment in a way that leaves intact grounds for the existential judgment 'something is F'. Although we already find some hints towards this idea in Evans (e.g. 1982: 188), it was Wright who first explicitly defined immunity this way:

> A claim, made on a certain kind of ground, involves immunity to error through misidentification just when its defeat is *not* consistent with retention of grounds for existential generalization in this kind of way. (Wright 1998: 19)

Putting the same idea a bit more intuitively, we can say that a judgment of the form '*a* is F' is immune to error through misidentification if and only if any ground for doubt that *a* is F is *ipso facto* ground for doubt that anything is F (cf. Pryor 1999: 283). This idea is probably the most common way of defining immunity in the present debate.[21]

To illustrate, consider again the case FORREST, which is not immune. My belief that Forrest is running can be defeated by the information that the person I am seeing is not Forrest. This is a kind of defeat that leaves intact grounds for the existential claim that someone is running, namely the person I see. Judgments that are immune cannot be defeated in that way. Consider again my belief that I see a canary. My belief may be mistaken, but it is hard to see how it could be defeated in a way that leaves intact grounds for the claim that someone is seeing a canary.

An important point which is not always made explicit is this: the existential generalization has to be based on the original grounds of the singular judgment, more precisely, on those parts of the original grounds that survived the defeater. The point is nicely brought out in Coliva's description.

---

[21] See e.g. Campbell 1999a: 89, Bermúdez 2003b: 216, Coliva 2006: 409f., Hamilton 2007, Wright 2012: 256f.

> [W]hen paradigmatic error through misidentification occurs, the *very same grounds* that support the singular judgment will, if the eventual judgment is defeated, *survive as grounds for a corresponding existential generalization*. (Coliva 2006: 409)[22]

If, after the defeat of a singular judgment, a corresponding existential judgment is based on grounds that are independent of the grounds of the original singular judgment, or if they are based on the defeater itself, we do not consider this a retreat to an existential claim.[23] Such an independent existential claim does not undermine the immunity of the original singular judgment.

Now, although this definition is structurally quite different from the definition of immunity in terms of identification-freedom, they can be construed as getting at the same idea. For, what is it to defeat a judgment in a way that leaves intact grounds for an existential claim? Really, this means that the grounds of the original judgment are separable into at least two components, a predication component and an identification component, and that solely the identification component is defeated while leaving the predication component untouched. Witness Pryor who expresses pretty much the same idea in the following passage:

> there is no "part" of your justification for believing that *a* is F which could offer you knowledge that *something* is F, while leaving it an open question for you whether *a* is F. Hence, any ground you acquired for doubting that *a* is F would ipso facto be a ground for doubting that anything is F […]. (1999: 283)

---

[22] However, in another respect this definition is slightly too narrow, or misleading. Not all defeaters leave intact grounds for an existential claim. Error through misidentification is given when it is *possible* to defeat the identification component only and thereby leave intact grounds for an existential claim; but of course it is always possible to only defeat the predication component.

[23] This is a trick Pryor uses to argue that memory-based judgments are liable to error through misidentification. See Smith (2006) for a critique.

The two definitions come out as equivalent if one makes the following assumption: when a judgment's justification contains an identification component, it is possible to defeat that identification component in a way that leaves intact the predication component. The close connection between identification-freedom and impossibility of defeat that leaves intact an existential claim is nicely brought out in the following passage from Wright, in which he turns the criterion of existential generalization into a positive criterion about the nature of the grounds. Concerning the judgment "My hair is blowing in the wind" (based on first-personal awareness), Wright points out that "[t]he nature of the evidence I have is that it is evidence that *my* hair is blowing in the wind [...] or it is not evidence of anything." (2012: 250) This subjective nature of the evidence is often alluded to as an explanation of Immunity. Being first-personally aware of the instantiation of a property, it is often said, simply is to be aware of that property *as one's own*. This intuitive idea is intimately connected both to identification-freedom and to the impossibility of retreat to an existential claim.

2.5 Cross-Wiring

Both definitions, the one in terms of identification-freedom and the one in terms of existential generalization fare very well as long as we look at normal cases. In fact, as I have shown, given a plausible seeming assumption they turn out equivalent. However, things become less clear when we look at the more tricky cases. I now turn to the discussion of cases in which a subject's grounds have a causally deviant origin. I dub them cross-wiring cases. Very roughly, these are cases in which a subject experiences, say, an episodic memory, a visual experience, or a bodily sensation, but in which the experience causally derives from somebody else's past experience, perception, or proprioception, respectively. Let me begin by explaining the idea of cross-wiring cases in more detail.

Memory, Perception, and Proprioception

First, consider self-ascriptions based on episodic memory. A subject has an episodic memory of an event F, say, drinking a particularly good coffee at

Bonanza Coffee Heroes, and based on that remembering self-ascribes the event in question, here: 'I once drank a really good coffee at Bonanza Coffee Heroes'. Importantly, we are interested here not in the self-ascription of the memory experience ('I remember drinking a coffee at Bonanza …') but in the self-ascription of the remembered event. Now, in discussing the question whether such self-ascriptions are immune, Shoemaker introduced the idea of quasi-memory. In quasi-memory, a subject has an episodic memory of an event which she did not actually experience herself. Rather, she remembers an event from somebody else's life. The idea is that the memory of the event has been somehow surgically or otherwise inserted into the subject's brain. However, from the subject's point of view, quasi-remembering another person's past experiences is indistinguishable from remembering one's own past experiences.[24] Thus, it seems to be conceptually possible to self-ascribe an experience based on episodic memory, and be wrong about the question whose experience it is (cf. Shoemaker 1970).

Next, consider self-ascriptions of visual perceptions. There are in fact two different kind of cross-wiring scenarios concerning visual perception. To distinguish between the two, I will assume a difference between *actually seeing* an object, construed as a factive state involving the correct functioning of one's visual perceptual system, and merely *having a visual experience* as of an object, which is involved in seeing, but which is also present in hallucination.[25] The first-kind of perceptual cross-wiring scenario involves a split between the subject who is actually perceiving an object (in the sense of light falling into her eyes etc.) and the subject who is having the corresponding visual experience. Since this is the kind of scenario that is typically discussed in the literature on Immunity let me call it the *tradi-*

[24] Shoemaker defines quasi-memory as the encompassing class of *both* memories of one's own experiences *and* causally deviant memories (1970: 271). I prefer to construe the notion of quasi-memory as referring solely to the causally deviant memories.

[25] For the distinction between factive and non-factive perceptual states see also Langland-Hassan (forthcoming: 4f.).

tional variant.[26] Suppose, for instance, that you are standing on Oberbaumbrücke, but that your visual system is wired up to my brain in such a way that I am having a visual experience as of standing on Oberbaumbrücke. Based on that visual impression, I then judge that I am standing on Oberbaumbrücke. Cases like this have been discussed as possible counter-examples to the claim that external perception can be the basis of judgments about one's own relation to external objects. It may seem that in this case, I get it right that someone is standing on Oberbaumbrücke, but that I am wrong about it being me. It may seem, then, that my judgment is in error through misidentification.

Of course, this kind of scenario does not threaten the immunity of self-ascriptions of the visual impression itself. For, I really do have a visual impression as of standing on Oberbaumbrücke. It is just that the visual impression is the result of a deviant causal mechanism. Langland-Hassan (forthcoming) has suggested another type of cross-wiring scenario which is supposed to show that not even the self-ascription of the visual experience is immune to error through misidentification.[27] Let me dub this the *radical variant* of perceptual cross-wiring. To make sense of radical perceptual cross-wiring, we have to assume that *having* a visual impression is independent of *being aware* of that impression. Then we can make sense of the idea that the cross-wiring could pertain not to the causal basis of the visual impression, but could affect the link between the having of the impression and the awareness of the impression. For instance, if we understand awareness in terms of higher-order-thoughts or, very roughly, in terms of a

[26] See e.g. Smith 2006: 278, Chen 2009: 29ff. Not quite cross-wiring, but close enough, is Evans's case of undetectable headphones (1982: 184–189).

[27] To my knowledge, Langland-Hassan is the only writer to consider this type of cross-wiring (and to actually use it to challenge the corresponding version of Immunity): "my target is the more cautious (and more attractive) version of Introspective Immunity, relativized to non-factive mental states. […] So my goal will be to describe a plausible case where a person uses introspection to judge, e.g., 'I am having a visual experience as of an *x*,' and is mistaken for the sole reason that he has misidentified the subject of the experience." (forthcoming: 6)

monitoring mechanism, we can assume that my higher-order-thoughts or my monitoring mechanism is directed in a causally deviant way (telepathically or by means of some neural cross-wiring) not onto my own first-order mental states (in this case: onto my visual impressions) but onto yours. In such a case, I may self-ascribe having a visual experience which in fact not *I* am having, but which *you* are having. My self-ascription 'I am having a visual experience as of standing on Oberbaumbrücke' would be right about somebody's having that experience, but wrong about it being me.[28]

Finally, consider proprioception-based judgments about one's own body. Evans famously holds that these judgments are immune to error through misidentification in the same way as introspection-based self-ascriptions of mental states. For instance, my proprioception-based judgment that my legs are crossed is immune. A possible counterexample to this view is afforded again by the idea that your proprioceptive system could be wired up to my brain in such a way that I proprioceptively experience your legs' being crossed as mine (cf. Evans 1982: 221). My judgment 'my legs are crossed' would be right about somebody's legs being crossed, but it would be wrong about it being mine.

Let me make two general remarks. First, it is commonly assumed that in these scenarios, despite the strange connections between two cognitive systems, we can still speak of two distinct systems that have the same boundaries as they would have without the cross-wiring. The very description of the case as involving, for instance, *your* proprioception being hooked up to *my* brain, presupposes that, in spite of the cross-wiring, the two bodies belong to two distinct subjects, you and me. Although I am sympathetic to challenging this assumption, I will not discuss it further. To make the assumption more palatable, we can assume that the cross-wiring is not permanent, but just magically occurs for the briefest duration

[28] The case suggested by Langland-Hassan (forthcoming) is even more demanding. He argues that it is possible for me to *have* a phenomenal experience which occurs in *your* mind, so that my subsequent self-ascription of that experience is mistaken. (See § 6.2.)

necessary for the subject to make the judgment in question.[29] In what follows, I will assume that cross-wiring can be construed in a way that does not affect questions of personal identity.

Second, note that in all cases except for the radical variant of perceptual cross-wiring, what is self-ascribed is not the state that is first-personally experienced, but a state which, in normal conditions, is closely causally connected to that experience. In quasi-memory a past event is self-ascribed based on a present remembering, in traditional perceptual cross-wiring a physical relation to an external object is self-ascribed based on a visual experience, and in proprioceptive cross-wiring a bodily state is self-ascribed based on one's proprioceptive experience. The radical variant of visual cross-wiring is special in that it is the only case which involves the self-ascription of a mental state based on first-personal awareness of that particular state. It is also the most controversial case in that it assumes that introspective awareness of a visual experience and ownership of that experience can come apart. It seems that this case could not even get off the ground on views which hold that *being introspectively aware* of a visual experience simply means *having* that experience (in the sense of ownership).[30]

The idea that introspective awareness of a mental state can give rise to two different kinds of self-ascriptions has been put very well by Recanati (2007: 150–154). He distinguishes between the self-ascription of the occurrent conscious state and the self-ascription of standing in a certain relation to the world. In the case of visual perception, a visual experience can ground either the self-ascription of precisely that visual experience or it can ground the self-ascription of standing in a perceptual relation to the object in question. Traditional cases of cross-wiring show that self-ascriptions of

[29] For a similar move, see also Langland-Hassan (forthcoming: 19). Introducing more magic may not make the scenario more palatable all things considered, but it helps assuage this particular concern.

[30] But again, see Langland-Hassan (forthcoming) for a defense of radical cross-wiring (§ 6.2).

relations to other objects (such as standing in a visual perceptual relation to Oberbaumbrücke) are not or only *de facto* immune (to be explained shortly). They do not, without further argument, challenge the immunity of self-ascriptions of conscious occurrent mental states (such as the visual experience as of standing on Oberbaumbrücke). For, the cross-wired subject really does have a visual experience as of standing on Oberbaumbrücke and hence doesn't make a mistake in self-ascribing that experience. The difference between these two types of introspection-based self-ascriptions will be of some importance later on.

Now, with respect to all these cases, there is a lot of controversy surrounding the question whether the kind of mistake that is involved in cross-wiring is an error through misidentification or not. Roughly, Evans's view is that the judgments in question do not contain an identification component and that the mistaken judgments in cross-wiring cases are mistakes that resemble an illusion rather than an error through misidentification (cf. 1982: 184–188). Shoemaker, in contrast, accepts that in certain cross-wiring cases (more complicated ones, involving fission and fusion of persons) memory-based judgments can be in error through misidentification (cf. 1970). He seems to take this to imply that memory-based judgments in general are not identification-free, but presuppose that the subject who is remembering is identical to the subject who experienced the event in the past. In their discussion of cross-wiring, both Shoemaker and Evans hold on to the idea that immunity to error through misidentification is closely tied to identification-freedom. I will return to the question whether cross-wiring leads to error through misidentification in more detail below (§ 7.1).

Logical vs. de Facto Immunity

Let us, for the moment, assume that cross-wiring leads to error through misidentification. A common reply to this is to admit that these situations and mistakes are logically or conceptually possible, but to maintain that they do not actually occur. Employing a distinction between logical immunity and *de facto* immunity, it can then be maintained that the

judgments in question (memory-based self-ascriptions of past events, perception-based judgments about one's relation to external objects, and proprioception-based self-ascriptions of bodily states) still do enjoy at least *de facto* immunity.

The distinction between logical and *de facto* immunity has been introduced and applied by Shoemaker (1970). Yet, neither he nor other writers who appeal to the distinction ever offered a precise definition. The most intuitive way of understanding the distinction is this: a judgment *p* based on grounds *g* is logically immune to error through misidentification iff there is no logically possible world in which a judgment of type *p* based on *g* is in error through misidentification. A claim about *de facto* immunity, in contrast, can be understood as an immunity claim that applies not to all logically and conceptually possible worlds, but only to worlds in which certain contingent facts obtain. In other words, *de facto* immunity is immunity to error relative to certain conceptually and logically contingent facts. But what are those facts?

Coliva in her discussion of the distinction seems to assume that the *de facto* constraint simply restricts the scope of the claim to the actual world. She spells out Shoemaker's view that memory-based self-ascriptions are merely *de facto* immune as the view that

> memory-based self-ascriptions are immune to error through misidentification *in this world*, where memory information is stored in the usual way, [while] they would not be so in different metaphysically possible worlds where one is storing information deriving from someone else's past. (2006: 422f.)

That is to say, a judgment *p* based on grounds *g* enjoys *de facto* immunity iff in the actual world there is no judgment of type *p* based on *g* that is in error through misidentification. But this is unsatisfying. I take it that *de facto* immunity is meant to track a systematic epistemic feature of certain judgments, and not just a feature that certain judgments in this world happen to have. In a possible world in which, as a matter of pure chance, nobody every mistakes another person for themselves in the mirror, self-ascriptions based on mirror-reflections would be *de facto* immune. That

seems wrong. When it is claimed that proprioception-based judgments about one's bodily position are *de facto* immune, I do not take this as the claim that, as a matter of fact, no one ever proprioceptively mistakes another person's body for their own, but as the claim that, given the anatomy and technology (or lack thereof) that we have, it cannot happen that someone proprioceptively mistakes another person's body for their own.

Although it is quite common in the debate to simply say that this or that judgment is *de facto* immune, we should really, to make sense of this claim, specify the contingent facts with respect to which the judgment is *de facto* immune. To illustrate, suppose that proprioception-based judgments about one's bodily position are *de facto* immune, *given that cross-wiring of proprioception does not occur*. I take this to mean that in all possible worlds in which cross-wiring does not occur, no proprioception-based judgment about one's bodily position is in error through misidentification. This approach suggests that we can conceive of degrees of immunity that vary in modal strength. Logical immunity is immunity that holds in all logically possible worlds, conceptual immunity holds in all conceptually possible worlds, nomological immunity in all nomologically possible worlds, and different kinds of *de facto* immunity in different sets of worlds restricted by contingent facts that obtain in these worlds.

A similar restriction that closely resembles the *de facto* constraint is to assume the *normal functioning* of memory, perception, or proprioception. The Immunity Thesis, so restricted, would hold that the judgments in question are immune, when based on the normally functioning faculties. Cross-wiring cases would not amount to counterexamples of this claim since they do not involve normally functioning faculties.[31]

[31] An even more restrictive version of this approach is suggested by Evans. Given his notion of identification-freedom, he holds that judgments not only have to be based on these normally functioning faculties, but that the subject also has to believe that they are so based (cf. 1982: 189f.). We will later see how this move underwrites the epistemic nature of his approach.

Consider, for instance, the following argument by de Vignemont. De Vignemont claims that vision-based bodily self-ascriptions are immune when based on a "self-specific first-person visuo-spatial perspective" (2012: 239). If I judge, for instance, that my legs are crossed, based on looking down my torso onto my legs, de Vignemont would presumably hold that I cannot be wrong about whose legs are crossed. But there is an obvious counterexample to this claim, a case that resembles cross-wiring, but is not as metaphysically extravagant. In the so called body swap illusion, it is possible to induce in a subject fitted with virtual-reality goggles the belief that the body she is seeing when looking down is her own body, when really it is not (Petkova & Ehrsson 2008). Such a subject may judge that her legs are crossed, when really it is somebody else's legs that are crossed. Here is how de Vignemont replies to such a counterexample against her claim that "it is anatomically impossible that it could be another individual's body that I could see from this angle at this distance" (2012: 241, fn. 9):

> It may be possible only with some artificial tricks like in the body-swapping illusion. But this cannot constitute a counterexample given that it involves a deviant causal chain between the body that is seen and the visual experience (e.g. virtual reality system). (ibid.)

Of course we have to be careful not to put the restriction in question-begging terms.[32] Excluding virtual reality scenarios (i.e. excluding scenarios that exist in the actual world) as counterexamples may run danger of excluding counterexamples at will. I therefore think that the *de facto* constraint, although standing in need of precisification, is the more promising way of restricting immunity claims.

However, there is a further problem for the notion of *de facto* immunity. It is the question how it fits in with the idea that immunity can be construed

[32] For another example of this style of argument see Romdenh-Romluc 2013: 503f. She argues that, even if schizophrenic delusions involve error through misidentification, healthy subjects' judgments (i.e. judgments based on 'normal functioning') enjoy *de facto* immunity. This almost sounds as if she were saying that healthy subjects' judgments are immune because they normally get it right.

in terms of identification-freedom. If proprioception-based judgments are immune in this world, or in a properly restricted set of worlds, and hence identification-free in these worlds, how could it be that the same judgments, based on the same grounds, are in error through misidentification and *a fortiori* identification-dependent in other worlds (cf. Coliva 2006: 423)? Coliva answers the challenge as follows:

> Judgments [...] will be *de facto* immune to error through misidentification if *contingently* true identifications—identifications which may be false in other metaphysically possible worlds—feature as part of their *background presuppositions*.

But this answer has further repercussions for the idea that immunity is a matter of identification-freedom. She continues:

> [J]udgments that are (rather unhappily described as) *logically* immune to error through misidentification will be the ones that either have *no* identification component [...] or else, have an *a priori* true one—such as introspection-based self-ascriptions of occurrent psychological states, which at least some theorists construe as based on identification components like "I = the thinker of this thought." (Coliva 2006: 423f.)

This last amendment points to a very general issue that is brought up by cross-wiring, the question whether first-personally based self-ascriptions are based on identification components such as 'I am the thinker of this thought', 'I am the agent of this action', or 'I am the subject of this sensation'. The question pertains not only to the distinction between logical and *de facto* immunity, but also to the question whether the kind of mistakes we find in cases of cross-wiring should be classified as error through misidentification or not. Further, it is an interesting question also in the light of the pathological cases of alienation. We will come back to this point in § 7.1.

2.6 Different Approaches to Error through Misidentification

So far, I have talked as if there was one generally accepted notion of error through misidentification. It is time to note that really there are different views on the table. The fundamental question on which different approach-

es disagree is the question what it is exactly for a judgment to be in error through misidentification. I want to present three different takes on this question. The approach I have presented so far is what I want to call the *epistemic approach*. On this view, a judgment is in error through misidentification iff the judgment's justificational basis contains a false identification component. According to a different view, which I will dub the *representationalist approach*, a judgment is in error through misidentification iff the judgment's cognitive or psychological basis contains a false identification component. Finally, the view which I myself want to propose and which I label the *ontological approach* holds that a judgment is in error through misidentification iff the object from which the predication information derives differs from the object to which the property is ascribed.

The difference between the three approaches is most visible in the discussion of the particular cases to which they give diverging answers. It is not easy to characterize, in general terms, how and where the approaches differ. But very roughly, the difference can be characterized as pertaining to how the notion of the judgment's basis is construed. All approaches agree that the notion of error through misidentification has to be understood as a belief being in error relative to a certain basis. That is to say, a belief itself cannot be in error through misidentification, but rather it is the belief as founded on a particular basis which can be in error through misidentification. However, few authors have made precise what they mean by that basis. Depending on how one construes the basis at this point, one can end up with very different notions of error through misidentification.

The idea that different writers approach the phenomenon of immunity from substantially different angles, and that there are thus substantially different notions of immunity involved, doesn't receive a lot of attention in the debate. Although scattered remarks here and there suggest that participants are aware of there being fundamentally different approaches to immunity, I know of no attempt to systematically classify and compare these different views. In this section, I attempt to provide such a systematic classification myself and I discuss a number of cases in which the approaches come to diverging results.

Epistemic Approach

Clearly, the epistemic approach is the most dominant approach in the debate. Many of the issues introduced above already presuppose an epistemic approach to immunity to error through misidentification. The peculiarities of the epistemic approach will come to light shortly in contrasting it to the other two approaches.

The epistemic approach construes a judgment's basis as the subject's epistemic grounds or rational reasons. Whether a judgment involves an identification or not depends on whether the belief is epistemically based on an identification component. What this means exactly is contested among the proponents of this approach. One thing that is for sure, and that distinguishes the epistemic approach from the representationalist and ontological approach, is that a judgment's being identification-dependent in the epistemic sense does not depend on the actual cognitive or psychological genesis of the judgment or on the actual origins of the information on which the predication is based.

Let me mention one way of spelling out the idea of identification-dependence which brings out the nature of the epistemic approach most clearly. According to Pryor, a judgment is in error through *de re* misidentification, iff "[t]he subject's justification for believing this singular proposition [*x* is F] rests on his justification for [falsely] believing, of some *y*, and of *x*, that *y* is F and that *y* is identical to *x*." (1999: 274) That is to say, whether a judgment is in error through misidentification, on this view, turns on the question whether my *justification* for that judgment is based on *justification* for a mistaken identity belief.

Ontological Approach

I want to propose a somewhat new approach to immunity, which I label the ontological approach. The ontological approach construes a judgment's basis as the actual states of affairs that are responsible for the judgment in question. This approach does without the epistemic notion of identification and it does not assume a connection between identification-freedom and

immunity. It holds that a judgment is in error through misidentification iff the source object differs from the target object.

This is the approach that captures best what I take to be a core feature of error through misidentification: that one is getting it right about somebody's being F, but getting it wrong about who it is that is F. I know of only one other writer who professes this approach, viz. Simon Prosser who observes that "a judgment is immune to error through misidentification in just those cases where the source object and the target object cannot differ." (2012: 161f.)

To define ontological error through misidentification, we can say that a judgment of the form '*a* is F' is in error through misidentification if and only if the source object (i.e. the object from which the predication information derives) is different from the target object (i.e. the object to which the predicate is applied). To illustrate, my judgment that Forrest is running when based on my visual perception of someone running is in error through misidentification if and only if the person I see running (the source of the predication information) is not Forrest (the object to which I apply the predication).

On this approach, immunity to error through misidentification then simply is the characteristic that certain beliefs that are based on certain sources of information do not ever exhibit this kind of divergence. To illustrate, consider my judgment 'I want a coffee' based on my experiencing a desire for coffee. For this judgment to be immune to error through misidentification relative to the question who wants a coffee, on the ontological approach, just means that it is not possible that the person to whom the desire is attributed (the target) is distinct from the person who actually experiences the desire (the source).

This approach differs substantially from the epistemic approach. On the ontological view, error through misidentification does not have to do with what the subject is epistemically entitled to believe, or what the subject should rationally believe. Rather, error through misidentification is defined in terms of whether the property is ascribed to the right subject, where the

right subject is the subject whose actually instantiating the property in question is responsible for the predication. In most cases, the ontological approach will give the same verdicts as the epistemic approach. The difference can be revealed in looking at cases that are either epistemically or ontologically unusual. For instance, in cases of cross-wiring the ontological approach yields a clear verdict, namely that the judgments in question are in error through misidentification. For instance, my cross-wired judgment that I am standing on Oberbaumbrücke involves a divergence between source and target: it is *your* standing on Oberbaumbrücke that leads (via cross-wiring) to my judgment that *I* am standing on Oberbaumbrücke. Although some versions of the epistemic approach also construe cross-wiring as involving error through misidentification, they reach that result in a completely different way, namely via the assumption that there is an identification component in the justificational structure. A lot more will be said on the difference between the epistemic and the ontological approach later (§§ 6f.).[33]

Representationalist Approach

A few authors have approached the phenomenon of immunity to error through misidentification from a cognitive or psychological point of view. For the sake of completeness, I will briefly present this approach. However, it will not play any role in the rest of this thesis.

The representationalist approach construes a judgment's basis as that which is actually cognitively going on within the subject's head. Whether a judgment involves an identification or not depends on whether the object (or the objects, in case of misidentification) is represented by means of two distinct representations (or two different modes of presentation) in the cognitive processes that lead up to the judgment. Thus, a judgment is in error through misidentification iff the cognitive processes that lead to the

[33] Although I am contrasting the ontological approach with the epistemic approach, note that error through misidentification is still a genuinely epistemic phenomenon, also on the ontological approach.

judgment involve two distinct representations which are mistakenly taken to refer to the same object.

The most explicitly representational approach is presented by Gottfried Vosgerau (2009a, 2009b). Vosgerau's overall project is a representationalist theory of self-consciousness. In order to discern different levels of self-consciousness, he proposes a classification of different types of I-thoughts in terms of different types or errors that these thoughts are liable to. I-thoughts which are immune to error through misidentification, so the idea, constitute the most fundamental levels of self-consciousness. In his classification, Vosgerau distinguishes between error through misattribution ("Fehler durch Falschzuschreibung") and error through misidentification ("Irrtum durch Fehlidentifikation") (cf. 2009b: § 3.1). While a misattribution involves merely the attribution of a predicate to the wrong object, a misidentification involves a mistaken identification within the cognitive genesis of the representation. Vosgerau defines error through misidentification as follows. "Irrtum durch Fehlidentifikation [kann nur entstehen], wenn bei der Bildung der Repräsentation eine Identifikation von zwei Objekten involviert ist"[34] (ibid.: 113).

This view yields a notion of error through misidentification which differs substantially from the dominant epistemic notion. For instance, most judgments that are in error through which-object misidentification in Pryor's sense would not count as in error through misidentification in Vosgerau's sense, but as in error through misattribution. Witness Vosgerau's criticism of Pryor's notion of which-object misidentification:

> Mein Hauptkritikpunkt ist allerdings, dass hierbei eine falsche Sicht der Entstehung der Repräsentation zu Grunde liegt: Wenn wir eine Repräsentation der Form F(a) bilden (dem Objekt a kommt die Eigenschaft F zu), dann geschieht dies oft (ausgenommen Spiegel-Fälle und dergleichen) nicht durch einen Schluss der Form: es gibt ein Objekt mit der Eigenschaft F, dieses Objekt ist identisch mit a, also

[34] "Error through misidentification occurs only if an identification of two objects is involved in the genesis of the representation." (My translation.)

> F(a); vielmehr nehmen wir eine Eigenschaft F wahr, die wir direkt und ohne eine Identifikation von zwei Objekten dem Objekt a zuschreiben. In diesen Fällen ist also keine Identifikation involviert, und daher kein Irrtum durch Fehlidentifikation (in keinem Sinne) möglich. (ibid.: 111)[35]

But Pryor need not deny Vosgerau's claim about the genesis of thoughts. Rather, he can point out that his (Pryor's) notion of error through misidentification simply does not have any implications regarding the genesis of the representation in question. Pryor's claim about cases of which-object misidentification is not that subjects arrive at the representation F(*a*) via an identification process, but that the justification of F(*a*) is based on or involves justification for a singling out identification. In some sense, then, Vosgerau's critique misses its target.[36] Be that as it may, my main point was to illustrate how Vosgerau approaches the whole issue from a completely different angle, one that focuses on the cognitive genesis of the representations rather than on the rational justification of the representations.

Another author whom I take to be a proponent of the representationalist view is François Recanati. Since, unfortunately, he is not very clear about his position in print, let me illustrate his view by recounting an argument he proposed in a presentation.[37] Recanati argued to the effect that judgments

---

[35] "My main critique is that this is based on a mistaken view about the genesis of the representation. Normally (apart from mirror-cases and the like), generating a representation of the form F(*a*) (object *a* has the property F) does not involve an inference of the form: there is some object that is F, this object is identical with *a*, therefore F(*a*). Rather, we perceive a property F which we attribute to the object *a* directly and without the identification of two objects. Hence, these cases do not involve an identification and error through misidentification is not possible (in any sense)." (My translation.)

[36] One could further discuss which of the two notions really deserves the label *misidentification*, but that is a different question which amounts to nothing more than a terminological dispute.

[37] During the workshop "Immunity to Error through Misidentification and Essential Indexicality", September 21-22 2012, Konstanz.

based on close visual recognition are immune to error through misidentification because they are identification-free. To illustrate the notion of identification-freedom which underlies this claim, he offers the following example: upon entering the kitchen, I see my spouse sitting there and thus form the belief that my spouse is in the kitchen. According to Recanati, that thought is immune to error through misidentification of the spouse. But why is that? After all, couldn't it be that I am seeing a look-alike of my spouse? And wouldn't I have, in that case, misidentified the person I see as my spouse? Yes, Recanati agrees, but still I am representing the person solely *as my spouse* and never *as that person* (whom I see). I do not identify the person in the sense that my belief does not involve two distinct representations of my spouse. And in this sense of identification, my judgment is immune to error through misidentification.

Although Recanati does not explicitly present this view in print, the idea can be gleaned from some passages in his *Mental Files* (2012b). Witness the following claim about self-ascriptions that are not immune:

> [W]hen some information about ourselves is gained from outside, it goes into the SELF file only in virtue of a judgment of identity. [Footnote 5: Or, in mental file talk: the information goes into the self file in virtue of a 'link' between that file and some other file.] (2012b: 66)

The idea behind this claim is that a judgment is based on an identity judgment iff there is a linking process. Further, it is implied that if a judgment is based on an identity judgment then it is not immune. Later, Recanati explicitly states that in immediate recognition there is no linking process and no identity judgment:

> In immediate recognition, there is no linking of files, as there is a single file (based on a composite relation). To be sure, it is *presupposed* that the object which the subject stands in the demonstrative relation to is the same object he has been acquainted with before and remembers; but the subject does not *judge* that the identity holds. Rather, the identity is established, at the sub-personal level, through the subject's non-conceptual capacity to recognize the object and track it over time. (ibid.: 87)

This implies, even though Recanati does not explicitly draw the conclusion, that immediate recognition (in contrast to what he calls 'slow recognition') delivers immune judgments. As far as I can see, Recanati's main point is that, in immediate recognition, on the personal (conscious?) level we find no identity judgment. For, Recanati acknowledges that the identity of the person in question is presupposed. But what does that mean? Recanati alludes to the sub-personal non-conceptual capacity to recognize the object and track it over time. But there seem to be different kinds of capacities involved in the case. The capacity to recognize an object as an object and track that object over time is different from the capacity to recognize, or, re-identify an object as a previously known object. Certainly, immediate recognition involves re-identification, even if on the sub-personal non-conceptual level. But this is precisely why it is usually taken to be vulnerable to error through misidentification. Recanati's approach construes error through misidentification in an importantly different way from the epistemic approach. His point that immediate recognition does not involve the linking of files can be translated into language of representation as saying that judgments based on immediate recognition do not involve two distinct representations.

Visual recognition normally serves as a canonical illustration of error through misidentification and liability to error through misidentification. Recanati's claim that judgments based on immediate visual recognition are immune reveals that Recanati construes error through misidentification in a radically different way. Like Vosgerau, Recanati is interested in the genesis of the representation and in particular in the question whether it involves two distinct representations. Proponents of the epistemic approach, in contrast, are interested in the justificational basis of the judgment. They would presumably reply to the above argument: what Recanati calls a presupposition (that the object referred to in thought is one and the same as the object acquainted with before) is exactly what we mean by identification. It is in virtue of the object's characteristics (appearance, location, etc.) that it is perceived by the subject as being this particular object. The fact that the identity is established at the subpersonal level does

not make the judgment identification-free in an epistemic sense. There may not be an identification on the personal level, but the judgment certainly depends, as it is often put, on the recognition of a particular person in the sense that criteria of recognition are brought to bear.

To sum up, the representationalist notion of error through misidentification differs substantially from the epistemic notion of error through misidentification in several respects. A judgment can be identification-free in the representational sense (its genesis does not involve two distinct representations) but identification-dependent in the epistemic sense. The representationalist approach also cuts across the idea that a judgment is immune iff any ground for doubting that *a* is F is *ipso facto* ground for doubting that anyone is F. In judgments that are based on immediate recognition, for instance, one can certainly challenge the subject component separately from the predication component. If Recanati were told that this person is not his spouse, he would still have reasons to believe that *someone* is sitting in the kitchen. Further, the cases of cross-wiring do not affect questions of representationalist misidentification at all, as the cases are usually construed in a way that does not make a difference to the question whether the genesis of the belief involves two distinct representations or not.

2.7 Summary

I have introduced a number of issues surrounding immunity that will at some point or other be important to my discussion. The most fundamental issue is my distinction between three different approaches to the phenomenon of immunity to error through misidentification. While I do not further discuss the representationalist approach, the differences between the ontological and the epistemic approach will not matter in the following chapter, in which I defend the Immunity Thesis against the counterexamples from pathological alienation. In later chapters (§§ 5–7), the distinction will be of central importance and I will have more to say on both approaches. For now, we shall work with a simple definition of error through misidentification: a judgment is in error through misidentification iff the judgment is right about someone's being F, but wrong about who it is that is

F.[38] For the sake of simplicity, I will often use the term 'grounds of a judgment' very broadly, i.e. open to interpretation according to the epistemic or ontological approach.

Another point that will play a central role in the following chapter is the idea that the scope of the Immunity Thesis is restricted (among others) by what I called the self-ascription constraint and the introspection constraint. That is to say, the thesis claims immunity only for those judgments that are self-ascriptions and based on first-personal ways of gaining knowledge. Any putative counterexamples that breach these restrictions fail to challenge Immunity simply in virtue of not falling within its scope.

[38] Again, this definition ignores cases of simultaneous mispredication and misidentification, but that will not matter for the following discussion.

3. Pathologies of Alienation[39]

I now turn to the discussion of the putative counterexamples, i.e. different cases of pathological alienation[40]. The dialectics of the following discussion will be this: the Immunity Thesis is the claim that certain judgments are immune to a certain kind of error. For any case to refute the thesis, it has to involve the right kind of judgment, i.e. the kind of judgment that is claimed to be immune, and it has to involve the right kind of error, i.e. error through misidentification. As I laid out above, the kind of judgments that are claimed to be immune are introspection-based self-ascriptions of mental states. In this chapter, I will ask for each of the putative counterexamples whether the belief in question is a self-ascription (i.e. whether it satisfies the self-ascription constraint), whether it is based on first-personal awareness (i.e. whether it satisfies the introspection constraint), and whether it is in error through misidentification. I argue that the pathological cases either do not fall within the scope of Immunity or do not involve error through misidentification and that hence none of the cases undermines Immunity.

3.1 Thought Insertion

The pathological phenomenon that has first been raised as an empirical counterexample against the Immunity Thesis is the phenomenon of thought insertion (Campbell 1999b), a first-rank symptom of schizophrenia (Schneider 1959). In thought insertion, subjects claim that certain thoughts they are experiencing are not their own thoughts. Typically they also claim that the thought in question is in fact somebody else's thought. Here is one widely cited report:

---

[39] Parts of this chapters appear also in my paper "Immunity and Self-Awareness" which is under review at *Philosophers' Imprint*.

[40] Note that 'pathology of alienation' or 'pathological alienation' is not a psychiatric term. I am using it here to loosely denote pathological phenomena in which a subject feels alienated in some way or other.

> Thoughts are put into my mind, like 'Kill God'. It's just like my mind working, but it isn't. They come from this chap, Chris. They're his thoughts. (Frith 1992: 66)

Campbell takes this to be an (at least) prima facie threat to Immunity.

> A patient who supposes that thoughts have been inserted into his mind by someone else is right about which thoughts they are, but wrong about whose thoughts they are. So thought insertion seems to be a counterexample to the thesis that present-tense introspectively based reports of psychological state cannot involve errors of identification. (1999b: 609f.)

The discussion of thought insertion requires a few conceptual clarifications. First, I distinguish two judgments typically found in thought insertion: the disowning judgment ('this is not my thought') and the external attribution ('this is Chris's thought') (cf. Stephens & Graham 2000: 152). Secondly, I appeal to an established distinction of what it means for a thought to be one's own. On the one hand, a thought can be my own in the sense that I am the one experiencing the thought, that the thought is introspectively available to me or that the thought appears in my stream of consciousness. Judging a thought to be one's own in this sense is what I call the *ownership* attribution. On the other hand, a thought can be my own in the sense that I am the producer, the active thinker, or the causal origin of the thought. Judging a thought to be one's own in this sense is what I call the *authorship* attribution.[41]

Basically, the idea is that thought-ascriptions of the form 'this is (not) my thought' are ambiguous; a thought can be mine, or yours, in at least two different ways. According to the standard interpretation of thought insertion, what subjects deny, when they disown a thought, is authorship but not ownership. Following this approach, we can disambiguate the judgments 'this is not my thought' and 'this is Chris's thought'. In what

[41] There is no universally accepted terminology, but roughly the same distinction can be found in Campbell (1999b), Gallagher (2000), Stephens & Graham (2000), Sousa & Swiney (2011), Vosgerau & Voss (forthcoming).

follows I am going to take them to express the beliefs 'I am not the author of this thought' and 'Chris is the author of this thought', respectively. A common response to the alleged counterexample is then to concede that self-ascriptions of authorship are not immune to error through misidentification, but to insist that self-ascriptions of ownership are (see e.g. Gallagher 2000, Vosgerau & Voss (forthcoming)). As we will see shortly, this concession is in fact too quick.

Now, of the two judgments involved in thought insertion, the disowning belief and the external attribution, Campbell never made explicit which one he takes to refute Immunity, so we will look at both. Note also that the whole debate is premised on the assumption that the subject is in fact the author of the inserted thought, which implies that both beliefs are false.

The external attribution ('Chris is the author of this thought') is not very convincing as a counterexample. While it is in error through misidentification, it fares poorly with respect to both scope restrictions. Obviously, the external attribution is not a *de se* self-ascription.[42] Regarding the introspection requirement, the question is what the basis is for ascribing the thought in question to a particular other person, say to Chris. Some find it conceivable that a thought can be first-personally experienced as some particular other person's thought (see e.g. Sollberger (forthcoming)). However, the standard view is that the external attribution is at least in part a form of rationalization, confabulation, or inference. What may be experienced first-personally, on this view, is that the thought is strange in some way or other, but not as being some particular other person's thought. If this is correct, the external attribution is not based solely on first-personal awareness of the thought in question.

Let us then turn to the disowning claim ('I am not the author of this thought'), which fares more promising with respect to the scope criteria. Does the disowning claim satisfy the self-ascription constraint? It may seem that it does not. After all, a property is denied rather than ascribed to

[42] Coliva (2002a: 30) first noted this point. See also Langland-Hassan (forthcoming: 7).

oneself. However, I suggest that the self-ascription constraint has to be understood broadly as restricting the thesis to self-concerning beliefs generally, be they positively self-attributing or negatively self-denying. I shall henceforth speak of *positive self-ascriptions*, which have the form 'I am F', and *negative self-ascriptions*, which have the form 'I am not-F'. My assumption is that the self-ascription constraint does not exclude negative self-ascriptions from the scope of the thesis, but mainly excludes external attributions. On this view, the disowning belief does satisfy the self-ascription constraint. (I defend this assumption in § 5.1.)

Does the disowning belief also satisfy the introspection constraint? The question is, in other words, whether the belief is based solely on first-personal awareness of the inserted thought. The answer depends on a controversial empirical issue that I do not have space to fully address here. The contested question is what the primary thought experience at the core of thought insertion is like. There are two views on the matter.[43] According to the *endorsement approach*, the thought is already represented in the primary thought experience as not one's own. The judgment 'I am not the author of this thought' simply expresses this experience. If this is correct, the judgment is based on introspective awareness of the primary thought experience and hence satisfies the introspection criterion. In contrast, according to the *explanationist approach*, the primary thought experience does not represent the thought as not one's own (although it may represent the thought as strange in other respects). The explanationist approach holds that the disowning is an attempt to explain or rationalize the occurrence of the thought (e.g. because it has an unwelcome or unfamiliar content); it is not based directly on the alien nature of the thought in question, but rather on confabulation or some kind of inference. If this is correct, it is debatable whether the disowning claim satisfies the introspection criterion.

[43] The following distinction has been introduced by Pacherie et al. (2006). Sollberger (forthcoming) applies it to thought insertion.

Let me explore this point a bit further. I do think that the main problem in determining whether the disowning belief satisfies the introspection constraint is that it depends on the controversial empirical issue regarding the genesis of the delusional belief. I want to stay neutral with respect to the two approaches. However, part of the problem may also be that it is not entirely clear what it takes exactly to satisfy the introspection constraint. The notion of introspection is notoriously hard to define and I do not want to take a stand on a particular view on introspection. But from a philosophical point of view one can ask, more generally, what the introspection criterion should be thought to require in the light of cases such as thought insertion.

The idea roughly was that for the delusional beliefs in question to be based on first-personal awareness, the alien nature (either not-ownership or other-ownership) that is attributed to the primary thought already has to be part of the primary thought experience. I suggested that, in contrast, the attribution should not count as based on introspection if it was guided by secondary delusional beliefs. But couldn't it be completely normal that secondary beliefs guide expectations regarding the attribution of authorship? What is unusual, in the pathological cases, is the content of the secondary beliefs. The mere fact that the attribution of authorship is influenced by such beliefs may not be strange at all. Behind this idea is the empirical insight that beliefs about agency and authorship are generally guided not purely by first-personal awareness of the state in question but by background beliefs and other factors (with respect to agency, see for example Wegner & Sparrow 2004). The fact that in thought insertion background beliefs play a role in the authorship ascription (even if these beliefs are strange and delusional) does not make the source of disowning beliefs problematic. Generally, this means that if one defends Immunity against the pathological type of counterexample by appealing to the introspection criterion one might equally be excluding from the scope of the thesis a number of regular self-ascriptions that have traditionally been held to be immune to error through misidentification. Certainly, that is a possible way to go. But it would mean defending a much more restricted

Immunity Thesis than has been traditionally held. I will not explore this route further, but rather grant to critics that the disowning claim satisfies the introspection constraint.

Granting that the disowning claim ('I am not the author of this thought') actually satisfies the two scope criteria, is it a successful counterexample? The final question is whether it involves an error through misidentification. As a matter of fact, it does not. For the case to be an error through misidentification, according to the simple definition, it would have to be the case that the subject is right about somebody's not being the author of the thought and wrong only in believing *of himself* that he is not the author. But there isn't anybody else involved whose property the subject could have misattributed to himself. He didn't mistake somebody else's not being the author for his own. Rather, the mistake is a mispredication, not a misidentification. The subject experiences the thought of which he actually is the author as alien and falsely believes of himself that he is not the author. Hence, the disowning belief fails as a counterexample because it does not involve an error through misidentification.

If this result seems surprising, it may be so because disowning and external attribution have not been kept properly apart. In some sense, it is true that thought insertion involves both a misidentification and a self-ascription: The external attribution involves a misidentification and the disowning involves a self-ascription, but there is not one single judgment that is both a self-ascription and involves a misidentification. Later, I discuss three ways in which critics of Immunity can attempt to save the case. These rejoinders bring out in different ways the idea that thought insertion challenges Immunity by showing that authorship-ascriptions are identification-dependent (see § 5).

3.2 Anarchic Hand Syndrome

Anthony Marcel claims that the anarchic hand syndrome refutes the Immunity Thesis regarding bodily agency (cf. 2003: 80f.). In anarchic hand syndrome, one hand of the patient performs "unintended but complex,

well-executed, goal-directed actions" which often compete with what the patient actually intends to do (ibid.: 77). The hand itself is not (necessarily) disowned by the patient; i.e. the patient (usually) acknowledges that it is his or her own hand. The crucial aspect is that the actions which this hand performs are not intended by the subject and appear to the subject as if not done by him. Marcel suggests that these subjects make an identification mistake in denying agency. The candidate counterexample therefore is some belief such as 'I am not the agent of this action', 'This is not my action', or 'I am not doing this action'.

Let me set one thing straight from the beginning. Patients of anarchic hand syndrome do not actually *believe* not to be the agent, but rather report that the action *feels as if* it is not theirs. As Marcel himself remarks, while one patient "said that he was not doing the anarchic actions. He quickly followed this by adding that 'of course I know that I am doing it. It just doesn't feel like me.'" (ibid.: 79) Marcel adds that patients are "often clear that their experience of the action as disowned is a 'seeming'." (ibid.) So perhaps there is no disowning *belief* involved in anarchic hand syndrome at all. Since Immunity is a thesis about beliefs, I will nonetheless discuss a potential disowning *belief* for now. I will later come back to the idea that it is the feeling of non-agency which may be thought to challenge Immunity.

I have two objections against Marcel's case. Before I present them, let me mention, just to put aside, a critique raised by Christopher Peacocke (2003: 109). Peacocke holds that the movements of the anarchic hand are not actions at all and that therefore there is no mistake involved in the subject's denial of agency. Peacocke's guiding assumption seems to be that for the movement to be a φ-ing it would have to essentially involve a trying to φ, which it does not. Since the following arguments can do without that assumption, I will grant to Marcel for the sake of argument that the hand's movement is an action.

The first objection is that the judgment is not based solely on first-personal awareness. Let us assume that there is something we may call first-personal action awareness. Plausibly, this kind of action-awareness would involve

introspective awareness of the action intention. However, as Marcel himself stresses, in the anarchic hand syndrome "awareness *from the inside* of relevant intention, effort, and will are lacking" (2003: 79). In what sense, then, can it be said that the subjects are first-personally aware of the anarchic action? Subjects may have proprioceptive (i.e. first-personal) awareness of the hand's *movements* (ibid.: 81). But that does not suffice for first-personal *action* awareness. For, it is the goal-directedness of the hand's behavior that makes it an action. But clearly, subjects are not aware of this goal-directness first-personally, via awareness of the intention, but are aware of the goal-directedness in a third-personal kind of way. For instance, when they see (and feel) the anarchic hand unbutton the shirt which their other hand is trying to button up, they perceive the hand's movement as an unbuttoning. But they do so in a third-personal way, just as a bystander would perceive the movement as an unbuttoning. Crucially, they are not first-personally aware of an intention to unbutton the shirt.

Critics of Immunity may want to rejoin that patients do after all have first-personal awareness of the hand's movement (somatic proprioception). And since the hand's movement *is* the hand's action, they do have first-personal awareness of the hand's action in some sense. But the fact that they perceive the hand's movement proprioceptively does not make them first-personally aware of the action *qua action*. Subjects do not disown the hand's movement *qua* movement, but the hand's action *qua* action. And when it comes to the immunity of action ascriptions, the introspection-criterion must be understood to restrict the scope to those beliefs that are based on first-personal action-awareness, that is first-personal awareness of actions *qua* actions. Hence, for the case to challenge Immunity, it would have to involve first-personal action-awareness. I argued that it does not.[44]

My second and more fundamental objection is, again, the point that the denial of agency does not involve a misidentification. The subject does not attribute the right property to the wrong person, but rather attributes the wrong property to the right person. Subjects suffering from anarchic hand

[44] Thanks to Gottfried Vosgerau for discussion of this point.

syndrome do not go wrong in figuring out who it is that lacks agency, rather they go wrong in the question whether they themselves are agentically responsible for the movement or not. Again, I will later discuss whether this objection can be resisted (see § 5). Before I move on, I briefly digress on two more general issues which are best discussed in the light of anarchic hand syndrome.

Absolute Immunity vs. Circumstantial Immunity

Related to the first objection, the objection that awareness of the action intention is an essential part of first-personal action-awareness, we can ask the following question. Was the Immunity Thesis ever meant at all to hold for the ascriptions of actions themselves, or was it perhaps just meant for the ascriptions of action *intentions*? The example given by Wittgenstein, and cited approvingly by Shoemaker, suggests the latter: "I *try* to lift my arm" (1958: 66; my emphasis, Wittgenstein's emphasis omitted). Obviously, the anarchic hand syndrome has no bearing at all on the claim that introspection-based ascriptions of action-*intentions* are immune to error through misidentification.

However, it cannot be denied that some authors have claimed immunity not just for the ascription of action intentions, but for the ascription of the action itself. Here is Shoemaker's example: "whereas the statement 'My arm is moving' is subject to error through misidentification, the statement 'I am waving my arm' is not" (1968: 557). Of course, Shoemaker must be taken to mean that the action ascription is immune only if it is based on first-personal action awareness, which, as I have argued, is not the case in anarchic hand syndrome.

Still, there is something in the quote of Shoemaker's example that makes anarchic hand syndrome an interesting case to discuss.[45] Why does Shoemaker deny immunity to the self-ascription of movements? After all,

[45] I do not think that Marcel, in bringing up the anarchic hand as a counterexample, had anything in mind even closely related to what follows. So, this is my own attempt to get as much out of the case as possible.

when I judge that my arm is moving based on first-personal (i.e. proprioceptive) awareness of my arm's movement then, it seems, I couldn't go wrong with respect to the question whose arm is moving. Of course, we can imagine cross-wiring cases in which the proprioceptive information I receive is coming from another person's arm. But Shoemaker doesn't have this kind of case in mind when he denies immunity to movement ascriptions. Following the quoted passage, Shoemaker discusses how the bodily self-ascription 'I am facing a table' is merely *circumstantially* immune in being based on the judgment 'I see a table', which is *absolutely* immune (for the distinction between absolute and circumstantial immunity see § 2.2). Similarly, he should say of the judgment 'My arm is moving' that it is circumstantially immune when based on the judgment 'I feel my arm moving', which is absolutely immune. The belief 'my arm is moving' is not absolutely immune because it could also be based on visual perception of one's arm, a case in which the judgment would not be immune, for one may mistake somebody else's arm as one's own.

So the reason why, in the initial quote, Shoemaker denies immunity to the movement ascription is that Shoemaker is talking in that passage about *absolute immunity*. His claim is that agency-ascriptions are absolutely immune, whereas movement-ascriptions are not. At this point, anarchic hand syndrome suddenly becomes an interesting case because it shows that, contrary to what Shoemaker claims, agency ascriptions are not absolutely immune. For them to be absolutely immune, it would have to be the case that there is no way of self-ascribing actions that is vulnerable to error through misidentification. In other words, it would have to be the case that actions can only be self-ascribed based on first-personal awareness. But anarchic hand syndrome shows that it is possible to be aware of one's own actions in a third-personal way (without also being first-personally aware of the action). And judgments based on third-personal awareness are generally vulnerable to error through misidentification.

Doxastic vs. Phenomenal Self-Ascriptions

Another general issue that can be raised in the discussion of anarchic hand syndrome is the question which kinds of states can have the property of being immune at all. So far I have been focusing on delusional beliefs as possible counterexamples. But couldn't it be that things come out differently if instead of looking at *judgments* of agency we consider the corresponding *feelings* of agency as threats to the Immunity Thesis?[46] In the case of anarchic hand syndrome, should we look at the subject's *feeling* that he is not the agent, rather than at the hypothesized disowning belief?

The supposition that Marcel may have the feeling of agency in mind as a counterexample, rather than a disowning belief, is underwritten by the following passage:

> By 'mistaken' I mean mistaken about whose [action] it *seems* phenomenally to be, rather than mistaken in logical or rational reflection about whose it *must* be. (Marcel 2003: 80)

The obvious problem with this suggestion is that the Immunity Thesis is clearly a thesis about beliefs, not about feelings or phenomenology. However, it would be interesting to explore the conjecture that the thesis can be extended to certain kinds of phenomenal states. I don't have the space to fully discuss this idea, but here is a rough illustration. What could it mean for feelings to be immune to error through misidentification? Suppose I experience a feeling as of me being hungry. While it is possible that the feeling of hunger misrepresents what state I am actually in (I might actually be tired rather than hungry), it is not possible for the feeling to veridically represent that someone is hungry, but to misrepresent who it is that is hungry. In that sense, it may be said that certain feelings, too, are immune to error through misidentification. The idea is not that my second-order belief about the feeling ('I feel hungry') is immune to error through misidentification; the idea is that the feeling itself cannot misrepresent who is in the state in question.

---

[46] For the distinction between judgment and feeling of agency see Synofzik et al. 2008.

Note that this kind of extension is built on the idea that some phenomenal states can be construed as carrying representational content the function of which is to represent the way the world is. Consider again my feeling of hunger. My feeling of hunger, I suggest, can be construed as representing me to be in a certain bodily state (something having to do for instance with an empty stomach and a low blood-sugar level). If actually I am not in a state of physical hunger, but still feel hungry, it can be said that my feeling misrepresents the state I am in. In that sense, we can say that the feeling 'gets something right' (the subject question) and gets something else wrong (the property question).

Let me now apply this idea to anarchic hand syndrome. If we construe the phenomenal experience as representing the world to be a certain way this experience comes out as having the same content as the delusional belief that I have discussed above, viz. the content 'I am not the agent'. Now, the reason why this whole move doesn't get us anywhere is that the same problem that has been raised against disowning *beliefs* resurfaces for the *feeling* of non-agency. Just like the *judgment* of non-agency does not involve a misidentification, the *feeling* of non-agency does not involve a misidentification either. To wit, the feeling of non-agency (i.e. the feeling with the content 'I am not the agent') does not ascribe the right property to the wrong subject, but rather ascribes the wrong property to the right subject. The general point is this: Even if we allow the phenomenal aspects of alienation to fall within the scope of Immunity, they do not add anything to the debate. For, we will always be dealing with the representational content of the experience, and that content is the same as the representational content of the corresponding judgments.

To complete the discussion of doxastic vs. phenomenal self-ascriptions, let me briefly turn to a question that has received a lot of interest recently, the question whether the mental states we tend to call 'delusional beliefs' really are beliefs in the proper sense of the term. Critics have argued that delusions cannot truly be regarded as beliefs since they are not properly integrated into the subject's belief system, but rather lead a life of their own, so to speak. For instance, a patient's delusional belief that he is Napoleon

will typically not interfere with the patient's tendency to follow the psychiatric ward's directions. Based on this phenomenon, also known as double book-keeping, critics have challenged the hitherto commonplace assumption that in delusions we are dealing with delusional beliefs, properly so called. Rather, they argue, these delusional states should be described, for instance, as seemings. It may then be objected, on the basis of this idea, that only beliefs proper fall within the scope of Immunity and that delusions fail as counterexamples simply because they are not beliefs. However, I do not find such a response very persuasive, particularly not in the light of the idea that perhaps the thesis can even be extended to more clearly phenomenal states. Whatever the proper classification of delusions is, they are close enough to being beliefs that we should consider them as the kinds of states that principally fall within the scope of Immunity.

Summary

I argued that anarchic hand syndrome refutes Shoemaker's claim that self-ascriptions of actions are *absolutely* immune to error through misidentification. However, I know of no other authors who defend this strong claim anyhow. The much more plausible immunity claim, when it comes to actions, is the claim that self-ascriptions of actions are immune *when based on first-personal action awareness*. That claim, I argued, is not challenged by anarchic hand syndrome since subjects are not aware of the action intention. In a second step, I explored the possibility of construing Immunity as a claim about phenomenal awareness (rather than belief) and argued that this move doesn't help critics of Immunity.

I want to now look more closely at the action intention, which is plausibly the core of first-personal action awareness, and at the corresponding immunity claim which says that self-ascriptions of action intentions are immune. Again, anarchic hand syndrome is not the right case to challenge this claim. But fortunately the rich repertoire of pathologies offers a somewhat similar case that will do better.

3.3 Made Impulses and Made Volitions

In so called made impulses and made volitions, subjects carry out goal-directed actions but believe that they are not the ones who are controlling and agentically doing these things. More precisely, in made impulses "[t]he impulse to carry out this action is not felt to be his own, but the actual performance of the act is", whereas in made volition the patient feels "completely under the control of an external influence", which is to say that movements are experienced as "initiated and directed throughout by the controlling influence, and the patient feels he is an automaton, the passive observer of his own actions." (Mellor 1970: 17)

Mellor offers the following two reports as examples of the two phenomena:

> [Made Impulse] A 26-year-old engineer emptied the contents of a urine bottle over the ward dinner trolley. He said, 'The sudden impulse came over me that I must do it. It was not my feeling, it came into me from the X-ray department [...] It was nothing to do with me, they wanted it done. So I picked up the bottle and poured it in. It seemed all I could do.'
>
> [Made Volition] A 29-year-old shorthand typist described her actions as follows: 'When I reach my hand for the comb it is my hand and arm which move, and my fingers pick up the pen, but I don't control them... I sit there watching them move, and they are quite independent, what they do is nothing to do with me... I am just a puppet who is manipulated by cosmic strings. When the strings are pulled my body moves and I cannot prevent it.' (ibid.)

The crucial question is whether, in contrast to patients suffering from anarchic hand syndrome, these subjects are first-personally aware of the intentional nature of their movements. In the case of made volition, it is not clear to me whether the intentional nature of the subjects' movements is part of the experience or whether the intentional nature of the movement is rather inferred third-personally by the subject who is just a passive bystander to his body's movements. In the case of made impulses the matter is clear. Subjects are first-personally aware of an action intention, but somehow experience that intention as not their own. Hence, let me

discuss made impulses as the more promising counterexample to the Immunity Thesis regarding intentions and actions.[47]

The crucial difference between anarchic hand syndrome and made impulses is that in the latter case the subject is introspectively aware of the action intention. The difference can be illustrated as follows. In made impulses, the subject knows what he is about to do, it is just that he feels compelled by an outside force to do it and that he cannot do anything to prevent it. In anarchic hand syndrome, in contrast, the subject does not know what he (or rather, his hand) is about to do. He experiences himself as automated and finds out what his hand does by watching it act, just like you find out what other people do by watching them act.

So, let's look at the above case.[48] The discussion will parallel that of thought insertion, so I will be brief. The case can be understood to involve both a disowning claim ('this is not my intention') and an external attribution ('this is the X-ray-department's intention'). The external attribution is not a self-ascription and thereby fails to target Immunity. What about the disowning claim? Again, I want to grant that it is a negative self-ascription in the sense of being about oneself. What the judgment is based on precisely, is again a controversial empirical question. I will assume, for now, that it satisfies the introspection criterion. After all, it cannot be denied that the subject is introspectively aware of the action intention. Where the judgment 'this is not my intention' fails as a counterexample is, again, in not being in error through misidentification. The subject does not attribute the right property to the wrong person, but rather attributes the wrong property to the right person. Subjects of made impulses do not go wrong in the question whether it is they themselves or somebody else who lacks the action

[47] Thanks to Gottfried Vosgerau for suggesting this case in the first place.

[48] Note that, as far as I know, made impulses have not actually been claimed by anyone to refute Immunity. However, they do get mentioned in the relevant literature on pathologies of alienation and fit in well with the other cases. I mainly discuss them here as a possible improvement of Marcel's critique from anarchic hand syndrome.

intention, rather they go wrong in the question whether they do or do not have the intention.

3.4 Somatoparaphrenia

Somatoparaphrenia is the disowning of a body part together with the attribution of the body part in question to another person. Timothy Lane & Caleb Liang (2011) argue that a particular case of somatoparaphrenia disproves the Immunity Thesis. In this case, a stroke patient (FB) suffering a lesion in the right cerebral hemisphere attributes her left hand to her niece (Bottini et al. 2002). Moreover, FB seems not to feel touches in her left hand (hemianesthesia), unless, that is, the touches are announced as delivered to the niece's hand. The phenomenon has been assessed by touching the blindfolded patient on her left hand in two conditions. In condition one, the touch was announced as being delivered to FB's left hand, in condition two the touch was announced as being delivered to the niece's hand. FB was asked to report with 'yes' and 'no' whether she felt any touch. In condition one, FB never reported feeling touched, in condition two, FB reliably reported the touch.

Now, what does this case show? Lane & Liang take it to show that FB represents the tactile sensation, of which she is obviously aware in condition two, to be experienced not by herself but by her niece. That is, according to Lane & Liang, FB misrepresents the owner of the tactile experience, not merely the owner of the hand or the location where she (FB) feels the touch. Note how this interpretation builds on a distinction between *being* the subject or owner of an experience (or simply: being the one who experiences) and *representing oneself as* the subject of the experience. While Lane & Liang (at least implicitly) accept that it is FB who is in fact experiencing the sensation, they claim that FB does not *represent* herself as the subject of the experience, but instead misrepresents her niece as the subject. The putative counterexample to the Immunity Thesis hence is FB's judgment 'my niece feels that touch'.

What is interesting and novel about Lane & Liang's argument is how they turn what looks like a misattribution of a *bodily* state into a misattribution of a *mental* state. However, this move comes at a cost, for they have to assume a very questionable interpretation of the case. What Lane & Liang claim is that FB misrepresents *who is experiencing* the tactile sensation. But a more conservative interpretation is available that explains the results equally well: FB does not misrepresent *who* is having the sensation, rather FB misrepresents *where she (FB) experiences* the tactile sensation. To see what is meant, consider being touched first on your right and then on your left hand. Both times, you are aware that it is you who is experiencing the touch. But you experience the first touch in your right hand and the second in your left hand. Similarly, the conservative explanation holds, FB is well aware that *she* is experiencing the touch. She just misrepresents *where* she is feeling the touch, namely in her niece's hand.[49]

Which of the two interpretation is the more convincing? While both interpretations are compatible with Bottini and colleagues' description of FB's case, the conservative interpretation is the much more natural and plausible. Consider that even Lane & Liang, in their own description of the case, inadvertently support the conservative interpretation: "When advised that the examiner was about to touch her niece's hand, however, upon actually being touched, she reported feeling tactile sensation." (2009: 664) So, Lane & Liang here effectively grant that FB believes that it is *she herself* who is feeling the tactile sensation. Where is the misidentification then? What Lane & Liang seem to have in mind is that FB believes of that sensation that it is her niece's. But what does that even mean? I can make sense of how ownership of a tactile sensation can be construed in two ways. First, for a tactile sensation to be one's own could mean that one is the subject who feels that sensation. Lane & Liang seem to grant that FB does not disown the sensation in this sense: she reports that she herself feels the sensation. Secondly, for a tactile sensation to be one's own could mean that it occurs in one's own body. The conservative interpretation holds that it is

[49] For a similar reply see Rosenthal 2010.

in this sense, that FB disowns the sensation. In this interpretation, no mental state is being disowned. It does not seem to be what Lane & Liang have in mind.

Rosenthal, in defending his HOT theory against Liang & Lane 2009, argues for something along the lines of the conservative interpretation. In his words: there are "two ways an individual can subjectively own a tactile sensation [...]. In addition to being aware of bodily sensations as one's own, we are aware of such sensations as having some bodily location" (2010: 272). Rosenthal argues that FB is well aware that it is her own experience. What she misrepresents is the bodily location at which she is feeling the sensation.

Lane & Liang (2010) do not have anything convincing to say against this interpretation. In a reply to Rosenthal, they mostly do burden of proof shifting. They also offer an argument to the effect that representation of bodily location is not a proper way of understanding mental ownership. Their idea seems to be: if this is what was going wrong in FB's case, then FB would not be making a mistake in ownership ascription. At this point, they clearly beg the question against the conservative interpretation. For, the simple reply in defense of the conservative interpretation is to say: yes, that's it, FB does not misrepresent ownership but location of the sensation; it's your turn to show that she is actually misattributing mental ownership. To support this reply, one may further draw attention to the fact that, as far as the description in Bottini and colleagues' paper goes, FB never explicitly disowns the sensation at all. All that we have, in terms of disowning, is that, when the touch is announced as a touch of the niece's hand, FB will report that she feels the touch. Apart from the question whether ownership of tactile sensation can be spelled out in terms of bodily location or not, what is most plausible is that it is the bodily location of the sensation that is getting misrepresented, rather than the bearer of the experience.

The question of how to properly interpret the case will not be decisive until later. For, even if we accept Lane & Liang's interpretation, the case does not refute the Immunity Thesis. Clearly, the judgment in question ('my niece

feels that touch') does not satisfy the self-ascription criterion. The property in question (feeling touched) is not self-ascribed but attributed to the niece.[50] What the case does show, given Lane & Liang's interpretation, is that it is possible to misattribute a mental state of which one is introspectively aware to the wrong person.

In an attempt to rescue the case, it could be assumed that FB makes an implicit self-ascription. That is, by attributing the sensation to her niece, she *implicitly* disowns the sensation. For the sake of argument, let's grant that the case involves an implicit disowning belief ('I do not feel the touch'). This belief, arguably, satisfies the scope criteria in involving a negative self-ascription and in being based on first-personal awareness. However, this judgment fails as a counterexample for the same reason as the other disowning beliefs discussed above, which should by now be familiar: the judgment does not involve a misidentification. When FB implicitly judges that she does not feel touched, she does not ascribe the right property to the wrong person, but simply ascribes the wrong property to herself. She does not go wrong in figuring out who it is that does or does not feel the touch, but rather just goes wrong in whether she does or does not feel the touch. The disowning belief is in error through mispredication, not in error through misidentification. (But see § 5.3 for a rejoinder to this objection.)

3.5 Dissociative Identity Disorder

Finally, let me briefly discuss, just to put aside, a disorder that invariably comes to mind when thinking of pathology and misidentification, namely *dissociative identity disorder* (DID, formerly known as multiple personality disorder). In DID there seem to be several identities, called alters, living in one human body. Typically, the different alters are not present (i.e. conscious) simultaneously, but rather take turns 'being in front' and controlling the patient. Moreover, the alters normally have no memories of what other alters have been doing.

[50] A similar objection is raised by de Vignemont (2012) who discusses somatoparaphrenia as a possible counterexample to bodily immunity.

The question whether DID is a medically valid disorder is highly controversial. Many psychologists argue that the phenomenon is rather therapeutically induced in suggestible patients mainly through methods of hypnosis.[51] Be that as it may, the phenomenon itself seems to be quite real and, looking at it as a potential counterexample to Immunity, the etiology and the medical validity do not matter to us.

It is easy to see that several alters' inhabiting a single body might create room for confusion. Accordingly, Rosenthal has suggested DID as a counterexample to Immunity (2012). He imagines a case in which one alter is aware of another alter's pain and goes wrong in ascribing that pain to himself. This argument has been convincingly rebutted by Langland-Hassan (forthcoming) as follows. First, Rosenthal would have to assume that DID involves not only distinct alters, but actually distinct persons. For the pain-ascription to be in error through misidentification, it would have to be the case that the pain is ascribed to a person that is distinct from the person who actually is in pain. The assumption that different alters actually are different persons runs counter to psychiatrists' view on the matter, which is the view that one person houses several personalities. Second, Langland-Hassan shows that Rosenthal has no convincing argument for the claim that the pain of which the alter is aware is not *ipso facto* that alter's pain. Given these two points, I take it that DID does not make a good case against Immunity and it will not play any further role.

3.6 Summary

Let me quickly recap the discussion of the putative counterexamples. None of the cases can be construed in a way so as to involve a belief that is based on first-personal awareness, is a self-ascription, and involves an error through misidentification (see fig. 1). Hence, none of the cases refutes the Immunity Thesis. Note that this is a preliminary result as I will discuss a refined version of the critique in §§ 5-6.

[51] See e.g. http://en.wikipedia.org/wiki/Dissociative_identity_disorder, retrieved April 12th 2014.

	first-personal awareness	self-ascription	misidentification
thought insertion			
'Chris is the author'	-	-	+
'I am not the author'	?	+	-
anarchic hand			
'I am not the agent'	-	+	-
made impulse			
'It's the X-ray department's intention'	?	-	+
'It's not my intention'	+	+	-
somatoparaphrenia			
'my niece feels that touch'	+	-	+
'I do not feel that touch'	+	+	-

Fig. 1: Assessment of putative counterexamples

Two general points emerged. First, there is an important distinction between external attributions and disowning claims. In thought insertion and made impulses, subjects explicitly make both types of judgments, in FB's case of somatoparaphrenia the disowning claim is present only implicitly, and in anarchic hand syndrome there is no external attribution. I argued that external attributions fail the self-ascription constraint (a more in depth discussion follows in § 5.2). I granted that disowning claims (except in the case of anarchic hand syndrome), construed as negative self-ascriptions, satisfy both the self-ascription constraint and the introspection constraint, but argued that they do not involve error through misidentification. When subjects disown a thought, intention, or sensation they do not

go wrong in ascribing a property to the wrong person, but go wrong in saying something wrong of the right person. Hence, disowning beliefs are in error through mispredication, not in error through misidentification.[52] (More discussion of this argument follows in § 5.3.)

As a second general point, we saw that the notion of ownership, that is the notion in terms of which a thought, action, intention, or sensation is being disowned or externally attributed, stands in need of interpretation. What does it mean for a thought, action, intention, or sensation to be (or not to be) one's own? With respect to actions, we can intuitively distinguish between a notion of bodily ownership (being the person whose body is moving) and a notion of agentic ownership (being the person who is initiating or controlling the movement). It is the latter notion that is relevant to anarchic hand syndrome. This very distinction is also used in motivating an analogous distinction regarding thoughts, namely between ownership and authorship. Again, it is authorship of thoughts that is in question in thought insertion. A similar distinction suggests itself accordingly for intentions in the case of made impulses. Finally, the question in which sense FB takes the sensation to be her niece's lies at the heart of Lane & Liang's controversial interpretation of the case. The question is whether FB really believes that her niece is the one who is experiencing the sensation, or whether she merely believes that she herself is experiencing the sensation in the niece's hand.

The results are by no means a coincidence, but reflect a fundamental aspect of immunity. To wit, they reveal a tight relation between self-ascriptions, first-personal awareness, and misidentification (see fig. 1). In all the cases, the same kinds of questions must be asked in determining whether we have a counterexample to Immunity or not. Does the judgment in question involve a self-ascription? Is the judgment based on first-personal awareness? And is the judgment in error through misidentification?

---

[52] A very similar rejection of the same cases can be found in Smith (forthcoming). The ideas presented here were developed independently and I only later found out that Smith presents a very similar argument at Consciousness Online 2013.

Comparing the four cases of alienation, thought insertion and made impulses strike me as the most interesting and most promising. Lane & Liang's challenge from somatoparaphrenia depends on a highly questionable interpretation of a single case, and Marcel's challenge from anarchic hand syndrome clearly fails the introspection constraint. The two remaining cases, thought insertion and made impulses, are quite similar and it will suffice to discuss only one of them in more depth. I shall from here on lay a particular focus on thought insertion, mainly because, of the two cases, this is the one that has been suggested as a counterexample in the debate and it is the one that receives a lot more discussion in the literature. I assume that a lot in the discussion of thought insertion is representative for other cases as well.

One particular aspect of thought insertion that I take to require more discussion is the notion of authorship. In my discussion so far I have been assuming that subjects make a mistake in disowning thoughts, actions, intentions, or sensations. That is to say, I have been assuming that the thoughts, actions, intentions, or sensations are in fact the subjects' own. But given that they do feel alienated from these experiences in some way or other, it can reasonably be asked whether this assumption is correct at all. Aren't they perhaps saying something correct when they claim that it is not them who are doing the thinking in thought insertion, when they say that it is not them but the alien hand who is unbuttoning the shirt, or when they say that it is not their intention to pour the urine bottle over the dinner trolley? Without the assumption that they are mistaken in disowning these events, the disowning claims obviously would not threaten the Immunity Thesis. I now turn to an in depth discussion of the question whether disowning claims in thought insertion are correct or mistaken.

4. Authorship in Thought Insertion[53]

So far, the discussion proceeded on the presumption that subjects make a mistake in disowning inserted thoughts. While it is beyond question that they do make a mistake in ascribing the thought to an external entity, one may ask in what sense they are mistaken in disowning the thought. After all, we can assume that, typically, inserted thoughts come unasked, simply come to mind and often do not express the subject's background psychology. Why should we insist, then, that subjects are the authors of these thoughts? The main task in answering the question whether subjects are the authors of inserted thoughts is to give an analysis of the relevant notion of authorship. Hence, the question of this chapter is this: What does it mean to be the author of a thought?[54]

4.1 Preliminaries

Let me start with a reminder of the conceptual framework underlying the discussion of thought insertion. Reports of thought insertion typically involve two claims: on the one hand, the thought is claimed not to be one's own ('that is not my thought'), I call this the *disowning* of the thought; on the other hand, the thought is attributed to another agent or entity ('it is so-and-so's thought'), I call this the *external attribution* of the thought (cf. Stephens & Graham 2000: 152). In my discussion I am focusing on the disowning aspect of thought insertion.

---

[53] Substantial parts of this chapter appear also in Seeger (forthcoming).

[54] To anticipate, an analysis of what it means for a thought to be one's own is also crucial to the explanation of Immunity which I will propose later. My explanation is based on the assumption that introspective awareness of a thought guarantees authorship of that thought. Assessing this assumption, just like assessing the assumption that subjects are the authors of inserted thoughts, requires that we have a good idea of what it means to be the author of a thought.

The Conceptual Puzzle

The question is: what exactly does 'not one's own' mean when it comes to thoughts? The subjects' claim to have thoughts which are not their own presents us with a puzzle. Isn't it outright incoherent to claim being aware of a thought which one is not thinking? I refer to this question as the conceptual puzzle raised by thought insertion. The desideratum shared by many participants to the debate is to solve this puzzle, that is, to provide a conceptually coherent interpretation of the reports, an interpretation which renders the disowning claim coherent—even if false.[55] Obviously, if the disowning claim is to have even the slightest chance of being a counterexample to Immunity, we have to assume that it is a coherent claim to begin with. Importantly, when I say that the disowning claim is a *coherent* claim, I do not mean to say that it is *rational*.

Note well that thought insertion gives rise to many further questions. The ones that receive most attention are the phenomenal question, asking what the experience of thought insertion is like, and the explanatory question, asking why this experience occurs.[56] It is important not to conflate these different questions. My sole concern is the conceptual question: how are we to give a coherent meaning to the claim 'that thought (which I have) is not my own'?

---

[55] See e.g. Graham & Stephens 1994: § 2, Campbell 1999b: 611, 619f., Stephens & Graham 2000: 149ff., and Fernández 2010: 68. Coliva (2002b) recognizes the conceptual puzzle, but—as a notable exception—does not see a need to solve it: she claims, in contrast, that reports of inserted thoughts are indeed incoherent.

[56] Regarding the phenomenal question, see e.g. Stephens & Graham 2000, Gallagher 2000 and 2007a, Bortolotti & Broome 2009, Pickard 2010, Fernández 2010, Sousa & Swiney 2011, Seeger (2013), and Sollberger (forthcoming). Regarding the explanatory question see Feinberg 1978, Frith 1992, Daprati et al. 1997, Campbell 1999b, Stephens & Graham 2000, Gallagher 2004a, Bayne & Pacherie 2007, Vosgerau & Newen 2007, Synofzik et al. 2008, Sugimori et al. 2011, Martin & Pacherie 2013, and Swiney & Sousa 2013.

The key to solving the conceptual puzzle lies in the distinction between the two senses in which a thought can be one's own: ownership and authorship. On the one hand, a thought can be one's own in the sense that one is the person introspectively experiencing the thought or that the thought is within one's stream of consciousness. I will refer to this aspect by saying that the person has *ownership* for the thought or is the *subject* of the thought. On the other hand, a thought can be one's own in the sense that one is the person actively bringing the thought about or that one is the (causal) source or originator of the thought. I will refer to this aspect by saying that the person is the *author* or *agent* of the thought. Using this distinction, the standard interpretation is to make sense of the disowning claim 'I have a thought that is not my own' as meaning 'I am the subject of that thought, but I am not its author'.[57] In what follows, I assume that this is the correct interpretation. The question I discuss in this chapter is how to exactly understand the agency/authorship side of the distinction. In other words, the question is what it is for a thought to be one's own in precisely that sense, in which subjects take inserted thoughts to not be their own.

A remark on terminology is in order. While the distinction itself is generally accepted, there is no established taxonomy. Campbell refers to the two different aspects as "two strands in the ordinary notion of the ownership of a thought" (2002: 36), Graham & Stephens distinguish between being the *subject* and being the *agent* of a thought (cf. Graham & Stephens 1994: 98; Stephens & Graham 2000: 152f.), Gallagher distinguishes between *ownership* and *agency* (cf. 2000: 203f.), and others distinguish between *ownership* and *authorship* (e.g. Hoerl 2001; Gerrans 2001; Vosgerau & Voss, forthcoming).

I will use the terms 'authorship' and 'agency' (and their cognates) interchangeably as these are most commonly used today. However, I do so without thereby meaning to import any of the connotations attached to the

[57] Do not confuse this standard interpretation of the disowning *belief*, i.e. the idea that what subjects deny is authorship, with the standard account regarding the *phenomenology*, i.e. the idea that subjects have a disturbed experience of agency.

terms (e.g. the idea that thoughts are (motor-)actions, or the idea that being an agent means being an intentional agent). Rather, I use them as placeholders to denote whatever property it is that subjects deny of themselves when they say things like "Thoughts are put into my mind" and that they attribute to others when they say things like "They come from this chap, Chris. They're his thoughts" (Frith 1992: 66).

Authorship vs. Sense of Authorship

Now, going back to the distinction between authorship and ownership, there is a parallel distinction between the *sense* of authorship and the *sense* of ownership. The sense of authorship is at the core of the discussion of both the phenomenal question (it is argued that thought insertion involves a disturbed sense of authorship) and the explanatory question (explaining the disturbed sense of authorship means explaining why subjects experience thoughts as inserted). In this thesis, I will have nothing to say about the sense of authorship. To distinguish questions regarding the sense of authorship more clearly from questions regarding authorship I will also refer to authorship as *metaphysical authorship*. The distinction between phenomenology and metaphysics of authorship allows for the possibility of mistaken phenomenology: it is possible to experience oneself as the author when really one is not, or to fail to experience oneself as the author when really one is.[58]

The whole debate suffers from the conflation of issues pertaining to authorship with issues pertaining to the sense of authorship, or, to say the least, of verbal conflations of the terms 'authorship' and 'sense of authorship'.[59] Examples abound, let me mention just one to underscore the importance of keeping this distinction in mind. Philip Gerrans, in reconstructing Stephens & Graham's (2000) view, says that

[58] Cf. Horgan et al. (2003). But see Gallagher (2007b: 1) for the assumption that agency presupposes a sense of agency.

[59] Especially in the empirical literature, the terms 'authorship' and 'agency' are often used as shorthand for 'sense of authorship' and 'sense of agency', respectively.

> [t]hey distinguish a *sense* of subjectivity from a *sense* of agency, or, as I shall put it, a *sense* of ownership from a *sense* of authorship. In cases of auditory hallucination and thought insertion, the agent has a psychological experience, which she *owns*, in the sense that it occurs within her mind, not the mind of someone else, but of which *she is not the author*. That is to say that the audition or cognition is *experienced* as somehow originating in the mind of someone else who causes the subject to *experience* it as occurring within her mind. (Gerrans 2001, 231; my emphases, Gerrans's emphases omitted)

Gerrans begins with the distinction between sense of ownership and sense of authorship in the first sentence, slips into talk of ownership and authorship in the second sentence, and goes back into talk of experience in the last sentence, which is supposed to elucidate the second. Should we take Gerrans literally in claiming that subjects are not the metaphysical authors of inserted thoughts? And should we take him literally in claiming that not *being* the author is the same as *experiencing* oneself to not be the author? I think the most charitable, actually the only viable interpretation is that he, like many others, uses the term 'authorship' here as shorthand for 'sense of authorship'. The point I am making here is that with this kind of confusion in the debate, it is often not entirely clear whether writers make claims about authorship or about the sense of authorship. This, in turn, is a challenge to my project of reconstructing views of what it is to be the author of a thought. In particular, it may be thought that some of the theories I discuss as views on metaphysical authorship are really intended as views on the phenomenology of authorship. I will address this worry where adequate.

So, to repeat, I am not asking what the phenomenology of authorship is like. I am also not asking why one experiences oneself as the author or not. The question is this: when somebody claims 'that is not my thought' (in the sense of 'not my thought' that is at play in reports of thought insertion), what does it take for that claim to actually be true?

Methodology

A natural starting point to address the conceptual question is to analyze reports of inserted thoughts. However, two points should be borne in mind when looking at patient reports. First, the current debate feeds on a very limited diet of reports; we find roughly a handful of examples that are cited over and over again.[60] Apart from the fact that the origins of the examples are rather obscure (e.g., it is not clear whether they are verbatim transcriptions or examiner's notes from memory), it is very unclear whether this sample is representative at all. Certainly, we would need more data to assess empirical claims such as Fernández's assumption that all inserted thoughts are beliefs (2010: 69f.). Secondly, it is not entirely clear how well the available reports express the subjects' beliefs. Most patients are simply not willing to talk about their experiences and beliefs. To get anything out of a patient at all, examiners often have to be quite suggestive. In so far as patients are willing to describe their experiences, it is often conceptually challenging for them to do so.[61]

Given these difficulties, I will try to appeal to reports as little as possible in arguing for an analysis of authorship. Of course, in providing an analysis of these reports I cannot ignore them entirely. But I will not base my analysis on idiosyncratic features of particular reports, but only on coarse features. Particularly, I appeal to two features of patient reports that I take to be common and not subject to the worries raised above: first, in some way or other a thought or thinking is disowned; second, the disowning belief and external attribution are often expressed using causal terminology.

Let me quickly explain what I have in mind in criticizing the appeal to idiosyncratic features by discussing some methodological remarks by

[60] The most prominent are 'Kill god' (quoted in Frith 1992: 66); 'Eamonn Andrews' (quoted in Mellor 1970: 17); and 'murder Lissi' (quoted in Mullins & Spence 2003: 295).

[61] Thanks to Martin Voss for arranging my sitting in on several psychiatric ward rounds. These experiences have made me very sensitive to the methodological issues pertaining to the analysis of thought insertion.

Fernández regarding the analysis of thought insertion. First, Fernández suggests we need to respect patient reports:

> The first constraint that we need to respect is the patients' reports. […] Any account of thought insertion must explain why patients with this disorder make claims of the general form 'I believe that such-and-such but that belief is not mine'. (2010: 71)

I agree that we should respect patient reports. But I disagree that these reports have the form Fernández takes them to have. Subjects do not disown beliefs, but thoughts. Even if some reports can be construed as a disowning of a belief (it is not even clear how all the reports adduced by Fernández fit this bill), this is hardly a general feature of reports.

Second, Fernández finds his own analysis of inserted thoughts (inserted thoughts are beliefs the contents of which are not endorsed by the subject) supported by the fact that it "accounts for certain details in some reports" (ibid.: 79). In particular, Fernández takes his analysis to be supported by the fact that it accounts for the details of an extraordinarily confused report in which one patient "describes the disowned thought as feeling like 'a piece of information'" (ibid.). Whether a thought is described as feeling like a piece of information or not is what I take to be an idiosyncratic feature of a report, a feature which should not inform a general notion of authorship.

Apart from appeal to the two very general features of disowning and causal language, I will further assume that patients want to express something extraordinary and bizarre. However, my main argument does not build on patient reports, but on paradigmatic cases of authorship and non-authorship, i.e. on cases of which everyone in the debate should agree that they involve authorship or not.

Paradigm Cases

Paradigmatic cases of authorship are thoughts resulting from consciously controlled or directed thinking processes. Examples from the literature are "[p]roblem solving, thinking through a set of instructions, and narrating a story" (Gallagher 2000: 225), "trying to comprehend, e.g., the proof of completeness for first order logic" (Vosgerau & Voss, forthcoming),

intentionally “imagining the place where I am going to spend my vacation” (Sousa & Swiney 2011: 4f.), reciting a poem silently to oneself or mentally rehearsing an argument (Stephens & Graham 2000: 150), and turning one’s “thoughts to a certain topic or project, such as discovering a counterexample to Quine's Indeterminacy of Translation” (Stephens & Graham 1994: 6). I refer to these paradigmatic cases as thoughts that are the result of directed thinking, or, for the sake of brevity, as *directed thoughts*.

Note that all these paradigmatic cases of authorship are at the same time paradigmatic cases of a profound *sense* of authorship. That is, all cases involve a sense of intentional guidance or control, perhaps even a sense of effort. These cases are frequently contrasted with thoughts or thinking processes in which one lacks any such sense of intentional agency: day-dreaming, bits of doggerel, catchy tunes stuck in one’s head, memories that impinge on one’s consciousness, fantasies, etc.[62] Such thoughts that simply come to mind may have an unwelcome content or may be annoying, but they may also be welcome and helpful (remembering an important appointment just in time). I refer to them as *unsolicited thoughts*.[63] While everyone agrees that there is a phenomenal difference between directed and unsolicited thoughts, it is typically left unclear whether unsolicited thoughts are authored by their subjects or not.

What, now, are paradigmatic cases of non-authorship? Surprisingly, one finds hardly any examples in the literature. Here are my own suggestions which, I take it, everyone should agree on. The simple idea is that, if thought insertion were indeed possible, that is if someone or something were able to telepathically or instrumentally insert a thought into another

[62] See e.g. Frankfurt 1988: 59, Gallagher 2000: 215 and 225f., 2004b: 90f., Langland-Hassan 2008: 375f., Pickard 2010: 62, and Stephens & Graham 1994: 6 and 2000: 150).

[63] Perhaps most of our everyday thoughts are neither clearly directed nor completely unsolicited, but lie somewhere in between. Nothing in my view depends on there being a clear-cut distinction; there may be a broad spectrum of thoughts being more or less directed. Thanks to Michael Sollberger for pointing this out.

person's mind, then the receiving subject would not be the author of that thought. By telepathic thought insertion I mean the transfer of thoughts by the sole power of the mind (think of the science fiction character Professor X or of a Cartesian Demon). By instrumental thought insertion I mean (again with a bit of science fiction) the direct triggering of the neural correlate of a thought, e.g. by means of something like transcranial magnetic stimulation or direct electric stimulation by sticking an electrode into the brain.[64] I dub these cases *truly inserted thoughts*.

Equipped with these paradigmatic cases and having identified the unclear cases (unsolicited thoughts), I assume that an analysis is adequate only if it classifies directed thoughts as authored by the subject and truly inserted thoughts as not authored by the subject. I take it to be an additional asset of an analysis if it provides a systematic and convincing classification of unsolicited thoughts.

4.2 The Agency Analysis

Lynn Stephens and George Graham, who pioneered the philosophical debate of thought insertion, present in various publications the idea that the claim 'that is not my thought' should be understood as the claim that the subject is not the *agent* or the *active thinker* of the thought. They base their analysis on Harry Frankfurt's influential distinction between actively thinking a thought versus passively experiencing a thought.

> In our intellectual processes, we may be either active or passive. Turning one's mind in a certain direction, or deliberating systematically about a problem, are activities in which a person himself engages. But to some of the thoughts that occur in our minds, as to some of the events in our bodies, we are mere passive bystanders. Thus there

[64] Think of the Penfield experiments. See Desmurget et al. 2009 for a study in which direct electric stimulation of the brain induced either motor intentions in the subjects or (at higher intensity) the false belief that they actually have moved. One patient, for instance, reported "I felt a desire to lick my lips" or "I moved my mouth, I talked, what did I say?", correspondingly (ibid.: 812).

> are obsessional thoughts, whose provenances may be obscure and of which we cannot rid ourselves; thoughts that strike us unexpectedly out of the blue; and thoughts that run willy-nilly through our heads.
>
> The thoughts that beset us in these ways do not occur by our own active doing. It is tempting, indeed, to suggest that they are not thoughts that *we think* at all, but rather thoughts that we *find* occurring within us. This would express our sense that, although these thoughts are events in the histories of our own minds, we do not participate actively in their occurrence. The verb "to think" can connote an activity – as in "I am thinking carefully about what you said" – and with regard to this aspect of its meaning we cannot suppose that thoughts are necessarily accompanied by thinking. It is not incoherent, despite the air of paradox, to say that a thought that occurs in my mind may or may not be something that *I think*. (1988: 59)

Note that this distinction between active and passive thinking is motivated by contrasting paradigmatic cases of directed thoughts with cases of unsolicited thoughts. The most salient difference between directed and unsolicited thoughts is that, in cases of directed thoughts, the thinker has an occurrent intention to think in a certain direction or about a certain topic, whereas in cases of unsolicited thoughts, the thinker doesn't have any such intention, but has thoughts simply come to mind. Hence, Stephens & Graham remark that active thinking "is something that I often feel I do voluntarily, even deliberately" whereas in passive thinking "I may feel that certain thoughts occur in me through no doing of my own." (Graham & Stephens 1994: 98f.) The main idea behind the Agency Analysis can be summed up thus: being the agent of a thought means intentionally or voluntarily bringing that thought about by directing one's thinking in a certain way. A very similar notion of mental agency has been proposed by Peacocke, according to whom a mental action φ essentially involves a trying to φ (cf. 2007: 361).

The distinction between active and passive thinking then allows to solve the conceptual puzzle: subjects accept that they are passively experiencing the thought in question, but deny that they are actively bringing it about. In Stephens & Graham's words: the person acknowledges that she is the

subject in whom the thought occurs, but denies that she is the active thinker of the thought (cf. e.g. Graham & Stephens 1994: 98).

Before I criticize their view, let me address a question regarding my interpretation of Stephens & Graham. There is a legitimate worry that they intend the distinction between active and passive thinking only to illustrate the *phenomenology* of authorship, and that I am misconstruing their view in applying the distinction to the definition of *metaphysical* authorship. I have several replies to that worry. First, they do make literal claims about mental agency, rather than about the sense of mental agency (see e.g. Graham & Stephens 1994: 98f., Stephens & Graham 1994: 4, and 2000: 150f.). Since they are very attentive to the distinction between agency and the sense of agency, these passages do not seem to be mere lapses. Relatedly, Frankfurt's distinction between active and passive thinking, on which they base their analysis and which they cite approvingly, clearly is a metaphysical distinction, not a phenomenal distinction. But most importantly, Stephens & Graham explicitly state that they want to solve the conceptual puzzle (cf. Graham & Stephens 1994: § 2, Stephens & Graham 2000: § 7.2). The puzzle arises from the subject's belief that the thought of which he is aware *is* not his. To provide a coherent interpretation of this belief in terms of a denial of agency, means to explain what it is to *be* the agent of the thought. It does not suffice to explain what it is to have a phenomenal experience of agency (phenomenal question) or to explain why subjects of inserted thoughts do not have this experience (explanatory question). Hence, if their distinction between active and passive thinking merely marked a distinction between the *sense* of agency and the *sense* of subjectivity (rather than between agency and subjectivity), then Stephens & Graham would not have provided a solution to the conceptual puzzle.

The exegetical worry highlights the central problem of the Agency Analysis, which is that it is built on the wrong contrast. The distinction between active and passive thinking is a distinction based on the phenomenal difference between directed and unsolicited thoughts. Applying this distinction to the analysis of authorship implies that both unsolicited and inserted thoughts are not authored by their subjects. It is the mark of

unsolicited thoughts that they are not entertained intentionally, so to speak. I assume that, very similarly, inserted thoughts are not entertained intentionally. The Agency Analysis thus implies that those who experience inserted thoughts are actually correct in disowning these thoughts. Although some theorists may be prepared to bite this bullet (see e.g. Carruthers 2012: 299 fn. 3), the implication sits rather uneasy with the idea that the disowning claim is a delusional belief which reveals a serious psychiatric disorder.

More generally, the Agency Analysis is implausible in treating unsolicited thoughts on a par with inserted thoughts. Remember that we are ultimately concerned with interpreting the claim 'that is not my thought'. Interpreting this claim in a way so that it is true of all our everyday unsolicited thoughts seriously trivializes the claim: what subjects of inserted thoughts claim, according to this interpretation, is no more than that they didn't think the thought intentionally. But that is not a strange or extraordinary claim at all.[65] This interpretation therefore does not do justice to the fact that subjects of inserted thoughts are typically very distressed and troubled by their experience of thoughts which they believe not to be their own. When they claim that certain thoughts are not their own they want to express more than the thought's being involuntary or unsolicited.

Finally, there is an internal tension in the views of both Stephens & Graham as well as Peacocke. Under the assumption that inserted thoughts are not intentionally brought about, it follows from both views that subjects are indeed not the agents of inserted thoughts. This implication, which the proponents of the Agency Analysis do not seem to be aware of[66], does not fit their view that thought insertion involves a disturbed sense of agency. If subjects actually are not the agents of the inserted thoughts, then they

[65] For analogous criticisms regarding the sense of agency see Gallagher 2000: 215; Langland-Hassan 2008: 371; Fernandez 2010: 69.

[66] Witnessed e.g. by the following remark: "No doubt her belief [that she is not the agent of the thought] is mistaken, but it is not incoherent or unintelligible." (Stephens & Graham 1994: 100)

should not experience a sense of agency for them. However, Stephens & Graham as well as Peacocke take the absence of the sense of agency to be a disruption; in fact, they take it to be the core of the pathology (cf. Stephens & Graham 1994: 100; Peacocke 2007: 368).

To be fair, there is a potential response Stephens & Graham may give to this critique. In some passages they express the idea that the thought is experienced as intended after all, just not as intended by oneself. As an analogy, they suggest the case of an anarchic hand writing a letter. The writing is experienced as expressing someone's intentions (it is a meaningful letter), yet they do not seem to be the subject's intentions (the subject does not even know the person to whom the letter is addressed). (Cf. Graham & Stephens 1994: 106, Stephens & Graham 1994: 7, 2000: 174f.) Analogously, it may be thought that inserted thoughts are experienced as expressing someone's intentions, yet not one's own. On this view, then, thought insertion is not really characterized by a *lacking* sense of agency (as it is often put), but rather by an *alienated* sense of agency. Such a view runs counter to the standard assumption in the debate, which is the assumption that subjects indeed are the agents, and the disorder is exhibited in a lacking sense of agency. The present view, in contrast, would have to hold that subjects indeed are not the agents, and the disorder is precisely that they do experience a sense of agency where none should be. While this is a possible reply, it is not clear whether Stephens & Graham would subscribe to this line of thought.

I have put it somewhat tendentious. Perhaps a fairer formulation would be this: the normal assumption is that the disorder consists in the subejcts' not experiencing themselves as the agents. The presently discussed reply, in contrast, construes the disorder as consisting in an alienated sense of agency. But putting it this way reveals more clearly the controversial nature of the reply. It is simply not clear what it should mean to have an alienated sense of agency for thoughts. The whole notion of a sense of agency for thoughts is controversial. Perhaps we can imagine a lacking sense of agency, but it is asking even more to accept that there should be a sense of agency for thoughts that represents the thought as being brought about by

somebody else. Note that the analogy to the anarchic hand is not particularly helpful here. The hand's writing is experienced as intentional in a solely third-personal way, namely because the subject realizes that meaningful things are being written. But in the case of thought insertion things are different. The above line of reasoning asks us to imagine a first-personal sense of alien agency. I am not going into the discussion whether this is possible or not. I am here simply noting that this is a particularly demanding line of response.

Summing up, the Agency Analysis implies that subjects are correct in disowning inserted thoughts and that the disowning claim amounts to nothing more than that the thought was not intended. Further, it does not fit well with the standard view that thought insertion is characterized by a lacking sense of agency. Given all this, the Agency Analysis does not provide a convincing notion of authorship in the analysis of thought insertion.

4.3 The Personality Analysis

A different approach that may appear promising construes authorship in terms of a thought's standing in a certain relation to the thinker's background intentionality. I will dub this the Personality Analysis. A remark by Campbell about the causal interwovenness of one's occurrent thoughts with one's background beliefs and desires points in that direction:

> What makes my occurrent thoughts mine [i.e.: what makes me the author] is […] the fact that they are products of my long-standing beliefs and desires, and that the occurrent thinking can affect the underlying states. (1999b: 621)

According to Campbell, authorship requires a two-way causal connection between the occurrent thought and one's background psychology. Let us bracket the latter direction, for obviously any conscious thought can in principle influence one's background psychology. I focus on the idea that for a subject to be the author of a thought means for the thought to *express* the subject's long-standing dispositional states (cf. ibid.: 620).

I think that Campbell is on the right track as far as the causal aspect is concerned. His analysis provides a sufficient criterion for authorship: if a thought is the causal product of a subject's background psychology, then the subject is the author of that thought. However, the success of the analysis depends solely on the causal aspect; restricting the causal source to long-standing dispositional states strikes me as too narrow. I can deliberately entertain a thought of any content just on a whim. Since I entertain it deliberately I am the author in the paradigmatic sense; but since the thought can have any kind of content my authorship cannot in any substantial sense depend on the thought's expressing my background psychology.

Further, the Personality Analysis implies, just like the Agency Analysis, that (many) unsolicited thoughts are not authored by the subject. For, many of one's everyday unsolicited thoughts are not the product of one's background psychology. Nor are inserted thoughts. Hence, Campbell's analysis is not much more convincing than the Agency Analysis. It faces the same problems as the Agency Analysis in implying, first, that disowning claims are (often) correct, and second, that disowning claims can be true in a non-bizarre everyday sense.

4.4 The Rationalist Analysis

For the sake of completeness, I now briefly discuss what I call a rationalist view of what it means for a thought to be one's own. This view takes the term 'authorship' quite literal in an everyday sense and holds that being the author of a thought means endorsing the thought, being committed to the thought, and being able to provide reasons for it.

Bortolotti & Broome, for instance, write that a subject "is the author of that belief, because she is in a position to provide reasons for supporting it" (2009: 207). The following passage is even more explicit:

> The notion of authorship we are proposing is primarily concerned with the endorsement of the content of a mental state, where the endorsement is measured in terms of the capacity for reason giving. In order to be the author of a belief in this sense, to take responsibility

> for it and be committed to its content, it is not necessary to assume that the reasons that justify endorsing the belief map onto the psychological causes of its formation. (ibid.: 212)

Note well that, according to rationalism, the actual (psychological) causes of the belief do not seem to be relevant for the question of authorship. In another passage they maintain that authorship "is not hostage to the way in which beliefs are *formed*" and that authorship is "not about *creation* ex nihilo" (ibid.: 213; my emphases). These passages strongly suggest that, according to rationalism, the cause of a thought is irrelevant to the question whether the thought is one's own.

Before I criticize the view in more detail, a caveat is in order. Passages similar to the ones above can also be found in Fernández (2010) and Pickard (2010). Although some passages are quite explicit, I think that neither Bortolotti & Broome nor Fernández or Pickard are actual or intentional proponents of the rationalist view of metaphysical authorship. More precisely, metaphysical authorship certainly is not what they are primarily concerned with. What these authors actually, or primarily, propose is a rationalist view concerning the phenomenal question. They argue against the agency view, which describes the phenomenal experience underlying thought insertion in terms of a lacking sense of agency for the thought. The rationalist view regarding the phenomenal question holds, in contrast, that the underlying experience is the subject's not endorsing and not being committed to the content of the thought. Why do I discuss the rationalist view as an analysis of metaphysical authorship, then? There are several reasons.

First, when taken literally, some of these authors make claims addressing the conceptual question (see quotes above). Of course, it may be that they just use 'authorship' as shorthand for 'sense of authorship'. But it may also be that they do not clearly separate authorship from sense of authorship. They may even think that the analyses of the respective notions go hand in hand (see below). In that case, they might be defending the rationalist view as an analysis of authorship after all.

Second, all three papers take as their departing point Moran's (2001) notion of first-person authority which is clearly meant as a notion of what it is to *be* the author, not a notion of what it is to *experience* oneself as the author. (Note however, that my criticism of the Rationalist Analysis is not intended as a criticism of Moran's view. Moran is interested in the epistemic privilege of self-knowledge. In my view, it is simply a mistake to apply his notion to the analysis of authorship in thought insertion.)

Finally, the rationalist notion of authorship is an interesting alternative to discuss independently of whether it is actually professed by the writers who motivate the view. Rationalists regarding the phenomenal question may have a prima facie motivation to embrace also rationalism regarding the conceptual question. For, suppose the rationalist view regarding the phenomenal question is correct, that is, the experience of thought insertion is the experience of a thought that one does not endorse and is not committed to. Then we have at least prima facie motivation to explore the possibility that what subjects mean by disowning the thought is that they are not committed to the thought. Note that, while there may be prima facie motivation to analyze the conceptual and the phenomenal question analogously, they do not necessarily go hand in hand. It may well be that the nature of the phenomenal experience does not correspond to the meaning of the disowning claim. For instance, a thought's being experienced as morally repulsive may lead to the disowning claim that one is not the causal origin of the thought.

Pickard is well aware that the answers to the phenomenal question and the conceptual question can come apart. Here is her take on the phenomenal question: "I propose that schizophrenics disown mental events that seem to be manifestations of mental states that they do not, for some reason or other, endorse." (2010: 67) The experience underlying thought insertion, according to Pickard, is the experience of a mental event that one does not endorse. This quote is followed by a passage that reveals her take on the conceptual question: the above experience explains "why they disown these manifestations – why they do not believe that these relentlessly occurring thoughts, impulses, or feelings are caused by their own mental states, and

so expressive of their own background psychology". That is to say, the disowning belief is spelled out by Pickard as the belief that the thoughts are not caused and thus not expressive of one's own mental states. Whether one endorses a thought and whether that thought is caused by one's own mental states are logically separate issues.

Fernández, in contrast, seems to falsely assume that the phenomenal experience underlying thought insertion must correspond to the disowning claim. For, his general strategy is to solve the conceptual puzzle by answering what he calls the what-question: "What is the experience E that a subject [of thought insertion] is trying to express with [the disowning claim]?" (2010: 69) Given his rationalist view regarding the phenomenal question and the assumption that the disowning belief expresses the phenomenal experience, we can expect Fernández to actually be a proponent of rationalism regarding metaphysical authorship.

Let us then get back to the rationalist view, or rather, to the criticism of it. The rationalist notion of authorship does not provide an adequate analysis of thought insertion. First, endorsement plus reason giving cannot be a necessary condition for authorship. Otherwise, only thoughts that one endorses would be one's own. Any thoughts that one merely entertains, without endorsing them, would *ipso facto* not be one's own. But that doesn't deliver the right picture. Cases of directed thoughts are paradigmatically one's own, but I one can clearly entertain thoughts which I do not endorse. Consider for instance goal directed problem solving: I may deliberate on several options and weigh them against one another. In the end, I might endorse one of them and reject the others. Does that mean that the rejected options do not express my thoughts, just because I do not endorse them? This can't be right. The solution which I eventually endorse is my thought in just the same way as the options which I eventually reject are my thoughts.

This shortcoming of the whole approach is reflected in the fact that the approach is limited to beliefs (and possibly desires). For other mental states, such as questions, commands, insults, or mere entertainings, the

analysis simply does not apply. But if we know anything about inserted thoughts, it is that the thoughts in question fall into these categories rather than in the category of belief. So the whole analysis does not even get off the ground. Fernández (2010) suffers from an almost grotesque dialectic as a result of this mistake. He assumes, in the very beginning of his analysis, that inserted thoughts are beliefs (cf. ibid.: 69). The best he has to offer by way of argument for this assumption is this: Inserted thoughts are suddenly entertained thoughts that are experienced as strange. Merely entertaining a thought, even suddenly and out of the blue, is not strange, it happens all the time. He goes on to explain the strangeness of these thoughts with the conjecture that they are beliefs the contents of which are not endorsed by the subjects. And surely it is a strange thing to find a belief in oneself the content of which one does not endorse. Or, to bring out the absurdity even more, it surely is a strange thing to believe something that one does not believe. What is the argument here? To support the conjecture that inserted thoughts are beliefs, Fernández assumes that subjects do not endorse the content of the inserted thoughts (cf. ibid: 78-80). But ironically he thereby just shows that the thoughts in question are not beliefs after all.

Second, endorsement plus reason giving also cannot be a sufficient criterion for a thought's being one's own. Consider a paradigmatic case of a truly inserted thought, for instance a thought that is electrically induced in a subject. Now suppose further that the subject endorses that thought and is even able to provide (however confabulated) reasons for it. According to Rationalism, the subject then is the author of the thought. Remember that the actual psychological origins of the thought do not matter according to Rationalism. But, again, that is clearly the wrong result. That is not the sense of 'my own thought' in which thoughts are disowned by subjects of inserted thoughts.

4.5 The Causal Analysis

The most intuitive analysis of the disowning claim in thought insertion understands authorship as a causal property. I dub this the *Causal Analysis*: a thought is the subject's own thought if and only if it causally originat-

ed in the appropriate way within the subject. Most participants to the debate indeed do describe authorship and the sense of authorship in causal terms.[67] Prima facie, there seems to be general agreement then. However, the Causal Analysis is rarely (if ever) spelled out in any detail or argued for. A particular challenge is raised by unsolicited and inserted thoughts. While the debate proceeds on the assumption that subjects make a mistake in disowning inserted thoughts, it is not at all clear in which sense these thoughts actually are their own. After all, we can assume that these thoughts, just like unsolicited thoughts, simply come to mind and express neither the subject's occurrent thinking intentions nor the subject's background psychology. Why, it may be asked, should we assume that inserted (as well as unsolicited) thoughts are the subject's own?

In developing an answer, it will be helpful to remember the shortcomings of the Agency Analysis and Personality Analysis. The Personality Analysis got it right about causation, but construed the causal base too narrowly. The Agency Analysis similarly construes mental agency too narrowly as intentional agency (in analogy to intentional bodily action). To properly analyze disowning claims we need a broader notion of mental activity, a causal notion.

A Broad Notion of Mental Activity

Sousa & Swiney (2011) distinguish between a broad and a narrow notion of mental agency. Stephens & Graham's Agency Analysis is cast in terms of the narrow notion of agency, i.e. the notion of intentional agency. Sousa &

[67] Here is a brief sample of how theorists interpret the reports (all emphases mine): subjects believe to have "knowledge of token thoughts which were *formed* by someone else" (Campbell 1999b: 620); they claim not to be "the person who *generated* that very thought" or "the person who *brought* that particular token thought *into existence*, the person who *formed* it" (Campbell 2002: 35); they claim not to be the "*causal source producing* the occurrence of the [thoughts]" (Sousa & Swiney 2011, n.p.) or not to be "the agent who *produces*" the thought (Graham & Stephens 1994: 100); instead, they claim to be "the passive recipient of alien thoughts that are the *products of alien thinking*" (Mullins & Spence 2003: 295).

Swiney object that thought insertion should be described in terms of the broad notion of agency, a notion according to which being the agent simply means being the causal originator of the thought. The central idea behind this approach is that subjects of inserted thoughts do not merely claim to not think these thoughts intentionally, they claim to not be causally involved in the thinking at all.

A good way to illustrate the idea is by analogy to bodily movements. In spelling out the Agency Analysis, Stephens & Graham themselves make strong use of analogies between mental actions and bodily actions. Primarily, their analogies are intended to make plausible the distinction between ownership and agency. Just as there can be movements of my body that are not my actions (e.g. when someone else lifts up my arm), there can be movements of my mind that are not my mental actions. This is to say, I can have ownership for events in my mind of which I am not the agent. But when it comes to the question, what it is to be a mental agent, Stephens & Graham are misled by their analogies in focusing too narrowly on intentional agency.

The right analogy to draw is not to intentional bodily agency, but to what we may call bodily activity, a class of events that comprises both intentional actions and mere doings. By 'doings' I mean motor-activities that are not consciously or intentionally performed and therefore do not count as intentional actions. Examples are sleepwalking, reflexes, breathing, eye blinking, gaze shifting, and perhaps also non-attended automatic skilled action such as steering on a long-distance drive. Both intentional actions as well as doings are things a person does (there may not even be a sharp line to distinguish between the two), both kinds of movements are the subject's own in the sense that the subject is the causal originator.

The analogous notion of mental activity is one that comprises both directed as well as unsolicited thoughts. Even if unsolicited thoughts are the result of subconscious, non-intended, automatic processes, they are brought about by the subject herself (rather than by somebody else). This notion of mental activity can be cashed out in causal terms by saying that a thought is a

subject's own just when it causally originates in the appropriate way within the subject. The contrast that needs to be drawn (in the interpretation of thought insertion) is not the contrast between intentional and non-intentional mental episodes, but the contrast between things done *by* me (be they done intentionally or not) and things done *to* me.[68]

Advantages of the Causal Analysis

The Causal Analysis holds that a thought is the subject's own if and only if it causally originated within the subject in the appropriate way. This analysis has a number of advantages. First, it takes the causal terminology seriously that we find both in many patient reports[69] as well as in most theoretical descriptions of thought insertion. It accommodates all the paradigm cases and provides a convincing classification of unsolicited thoughts as being the subject's own. In particular, inserted thoughts are the subjects' own and hence disowning claims are mistaken and bizarre.

The Causal Analysis also makes intelligible why subjects, apart from disowning the thought, also feel the need to attribute it to an external entity.[70] The Causal Analysis interprets the disowning as the belief that the thought did not originate within the subject herself. This belief naturally raises the question who or what then caused the thought. In contrast, the Agency Analysis interprets the disowning claim as the belief that one did not intentionally think that thought. But this belief does not explain the further postulation of an external entity; for non-intended thoughts are an everyday phenomenon.

Finally, the Causal Analysis not only provides the best interpretation of thought insertion, but provides a basis for the analysis of other passivity phenomena as well. Several authors have made the plausible suggestion

[68] Cf. Fulford 1989: 221-29, quoted in Stephens & Graham 1994: 7. Stephens & Graham approvingly cite Fulford's contrast, yet fail to draw the right lesson from it.

[69] For a list of diverse causal mechanisms that have been reported see Mullins & Spence 2003: 295.

[70] Thanks to Peter Langland-Hassan for pointing this out.

that the disowning in thought insertion is comparable to the disowning of emotions in the phenomenon of *made feelings* and to the disowning of agency in cases of *alien control* (see e.g. Graham 2004: 90f., Pickard 2010: 56f.). The Causal Analysis can aptly capture this similarity. According to it, thoughts are claimed not to be one's own in the very same sense in which emotions or actions are claimed not to be one's own, namely as not having causally originated within oneself.

Applying the Causal Analysis

I have said above that a thought is the subject's own thought if and only if it causally originated in the appropriate way within the subject. Spelling out, in general, what the appropriate way is, is a complicated matter. However, we can get quite far by way of example. So let us see how the Causal Analysis deals with the clear and the not so clear cases.

It is beyond question that directed thoughts causally originate in the appropriate way within the subject. It is equally clear that paradigmatic cases of non-authorship, truly inserted thoughts, do not causally originate in the appropriate way within the subject. Although they may be realized in the subject's brain, they are directly caused by a deviant external mechanism and are therefore in the relevant sense not the subject's own thoughts.

For a thought to causally originate within a subject, it does not suffice that the causal source is located within the body of the subject. Obviously, thoughts directly caused by a small alien living within the skull or by an externally controlled device implanted in the brain are not the subject's own thoughts. For the alien and the device are clearly not part of the subject. Putting it very simply, the thought has to causally originate within the subject's brain; putting it a bit more broadly, it has to originate within the subject's nervous system.

What about unsolicited thoughts, are they the subject's own? I take it that their causal genesis is comparable to that of directed thoughts, but that is an empirical question and I am not going to argue for it. Crucially, it is not a deviant external cause that triggers unsolicited thoughts. The idea behind

the Causal Analysis is that intention, consciousness and control do not matter for questions of authorship. Hence, among the appropriate ways of a thought's originating within a subject is a thought's simply popping into consciousness as a result from normal involuntary, automatic and subconscious cognitive processes within the subject's brain. Therefore, unsolicited thoughts are the subject's own thoughts.

I now turn to two kinds of cases that may seem to spell trouble for the Causal Analysis: perception and communication. In both cases, external objects play a salient role in the causal history of the thoughts and it may therefore appear that the thought has its causal origins outside the subject.

Suppose that, upon seeing a tree, I think 'this tree would provide enough firewood for the whole winter'. It may be said, in some sense, that the tree caused my thought. Does the Causal Analysis imply that the tree is the author of the thought then (cf. Vosgerau & Voss, forthcoming)? Of course not. Even though the tree plays an important role in the causal history of the thought, it is still me who is the thinker of this thought.

In cases of visual perception, the causal story is roughly this: light coming from the tree stimulates photoreceptive cells in my retina and the stimulus is processed in my brain. In the first step, there is a perception. Now, the neural processes that constitute the perception may trigger further neural processes which eventually constitute a thought. This is the kind of causal origination that is relevant to authorship of thoughts. It happens in my brain, in the right kind of way, and that is what makes the resulting thought mine.

Turning to communication, suppose that, upon your telling me that the lecture is at 4 pm, I entertain the thought 'The lecture is at 4 pm'. Some may hold that in this case, intuitively, the thought is yours rather than mine, or—to employ a terminology that proves particularly misleading in this case—that you, rather than I, are the *author* of the thought (see e.g. Vosgerau & Newen 2007: 37, Vosgerau & Voss, forthcoming). Now, in an everyday sense of 'authorship' it is indeed the sender who normally is the author of a communicated thought. However, this everyday notion of authorship has

largely to do with the thought type, i.e. with the content of the thought. Yet, what is disowned in thought insertion is not the thought type, but the token thinking episode. Hence, in the relevant sense of 'authorship', the receiver is the author of a token thinking episode resulting from communication.[71]

Vosgerau & Voss are well aware of the type/token distinction. Why do they find, then, that in communication the sender is the author? First, they seem to assume that that the everyday notion of authorship is informative for the analysis of thought insertion. The idea is that we do not need a special technical notion, but can just look at everyday cases of thought ascriptions. My assumption, in contrast, is that subjects of inserted thoughts want to express a bizarre and strange thing and that the everyday notion of thought ascription does not capture what they try to say. Secondly, they assume that this everyday notion of authorship sometimes refers to token episodes, for instance in ascribing communicated thoughts. All that being said, one may still worry that their intuition about the sender being the author is fueled by a notion of authorship that pertains to thought types. Otherwise, it would not be clear how to make sense of their further claim that in cases of misunderstanding the sender is *not* the author of the resulting thought. But the thought type (i.e. the content of a thought) should not play a role in the analysis of authorship in thought insertion. Hence, I do not think that the intuition about the sender being the author should inform the analysis of authorship in thought insertion.

Let me back up this claim by appeal to some remarks from the literature that deal with the question of authorship in communication. Campbell illustrates the irrelevance of content with the example of a secretary taking down notes at dictation. Although the writer does not determine the content, the written words are his in the sense that he produces them (cf. 2002: 37). It also does not matter that the sender in communication has the intention to cause certain thoughts in the receiver. According to Campbell, reports of thought insertion claim "something more immediate, where the intention [getting someone to think a particular thought] is executed

[71] For the type/token distinction see Vosgerau & Voss (forthcoming).

without going by way of any instrumental process [such as communication].” (1999b: 620)[72]

Similarly, Shaun Gallagher remarks that “when someone suddenly starts shouting instructions at me and causes me to start thinking […] I do not attribute my thoughts to someone else” (2000: 215). He defuses the intuition that communicated thoughts are the sender’s as follows: “it is possible to attribute thoughts in my mind to someone else, but in a very ordinary way–for example, in listening closely to a speaker, one might say that the speaker’s thoughts are being inserted into one’s mind” (ibid.: 233, fn. 12). The idea is that there may well be a notion according to which communicated thoughts are the sender’s, but that this ordinary notion is not adequate in interpreting the disowning claim in thought insertion.[73]

Now, how does the idea that the receiver is the thinker of a communicated thought mesh with the Causal Analysis, given that the sender plays an important causal role in the coming about of the receiver’s thought? The fact that receiving and processing information happens largely automatically conceals the fact that substantial work is done in communication by the receiver. As Gallagher puts it, although “someone else is causing me to think of certain things […] as I listen to a lecture […] my agency is involved in actively listening to the other person” (2004b: 91). This ‘agency’ involved in receiving and processing information is more visible when the receiver is not proficient in the language of communication or in cases of misunder-

[72] The view expressed by Campbell here is in tension with the Personality Analysis attributed to him above. I believe this tension is genuine to his writing and symptomatic of the fact that the notion of authorship has not been developed in sufficient detail.

[73] In the quoted passages, Gallagher is not explicitly addressing the conceptual question, but the phenomenal question. His point is that the *experience* of inserted thoughts is not to be understood simply as the experience of unintended thoughts. However, his observation can naturally be extended to the conceptual question: The *claim* of inserted thoughts is not to be understood simply as the claim of unintended thoughts.

standings. Now, it may be questioned whether my hearing you speak can reasonably be labelled agency. But that is a merely terminological question. The point is that in communication the receiver plays an important role and the notion of causal origination is intended to track this aspect.

This point can also be applied to some rather special instances of communication: it allows to make sense of a very broad notion according to which even thoughts resulting from brainwashing or subliminal priming are the subject's own (cf. Young 2006: 824). For, even in these cases there is cognitive work to be done on the subject's part for the stimuli to result in thoughts. Although the stimuli and the processing do not rise to consciousness, the resulting thoughts are the subject's own in the sense that they causally originated in the appropriate way within the subject.

It may seem odd that among the ways of causal origination that make a thought one's own we should also include subliminal priming and brainwashing. After all, these cases do involve a sort of manipulative external influence and one may be reluctant to attribute these thoughts to the subject. My reason for attributing authorship in these cases is to make room for the distinction between thought insertion and influenced thinking. According to the psychiatric literature, subjects who report influenced thoughts accept that they are the ones who think these thoughts, but claim to be in some sense manipulated. The sense in which influenced thoughts are disowned is the sense in which a brainwashed subject's thoughts are not hers. In contrast, subjects of inserted thoughts claim not to be thinking these thoughts at all.[74]

Addressing Objections

I now address four worries that may remain. First, someone may object that my presentation of the Causal Analysis is not very informative and leaves to be desired a general idea of how a thought has to be caused in order to be one's own, i.e. a general idea of how to spell out the appropriateness

[74] See e.g. Stephens & Graham 2000: 120f., Mullins & Spence 2003: 294f., Gallagher 2004b: 95f.

condition. I admit that a lot more needs to be said on authorship and that one can come up with a number of unclear cases. For instance, what about thoughts that result from a brain tumor or from the influence of drugs? Do these thoughts causally originate within the subject in the appropriate way?

Even if my discussion does not provide an answer regarding these borderline cases, it contributes to a better understanding of authorship in several ways. The examples discussed above give us sufficient grip on what it means for a thought to causally originate in the appropriate way within a subject to answer the main question, namely the question whether inserted thoughts are the subject's own. And while I haven't spelled out exactly what the appropriateness consists in, several requirements became clear in the discussion. The internal/external distinction by itself is not sufficient as the cases of perception and communication show. Rather, a subject lacks authorship when external causes are involved in a particularly deviant or direct manner. A promising way of spelling out the required relation might be to say that the thought must not be directly caused by an external event, but be the product of other mental states of the subject.[75] More importantly, as the arguments against the Agency Analysis have shown, whether a thought causally originated within the subject in the appropriate way does not depend on the thought being the product of intentional, controlled, or conscious processes. Given this rough characterization, we can already see that the proposed analysis, unlike competing approaches, promises to accommodate all the paradigm cases and provide a plausible classification of the unclear cases.[76]

Second, someone may take issue with the fact that, according to the Causal Analysis, probably every existing thought is authored by its subject; non-

[75] See e.g. Campbell's idea of unmediated influence (1999: 620, and 2002: 36).

[76] Spelling out what appropriateness consists in is a challenge not just for the Causal Analysis of mental agency, but for causal theories in general. See e.g. the causal theory of perception (Grice 1961) or the causal theory of action (Davidson 1963). Even after decades of discussion, proponents of these theories have not come up with precise criteria, but rely heavily on examples.

authored thoughts virtually do not exist. Doesn't the Causal Analysis thereby forfeit the distinction between authorship and ownership (cf. Vosgerau & Voss, forthcoming)? The simple answer is: no. Even if *de facto* ownership and authorship never come apart, the Causal Analysis allows for the conceptual distinction. The fact that this distinction expresses a divergence which–even if conceptually possible–never actually takes place only underwrites the bizarreness of what schizophrenics are claiming. If the coming apart of ownership and authorship were an everyday affair, instantiated in communicated and perceptual thoughts, then disowning claims would be substantially trivialized.

Third, it may be objected that I am imputing a complicated view or a particular metaphysical belief about thought causation to the subjects who make the disowning claims. This is not the case. Any account of what it is to be the causal originator of a thought will depend on metaphysical assumptions about the mind. A naturalist will have a very different picture of the causation of mental events than, say, a substance dualist. My discussion assumed a naturalist view of the mind, according to which thoughts are realized in the brain in having neural correlates. The neural correlates do not cause thoughts, but rather constitute thoughts. While the Causal Analysis is equally compatible with dualism, a different story would have to be given to explain in dualist terms what it is exactly to be the cause of a thought. I offered an interpretation of disowning claims that renders them conceptually coherent. This does not imply that subjects of inserted thoughts believe or want to express *all that is said here*.

Finally, it may seem that the Causal Analysis faces another objection from the idea that causation simply is not accessible to introspection and that the Causal Analysis therefore implies that authorship is not accessible to introspection. A point along this line is raised by Coliva. She considers, but then rejects a view that comes very close to the Causal Analysis: "X is the producer of a given token thought if and only if X is the proximal *physical* cause of that token thought." (2002b: 43) Her objection is that the property of authorship, so construed, is not a property that can be known to apply introspectively.

> [A] thought one is introspectively aware of does not seem to disclose its proximal physical cause. [...] A fortiori, the mere having of a thought in one's stream of consciousness cannot manifest the individual to herself as its proximal physical cause. (ibid.)

However, Coliva's actual target is not the Causal Analysis, but the idea that subjects of inserted thoughts could be considered rational in disowning a thought. She continues:

> So, even if the notion of being the producer of a given thought so understood makes sense from a third-person perspective, it is not a notion that, as such, can be used to make rational sense of the relevant reports from the *first-person perspective*. For it is just not a notion the warranted application of which can be licensed by the subject's own experience. (ibid.)

According to this argument, it is not the disowning claim per se that is irrational or even incoherent. Rather, it is claiming to be *introspectively aware* of not being the proximal physical cause of a thought which is irrational. If I understand Coliva correctly, the introspection-based disowning claim is about as irrational as claiming that one is introspectively aware of, say, being the tallest person in the room.

Although I find Coliva's position debatable, I will not engage with it here, for it is not relevant to the Causal Analysis. Coliva targets the idea that introspection-based authorship-ascriptions, construed as ascriptions of physical causation, are, as it is often put, *rational* reactions to fundamentally strange experiences. But this project is not part of the Causal Analysis. The main point of the Causal Analysis is to make sense of one side of the ownership-authorship distinction. This distinction, in turn, is needed to solve what I called the conceptual puzzle: Is it even conceptually possible to have a thought that is not one's own? The suggested view provides an interpretation of the disowning claim that is conceptually coherent. The view does not imply that subjects are rational or even justified in holding these beliefs.

4.6 Summary

Let me sum up the discussion of metaphysical authorship. In my discussion of the Agency Analysis, the Personality Analysis, and the Rationalist Analysis, two general points emerged. First, it is important to clearly separate questions regarding authorship from questions regarding the sense of authorship. The Agency Analysis seems to result from drawing the wrong contrast, the contrast between thoughts that paradigmatically involve a sense of intentional agency and thoughts that don't (directed thoughts vs. unsolicited thoughts). To analyze the notion of authorship, we need to contrast paradigmatically authored with paradigmatically non-authored thoughts.

Second, authorship is logically independent from the thought's content. In paradigm cases of authored thoughts, a subject can deliberately entertain thoughts of any content; in paradigm cases of non-authored thoughts, thoughts of any content could be inserted in a subject's mind. Therefore, the content of the thought does not play a role in the analysis of authorship. In particular, it does not matter whether the thought fits the subject's occurrent intentions or her background psychology, and it does not matter whether the thought is endorsed by the subject. While the thought's content may influence whether a thought is *experienced* as one's own, it does not bear on its *being* one's own.

I argued for the Causal Analysis which holds that a thought is the subject's own thought if and only if it causally originated in the appropriate way within the subject, which is more or less to say that the thought causally originated within the subject's nervous system. This means that the disowning claim in thought insertion has to be understood as the claim that the thought did not causally originate within the subject's nervous system. As a matter of fact, inserted thoughts do, of course, originate within the subject and are hence the subject's own thoughts.

This result underwrites my discussion of thought insertion in § 3.1, which was based on the assumption that disowning claims and external attributions are mistaken. I argued, however, that thought insertion does not

undermine Immunity since external attributions are not self-ascriptions and disowning claims, construed as negative self-ascriptions, do not involve an error through misidentification.

Nonetheless, given a causal account of authorship, it is clear that introspection-based self-ascriptions are liable to error through misidentification. Even if thought insertion itself does not refute Immunity, it leads the way to imagining a counterexample. Very simply, the cases of truly inserted thoughts refute the idea that self-ascriptions of authorship are logically immune to error through misidentification. That is to say, it is conceivable that someone is introspectively aware of a thought which causally originated within a different subject (think, for instance, of telepathy). Thus, self-ascriptions of authorship are at most *de facto* immune to error through misidentification, given that telepathy and the like are not possible. We can assume that this holds not just for introspection-based self-ascriptions of authorship, but that it holds in general for introspection-based self-ascriptions of factive mental states and self-ascriptions of physical relations (in contrast to self-ascriptions of, e.g., phenomenal experiences). I now turn to a refined critique of Immunity which aims to show that self-ascriptions of factive and physical states do not even enjoy *de facto* immunity.

5. A Refined Critique of Immunity

My defense of Immunity in § 3 stands on broadly two legs. I have argued that judgments in the pathological cases either do not fall within the scope of the thesis, or, if they do, do not actually involve an error through misidentification. In this chapter, I present a refined critique of Immunity which challenges both pillars of my defense.

Before entering the discussion, a general remark on dialectics is in order. Whether the refined critique is successful depends essentially on the question how one construes the Immunity Thesis. Given certain construals, the pathological cases challenge Immunity, given other construals, they miss their target. Now, there is not *the one correct* way to construe the thesis. What I am searching for, then, is the construal of the thesis that has the best chances of being true. Or, as Peter Langland-Hassan puts it in challenging the strongest version of Immunity: "my target is the more cautious (and more attractive) version of Introspective Immunity" which is restricted in scope to "the kinds of judgments that people are most likely to find absolutely [rather: logically] immune to error through misidentification" (forthcoming: 6). I will argue that on the epistemic approach the scope of the thesis can be restricted in a way that allows the thesis to be defended against the pathological cases. I will argue that these restrictions are not *ad hoc* moves to fend off counterexamples but can be motivated independently of the counterexamples as a reasonable interpretation of Immunity. However, it will also become clear that the counterexamples do make interesting and important points against the epistemic version of the Immunity Thesis.

As a quick reminder, on the epistemic approach, Immunity is closely tied to the idea of epistemic identification-freedom. On this view, a judgment is in error through misidentification if and only if it is epistemically based on a false identification assumption, and a judgment is immune to such error if and only if it is not epistemically based on any identification assumption. On the ontological approach, in contrast, a judgment is in error through misidentification if and only if the source of the predication information is

different from the object to which the predicate is applied; a judgment is immune if and only if such divergence is not possible.

The refined criticism that I will present consists of three arguments against Immunity. First, I explore the idea that Immunity is not restricted to self-ascriptions, so that external attributions do fall within its scope after all. Second, I present a different way of construing disowning claims, a way in which they are in error through misidentification after all. Finally, I discuss the idea that the pathological cases undermine Immunity indirectly, namely by showing that introspection-based self-ascriptions are identification-dependent. I argue that there is a plausible version of Immunity that can be defended against all three arguments.

But before I present the three arguments, I deliver an argument for one crucial assumption that has been made in § 3, namely the assumption that disowning claims, construed as negative self-ascriptions, satisfy the self-ascription constraint. The discussion of this argument, which really is a discussion of the self-ascription constraint, makes for a perfect introduction to this chapter as the first two arguments of the refined critique also hinge essentially on the question how to understand the self-ascription constraint.

5.1 Negative and Positive Self-Ascriptions

The pathological cases involve two different kinds of judgments: disowning claims and external attributions. I argued that external attributions fail as counterexamples since they do not satisfy the self-ascription constraint (but see § 5.2). In this section, I focus on disowning claims. They were involved in all cases either explicitly or implicitly: 'I am not the author of this thought', 'I am not the agent', 'It's not my intention', or 'I do not feel this touch'. In the discussion above, I simply assumed that disowning claims, contrary to first appearances, do satisfy the self-ascription constraint. But do they really? After all, a property is denied rather than attributed to oneself. In this section, I defend my assumption.

Now, there is no question that disowning claims differ in their surface structure from the kind of self-ascriptions that are paradigmatically

immune to error through misidentification. The structure of the disowning claims' natural linguistic expression is such that subjects deny having a certain property rather than ascribe a property to themselves. In contrast, in paradigmatic cases of immunity to error through misidentification subjects ascribe properties to themselves rather than deny having them. To rehearse, I have labelled judgments of the form 'I am F' *positive* self-ascriptions and judgments of the form 'I am not-F' *negative* self-ascriptions. Given that all paradigmatic examples of immune judgments are positive self-ascriptions, it is quite tempting to construe Immunity as a thesis solely about positive self-ascriptions. If the thesis is construed this way, the negative self-ascriptions we find in pathological alienation do not fall within the scope of the thesis and hence fail as counterexamples.

Here are two instances from the literature in which such a move could be seen. Langland-Hassan mentions thought insertion and FB's case of somatoparaphrenia as cases that fail to target Immunity.

> Whatever we make of these cases, they are not counterexamples to any version of IEM that is relativized to self-ascriptions using the first person pronouns. For the misidentifications occur when the introspectively detected state is ascribed to someone *other than* the person making the relevant judgment and ascription. (forthcoming: 7)

I do agree with Langland-Hassan that the external attributions involved in the two cases fail to challenge Immunity. But his general claim that these two cases are not counterexamples against Immunity ignores the disowning judgment that is explicitly present in thought insertion and arguably is also implicitly present in FB's attribution of the hand. Possibly, the disowning judgments have been overlooked due to the very fact that they involve negative self-ascriptions.[77]

[77] To be fair, possibly Langland-Hassan does not consider disowning judgments for the same reason that I reject them for, namely that they do not involve misidentification. This could be thought to be implicit in the remark that "the misidentifica-

As another instance, consider de Vignemont's claim that Immunity is a claim about *false positives* only, not about *false negatives*.

> There is a *false negative* if one does not self-ascribe properties that are instantiated by one's own body. [...] There is a *false positive* if one self-ascribes properties that are instantiated by another individual's body. The hypothesis of IEM clearly concerns false positives. (de Vignemont 2012: 229)

To be fair, her aim in this passage only is to exclude external attributions from the scope of Immunity (personal correspondence). But the way she puts the distinction indeed suggests that the types of beliefs I call negative self-ascriptions do not fall within its scope. Putting it slightly differently, de Vignemont's distinction overlooks that there are two importantly different kinds of negative judgments, i.e. judgments in which "one does not self-ascribe properties that are instantiated by [oneself]". I can make a false negative either by claiming that the property in question, which really is mine, is not mine (disowning claim), or by claiming that the property in question is somebody else's (external attribution). Both judgments are false negatives in the sense that the subject makes a mistake in not self-ascribing a property that really is his own. While de Vignemont is completely right that Immunity is not about external attributions (but, again, see § 5.2), I think it would be wrong to generally exclude negative self-ascriptions from its scope (again, this is presumably not her aim in this passage, but rather an unintended implication of the way she defines false negatives).

So why should we consider negative self-ascriptions as falling within the scope of Immunity? Firstly, if one wanted to exclude negative self-ascriptions from the scope of the thesis, one would require a criterion to distinguish positive claims (i.e. claims to the effect that one instantiates a property) from negative claims (i.e. claims to the effect that one does not instantiate a property). The problem is that the grammatical structure of a thought's linguistic expression is not a good guide to whether a thought

---

tions occur [in the external attributions]". However, he does not explicitly address the matter of disowning claims.

involves a positive self-ascription or not. Consider a thought that could be expressed equally well by the sentences 'I don't want to go to sleep yet' and 'I want to stay up longer'. Grammatical structure suggests that the first sentence denies a property whereas the second sentence ascribes a property to the speaker. But it is not clear that this superficial difference tracks an interesting difference between two corresponding thoughts.

More importantly, it is not at all clear why there should be an epistemic difference between positive and negative self-ascriptions when these claims are based on the same kind of information channel. What matters regarding the scope of Immunity is not whether I ascribe or deny a property; what matters is whether I make a judgment about myself or about somebody else, and whether the judgment is based on first-personal grounds. The fact that all the paradigmatic cases are positive self-ascriptions is misleading. Consider Wittgenstein's examples for the use of 'I' as subject: "'*I* try to lift my arm', '*I* think it will rain', '*I* have toothache'" (1958: 66). The fact that properties are ascribed rather than denied is not essential to these cases. The point could be made equally well with the judgments 'I do not try to lift my arm', 'I do not think it will rain' and 'I do not have a toothache'. When these negative claims are based on introspection, there is no way that I could mistake somebody else's not trying to lift their arm for my own not trying, or somebody else's not being in pain for my own not being in pain. Hence, Immunity is not restricted to positive self-ascriptions, but to self-ascriptions (positively or negatively). In other words, the self-ascription restriction excludes those judgments that are about somebody else, but does not exclude negative self-ascriptions. In particular, it does not exclude judgments such as 'This is not my thought' or 'This is not my intention'. Although I will somewhat qualify this claim in § 6, that qualification will not affect the present argument.

5.2 The Argument from Subject-Neutral Immunity

In this section, I take a second look at the claim that external attributions do not fall within the scope of Immunity. I have argued above that the self-ascription criterion restricts the scope of Immunity to self-ascriptions. This

point may seem hardly disputable. After all, consider how Wittgenstein distinguishes two different uses of the word 'I' (1958: 66f.) and how Shoemaker's claim is about "error through misidentification *relative to the first-person pronouns*" (1968: 556; my emphasis). To say the least, this strongly suggests that a judgment falls within the scope of the Immunity Thesis only if the subject self-ascribes a property in a *de se* mode.[78] I will dub the Immunity Thesis, so construed, *First-Person Immunity*. First-Person Immunity is silent about judgments in which a property is attributed to another person. That is to say, the external attributions we find in pathologies of alienation do not challenge First-Person Immunity for the simple reason that they fall outside its scope. To challenge First-Person Immunity, one has to provide a case in which a mental state is self-ascribed rather than ascribed to someone else. But is it possible that critics of Immunity who suggest external attributions as counterexamples, such as Lane & Liang, have so crudely overlooked this restriction?

Certainly, the point has often been overlooked. Consider, for instance, Romdenh-Romluc (2013) who on the one hand explicitly construes the Immunity Thesis as applying solely to self-ascriptions, yet on the other hand discusses at great lengths a counterexample suggested by Jeannerod & Pacherie (2004) which involves an external attribution. Roughly, Jeannerod & Pacherie's argument is that in hearing voices, subjects misattribute their own intentions for inner speech to an external entity. But why does Romdenh-Romluc even consider this a promising counterexample, given that it clearly does not contain a self-ascription? She comes up with very elaborate arguments to defend Immunity, but seems to miss the very simple point that the misattribution of one's own intention to an external source is not a self-ascription and thus does not fall within the scope of Immunity to begin with.

---

[78] I should note that this way of putting the idea does not sit well with the expressivist approach to Immunity, often attributed to Wittgenstein. For, according to expressivism, sentences involving the word 'I' as subject do not ascribe a property to the subject, but are rather *expressions* of the mental states in questions.

A more charitable way of looking at criticisms such as Lane & Liang's or Jeannerod & Pacherie's is to say that they really had a version of the Immunity Thesis in mind which is not restricted in scope to self-ascriptions.[79] Let us dub this version *Subject-Neutral Immunity* (cf. Langland-Hassan, forthcoming: 6f.). It differs from the First-Person version solely in not being restricted in scope to self-restrictions, so all we need to do in terms of changing the definition is to drop the self-ascription constraint: Ascriptions of mental states that are based on first-personal awareness of these states are immune to error through misidentification. More intuitively put: when one is introspectively aware of a mental state, one cannot go wrong in ascribing that state to the wrong subject. Against this version of Immunity, the external attributions in the pathological cases make a much better case: they cannot be dismissed as failing the self-ascription constraint. (Whether they actually undermine that version of Immunity still depends on whether they satisfy the introspection constraint.) In other words, when Immunity is construed as a subject-neutral claim, external attributions fall within the scope of the thesis (granting that they satisfy the introspection constraint). Since they are in error through misidentification, the external attributions plausibly refute Subject-Neutral Immunity. I dub this the *Argument from Subject-Neutral Immunity*.

I will not further discuss whether the pathological cases do successfully undermine Subject-Neutral Immunity. Rather, I will discuss the question whether Immunity should be construed as a subject-neutral thesis or not. Concerning the dominant epistemic approach, we find a tension: What proponents of Immunity have in mind clearly is First-Person Immunity, but what they should claim, given their explanatory framework, is Subject-Neutral Immunity.

First, let me explain in more detail how the two variants differ. First-Person Immunity is the claim that introspection-based *de se* self-ascriptions of

[79] Arguably, this holds also for Marcel's criticism, even though his case of anarchic hand syndrome does not actually involve an external attribution. More on this later (§ 6.1).

mental states are immune. For the sake of perspicuity, let me simplify this claim a bit by assuming that a self-ascription would be in error through misidentification if and only if the property that is ascribed to oneself is in fact not one's own, but somebody else's. With this notion of misidentification, we can simplify the above definition. First-person Immunity is the claim that, when one self-ascribes a mental state based on introspection, then that state is one's own. Subject-Neutral Immunity, in contrast, is the claim that when one ascribes a mental state (to self or other) based on introspection, then that state is the person's state to whom it has been ascribed.

Now, Subject-Neutral Immunity can be understood as the conjunction of First-Person Immunity and what I would like to call *External Immunity*. The latter is simply the claim that if one ascribes a mental state to another person based on introspection, then that state is that other person's state. Construing Subject-Neutral Immunity as the conjunction of First-Person Immunity and External Immunity allows us to ask more precisely whether one should construe Immunity as a subject-neutral claim. The crucial question is whether, over and above First-Person Immunity, one should also profess External Immunity.

On the one hand, External Immunity is a pretty strange claim to begin with, one that Shoemaker and others plausibly did not have in mind. It is strange in that the antecedent specifies a condition that does not normally come about, namely being introspectively aware of a mental state but ascribing that state to some other person. In fact, Shoemaker thinks that it is impossible for the antecedent to be true. So, although presumably he could have accepted External Immunity as a vacuously true claim, this is certainly not an immunity claim that he had in mind when proclaiming the immunity of introspection-based self-ascriptions.[80] So, if Shoemaker and others had First-Person Immunity in mind, rather than Subject-Neutral Immunity, the Argument from Subject-Neutral Immunity misses its target in construing the Immunity Thesis in a deviant way.

[80] Thanks to Gottfried Vosgerau for helping me get this point straight.

On the other hand, it has to be admitted that, given the view that Immunity is closely linked to identification-freedom, one would actually expect proponents of Immunity to construe it as a subject-neutral claim. For, if the error in question is taken to be impossible because of the judgment's identification-freedom, it should be impossible in either direction (self-ascription and other-ascription). The idea one would expect is this: whenever a judgment is based on introspection it is identification-free, and whenever it is identification-free it is immune to error through misidentification. On this view, both self-ascriptions and other-ascriptions should be immune to error through misidentification.

Given the idea that immunity is equivalent to identification-freedom, it is not surprising that it is not only critics of Immunity who construe the thesis as a subject-neutral claim. Consider the following passage drawn from the introduction of a collection on self-consciousness.

> [T]here are ways of acquiring knowledge about one's properties that are *immune to error through misidentification relative to the first person pronoun*. What this means is that any knowledge about properties gained in these ways cannot be known to apply to an individual without one ipso facto knowing that they apply to oneself. (Eilan et al. 1995: 22f.)

Using the term 'knowledge' in explaining immunity is intuitive, but can also prove misleading.[81] What I take Eilan and colleagues to mean is that one cannot believe that an introspected property is instantiated without believing that it is instantiated in oneself. But this means that when a mental state is ascribed based on introspective awareness, one can neither go wrong in falsely ascribing someone else's mental state to oneself, nor can

[81] Due to the factivity of knowledge, for instance, one cannot *know* that one's own properties apply to somebody else, simply because they don't. What Eilan and colleagues are after, I take it, is rather the point that one cannot *believe* that one's own properties apply to somebody else.

one go wrong in falsely ascribing one's own mental state to someone else.[82] And that it is precisely what Subject-Neutral Immunity says.

I will argue in the next chapter that, given the theoretical framework of the epistemic approach, it would be coherent to construe Immunity in its subject-neutral variant, whereas, on the ontological approach, Immunity would be most naturally construed as a claim about self-ascriptions only. If this is correct, the Argument from Subject-Neutral Immunity presents a challenge to the epistemic approach, but not to the ontological approach.

5.3 The Argument from Disidentification

I argued in § 3 that disowning beliefs fail as counterexamples since they are not in error through misidentification. Subjects who disown a thought, intention, or sensation (construed as a negative self-ascription) do not make an error of ascribing a property to the wrong person, but rather make an error of saying something wrong of the right person. Hence, I argued, disowning beliefs are in error through mispredication, not in error through misidentification. I now discuss a rejoinder which suggests construing disowning beliefs in a different way, namely in a way so that they do involve misidentification.[83]

Someone Is F, It's not Me

In the above discussion, I construed the negation in the disowning claim to pertain, so to speak, to the predicate. Roughly, I construed the claim as having the structure '(~F)(*i*)' (I am not-F). I did so because this is the most obvious way in which the claim can be understood as a self-ascription in spite of disowning a property rather than positively self-ascribing a property. The idea was this: when a subject believes of herself that she is F, or that she is not-F, based on being introspectively aware of F-ness, or of not-F-ness, then she cannot be wrong about it being her that is, or is not, F.

[82] Actually, the impossibility of the first kind of error is not explained by identification-freedom (see § 7.2).

[83] I am greatly indebted to Daniel Stoljar for raising this argument.

This is why we should consider the Immunity Thesis to hold also for what I called negative self-ascriptions.

The objection I want to consider now holds that the negation in the disowning claim pertains to the subject, rather than to the predicate. Roughly, the disowning claim is suggested to have the structure 'F(~*i*)' (roughly: it is not me who is F).[84] This construal seeks to capture the intuition that the disowning claim involves a misidentification after all. The basic idea is this: what the subjects are really saying, when they disown a mental state, is that there is the mental state in question, but that it is not theirs. In a nutshell, the claim 'I am not the author of this thought' has to be understood as the claim 'Someone is thinking this thought, but it's not me'. Mutatis mutandis for the other cases. There is an intuitive sense in which the disowning claim, so construed, is in error through misidentification. The intuitive idea is that subjects get the predication aspect right, but get the subject aspect wrong: They are right that someone is thinking the thought, they are wrong that it is not them. I will call this type of mistake an *error through disidentification*.

Consider the following analogy. Suppose I wonder whether I should have a bite to eat. I feel inside, so to speak, and come to believe that I am not hungry. However, in fact I am hungry. In this case, my judgment 'I am not hungry' is clearly not in error through misidentification. I am simply wrong about whether I am hungry or not. In the previous discussion, I construed disowning claims in the pathological cases analogously to that judgment. The present approach asks us to construe the disowning claim rather on the following model: Suppose, again, that I really am hungry and that I also feel hungry. However, for some strange reason I believe that it is not my hunger which I am feeling. In other words, I come to believe that I am not the hungry one, or, that there is hunger, but that it's not mine. Let us call this type of judgment a *disidentification*. This judgment certainly is not in error

[84] I was told that, as far as logic notation is concerned, denying the object doesn't make any sense. I believe that the idea is clear anyhow.

through mispredication. After all, I am not wrong about there being hunger. Is it a case of misidentification then? This will have to be discussed.

Before I asses how this way of construing disowning claims bears on Immunity, let me note that the model does seem to capture very well what is going on in the pathological cases. With respect to thought insertion, the idea is this: Subjects of inserted thoughts are aware of a thought that appears to be somebody's thought (in the sense of authorship), but come to believe somehow that it is not theirs. When I say that the thought appears to be *somebody's* thought, all I mean to say is that the thought appears to have an author, for instance in the sense that the thought appears to express somebody's intentions. This does not imply that the thought appears to be some *particular* person's thought. (See e.g. Stephens & Graham 2000: § 8.3.)

Relatedly, the demonstrative nature of disidentifications also matches the judgments we find in the pathological cases. Obviously, the judgment 'I am not-F' cannot actually be justified solely on being aware of someone's being F and believing of oneself that one is not the witness of that property. The belief that this particular instantiation of F-ness is not mine does not tell me anything about whether I am F or not. This fits the fact that the judgments we find in the pathological cases are best understood as demonstrative judgments which disown a particular mental episode. So the correct analogy to pathological disowning claims is the demonstrative claim 'this (particular hunger) isn't my hunger' rather than the general claim 'I am not hungry'.

The second part of the refined critique is what I dub the *Argument from Disidentification*. According to this argument, disowning claims, construed as disidentifications, show that introspection-based self-ascriptions are liable to error through misidentification. There are two crucial assumptions behind this argument. First, it must be assumed that an error through disidentification is a form of error through misidentification. Second, it must be assumed that judgments of the form 'F(~*i*)' are self-ascriptions. I will address both issues in turn.

Error through Disidentification vs. Error through Misidentification

On the present approach, disowning claims are construed as claims of the form 'It is not me who is F'. More generally, they have the form 'It is not *a* who is F', which is importantly different from the judgments discussed thus far that all had the form '*a* is F' or '*a* is not-F'. On this view, disowning claims can be understood to be composed of, or based on, the two beliefs 'someone is F' and 'it's not me', where the 'it' is anaphorically linked to 'someone'. The latter belief really is the negated identification 'I ≠ the person who is F'.

To illustrate, consider an adapted version of FORREST. Suppose, again, that I see Tom running, but do not recognize him. To turn this into a case that involves a disidentification, suppose further that, rather than mistaking him for Forrest and coming to believe that Forrest is running, I neither recognize him as Tom, nor mistake him for someone else, but simply come to believe 'this isn't Tom running'. Intuitively, the error I just made is quite similar to the error I made in the case in which I judged 'Forrest is running' and one may therefore want to say that it is an error through misidentification.

The difference between the new case and the standard case is this. In the case of error through disidentification, we have a judgment's being based on a falsely *negated* identification, in the standard case of error through misidentification, we have a judgment's being based on a falsely *affirmed* identification. The question then is whether disidentification is a distinct phenomenon or whether disidentification is really just another kind of misidentification.

Before I address the classificatory question, note that corresponding to the distinction between *de re* and which-object misidentification, we can distinguish two types of disidentification. A judgment can be based on a falsely negated *de re* identification, as in the case above where my judgment is based on the belief 'this man ≠ Tom'. But a judgment can also be based on a falsely negated singling out identification. Consider, an adaptation of SKUNK, where I see the skunk that is actually responsible for the smell in my

garden, but, due to its unusual coloring, think that it is not a skunk, and therefore falsely judge 'this animal is not the skunk in my garden'.

Now, on the one hand, a number of intuitive points can be made to say that disidentification is really just another form of misidentification. The error pertains to the identification component (which involves a negation, all right, but still), rather than to the predication component. The subject is right about somebody being F, but wrong about who it is that is F. Further, the sensibility of the who-question plausibly serves as an indicator of immunity. Judgments of the form '*a* is F' are typically immune to misidentification when it does not make sense to ask 'Are you sure it is *a* who is F?' Analogously, judgments of the form 'it is not *a* who is F' can be expected to be immune to disidentification when it does not make sense to ask 'Are you sure that it is not *a* who is F?'.

On the other hand, it could be stressed that disidentification really is a distinct phenomenon in that it does not involve the ascription of a predicate to an object. It is a fundamentally different matter whether an object is identified as the witness of a property or whether an object is excluded as a possible witness of the property. As far as I know, the literature on immunity does not contain a single case that involves a disidentification (except, of course, the pathological cases of alienation).

It is difficult to find a neutral criterion to decide the matter. Luckily, for the purposes of this discussion we do not need a neutral criterion. What matters, in the end, is whether the Immunity Thesis is construed in a way that rules out errors through disidentification. This question can be answered independently of the classificatory question. Let me explain what I mean by that. If we assume that disidentification is *distinct* from misidentification, it is still an open question whether Immunity should be understood to rule out both types of errors or only error through misidentification. If, in contrast, we assume that disidentification is *not distinct* from misidentification, that is if we assume that disidentification is a kind of misidentification, it is still an open question whether Immunity should be understood to rule out both kinds of misidentification or only the positive

kind. I therefore suggest to ignore the classificatory question as pertaining largely to terminology and to move on to the more fundamental question, the question whether Immunity should be construed to rule out error through disidentification or not.

What we have is two options of construing Immunity: in a narrow sense as ruling out only error through positive misidentification, or in a wider sense as also ruling out error through disidentification. I will argue below that, on the epistemic approach, we should construe Immunity more widely as a claim that rules out disidentification, and that, on the ontological approach, we should construe Immunity more narrowly, as not ruling out disidentification. This means that the Argument from Disidentification spells trouble for the epistemic approach, but not for the ontological approach.

Error through Disidentification and Self-Ascription

In § 5.1 I argued that disowning claims, construed as negative self-ascriptions, satisfy the self-ascription constraint. I now ask whether disowning claims also satisfy the self-ascription constraint when construed as disidentifications.

When I argued that negative self-ascriptions satisfy the self-ascription constraint, my idea was that the negation of the predication aspect does not matter to the immunity of the judgment in question. The judgment 'I am in pain', based on one's awareness of pain, should not be considered any different with respect to immunity than the judgment 'I am not in pain', based on one's awareness of being pain-free. For this idea to apply to the disowning claim, it is crucial to construe it as having the structure 'I am not-F'. Since on the present construal disowning involves a negated identity claim, we have to ask anew whether the disowning claim satisfies the self-ascription constraint. The question is: are claims of the form 'It is not me who is F' self-ascriptions?

Just like the question whether error through disidentification is a kind of error through misidentification, the question whether a disidentification is a self-ascription proves difficult to answer. It seems to me that disidentifi-

cations lie between two clear cases. On the one hand, we have self-ascriptions proper. They have the form 'F(*i*)' and clearly satisfy the self-ascription constraint. On the other hand, we have external attributions. They have the form 'F(*a*)' (where $a \neq i$) and clearly fail the self-ascription constraint. In between, we have disidentification claims. They have the form 'F(~*i*)' and it is quite unclear whether we should consider them self-ascriptions or not. On the one hand, they do concern the self in that the subject says something about herself. This would speak in favor of counting them as self-ascriptions. On the other hand, what the subject actually says of herself is precisely that it is *not she* who instantiates the property in question. In this respect, disidentifications are much more similar to external attributions than to self-ascriptions. Disidentifications, in a sense, imply a corresponding external attribution (they imply that someone other than oneself is F) and rule out the corresponding self-ascription.

Again, I want to suggest that we can ignore the mainly terminological question, whether claims of the form 'it is not me who is F' are properly described as self-ascriptions or not. Rather, the crucial question is, again, whether we should understand Immunity as a claim that rules out error through disidentification or not. That means that we have again two options of construing the thesis. Construed more narrowly, the self-ascription constraint is satisfied only if a subject claims of herself that she has a property (or lacks a property). On this view, the thesis does not apply to disowning claims and hence the pathological cases miss their target. Construed more broadly, the self-ascription constraint is also satisfied if a subject says of a particular instantiation of a property, that it is not she who instantiates it. On this view, the thesis does apply to disowning claims and the pathological cases hence present a serious challenge[85]. And, again, I will argue that the epistemic approach lends itself more naturally to the broader interpretation which is vulnerable to the Argument from Disidentification,

[85] Again, whether they actually present successful counterexamples depends further on whether they satisfy the introspection constraint.

whereas the ontological approach is most naturally construed as the more narrow claim that is not subject to this type of counterexample.

5.4 The Argument from Identification-Dependence

The Argument from Subject-Neutral Immunity and the Argument from Disidentification are both attempts to provide direct counterexamples to the Immunity Thesis. Both arguments presuppose an approach which ties immunity to identification-freedom. As a third part of the refined critique, I now turn to an argument that targets this broader idea. This argument is compatible with granting, pace the previous two arguments, that the pathological cases do not directly refute Immunity. Rather, it urges that the pathological cases provide an indirect argument against Immunity. The pathological cases show, so the main idea, that introspection-based self-ascriptions are not identification-free. I dub this the *Argument from Identification-Dependence*.

Very intuitively, the idea is that there seems to be some kind of identification mistake involved in the pathological cases after all. The mistake in question is that subjects are introspectively aware of a thought, an intention, or a sensation, but deny that it is theirs. But this contradicts the dominant explanation of Immunity, which assumes a close connection between Immunity and identification-freedom. Introspection-based self-ascriptions of mental states are assumed to be identification-free in the sense that subjects do not need to figure out whose mental states they introspect. Rather, the mental states are necessarily presented in introspection as one's own. This is what is typically assumed to guarantee that introspection-based self-ascriptions are immune. (I will say a lot more on this shortly.)

The pathological cases show that there is something wrong with this picture. Patients do go wrong in ascribing mental states of which they are introspectively aware to the wrong person. Doesn't that show that introspection-based self-ascriptions of mental states are not identification-free after all? And if so, doesn't it follow that these self-ascriptions are not

immune either? An argument along these lines is suggested for instance by Gallagher:

> If Campbell is right that something like schizophrenic experiences of thought insertion violate the immunity principle, then the claim is more serious than simply finding a counterexample or an exception to the rule. It would involve admitting (in contrast to Shoemaker's characterization) that first-person self-awareness as subject does involve identification, that schizophrenics get it wrong and that normal subjects get it right. So if the immunity principle is subject to exception in the case of schizophrenia, then the principle itself is threatened. (Gallagher 2000: 208)

Note that, even if the scope of the Immunity Thesis can be restricted in the way I suggested in the two preceding sections, that is, even if the pathological cases do not directly refute Immunity, the challenge pointed out by Gallagher can be sustained. The challenge is this: the pathological cases show that introspection-based self-ascriptions of mental states are not identification-free after all. On the view that identification-dependence entails liability to error through misidentification, this goes to show that introspection-based self-ascriptions are principally not immune to error through misidentification. It could be maintained, at most, that misidentifications happen only under pathological circumstances and that as a matter of fact we normally get the identification aspect right.

It should be obvious that also this third part of the refined critique targets the epistemic version of Immunity only, not the ontological version, for this version simply does not appeal to the idea of identification-freedom. Thus, my claim is that all three arguments of the refined critique challenge the epistemic approach to immunity, the one that ties immunity to identification-freedom, but do not challenge the ontological approach. This claim shall now be defended.

6. Two (and a half) Approaches to Immunity[86]

In the previous section I have presented three arguments against the Immunity Thesis which build on the pathological cases in a more sophisticated way: the Argument from Subject-Neutral Immunity, the Argument from Disidentification, and the Argument from Identification-Dependence. I now show that these arguments do challenge Immunity on the epistemic approach, but not on the ontological approach. To do so, I will have to present the differences between the two approaches in more detail.

The epistemic approach and the ontological approach are based on two different notions of what it means for a judgment to be in error through misidentification. On the epistemic approach, error through misidentification is construed as a judgment's being epistemically based on a mistaken identification.[87] On the ontological approach, error through misidentification is the divergence of source and target (i.e. the divergence of the object from which the predication information derives from the object to which the property is ascribed). Accordingly, the Immunity Thesis of the epistemic approach (henceforth: *Epistemic Immunity*) is different from the Immunity Thesis of the ontological approach (*Ontological Immunity*).

Furthermore, the traditionally held explanatory background assumptions of the epistemic approach differ substantially from the explanatory assumptions that I will offer in support of the ontological approach. To wit, Epistemic Immunity is traditionally explained in terms of identification-freedom. The idea is that if the judgment in question is not based on an identification, it cannot involve any identification errors, but if a judgment is based on an identification, it is necessarily open to error through misidentification. Ontological Immunity, in contrast, must be explained in

[86] Parts of this chapters appear also in my paper "Immunity and Self-Awareness" which is under review at *Philosophers' Imprint*.

[87] A quick reminder: by identification I mean both *de re* identification and singling-out identification. Since the difference between the two types does not matter for the ensuing discussion I simply speak of identification to cover both.

terms of an introspection-ownership link. The idea is this: when a mental state is self-ascribed, based on one's introspective awareness of that state, the state actually is one's own state. This can be explained by the assumption that one has introspective access only to one's own mental states. If this is true, there can be no divergence of source and target and that is precisely what it means, on the ontological view, that these self-ascriptions are immune. In the following two sections I spell out these ideas in more detail.

6.1 Epistemic Immunity

On the epistemic approach, a judgment is in error through misidentification iff its justification contains a false identification component. For a judgment to be immune to such error means for the judgment to be identification-free.[88]

I have already presented the idea that Immunity is closely tied to identification-freedom in § 2.4. Here is how this all applies to introspection-based self-ascriptions of mental states. Witness Coliva:

> The reason why one cannot make an error through misidentification when one is self-ascribing a mental property on the basis of one's introspective awareness of that mental property is that, minimally, introspective awareness is a form of awareness that does not involve either observation or inference. [...] [T]he important point is that because the self-ascription is not based on the observation of oneself, then it cannot be grounded on any identification component and, therefore, it cannot be affected by EM. (2002a: 28)

In the light of the pathological cases one may now wonder: if introspection-based self-ascriptions are not based on observation, what are they based on then? The idea behind the epistemic approach is that introspection necessarily presents mental states as one's own. Shoemaker famously

[88] More precisely, perhaps, being immune means that a judgment is identification-free *or based on an a priori identification assumption* (cf. Coliva 2006: 424, fn. 37).

claimed that "in being aware that one feels pain one is, tautologically, aware, not simply that the attribute *feel(s) pain* is instantiated, but that it is instantiated *in oneself*" (Shoemaker 1968: 563f.). Let us call this claim *Self-Awareness*: If one is introspectively aware of a mental state then one is necessarily aware of that state as one's own state.[89] Evans professes a corresponding claim about the proprioceptive awareness of bodily states:

> There just does not appear to be a gap between the subject's having information (or appearing to have information), in the appropriate way, that the property of being *F* is instantiated, and his having information (or appearing to have information) that *he* is *F*; for him to have, or to appear to have, the information that the property is instantiated just is for it to appear to him that *he* is *F*. (1982: 221)

Self-Awareness is taken to guarantee that introspection-based judgments are identification-free and therefore immune to error through misidentification. Since one need not identify the owner of an introspected mental state one cannot make an error through misidentification.[90]

The pathological cases undermine this explanation. Subjects are introspectively aware of a mental state, but deny that it is their own state. Note that there are principally two ways to understand their mistake. That is to say, the error could be located in two different places. Either, Self-Awareness is false, which is to say that subjects are introspectively aware of a mental state without being aware of it as their own, or Self-Awareness does not guarantee identification-freedom, which is to say that being introspectively aware of a state as one's own is not sufficient for its correct self-ascription. In other words, the error can either be due to a disruption at the phenomenal basis, or due to a disruption concerning the move from the phenomenal basis to the doxastic level. In both cases, the pathological phenomena undermine the explanation of Immunity in terms of identification-freedom.

---

[89] The sense of 'one's own' (e.g. thought ownership vs. thought authorship) will be discussed below.

[90] See e.g. Shoemaker 1996: 196, Gallagher 2000: 205, Bermúdez 2003b: 217, and Smith 2006: 275f.

They show that there is some sense in which introspection-based self-ascriptions involve an identification. And if there is identification, so the traditional view, there is room for misidentification (see e.g. Shoemaker 1968: 561f.).

Self-Awareness and Alienation

Taking a closer look at the literature reveals that the critics of Immunity may actually intend to target Self-Awareness with their counterexamples, rather than Immunity. On that reading of their critique, they do seem to have a good point. It seems that subjects are aware of thoughts, intentions, or sensations, without being aware of these states as their own.

Consider again the case of FB. According to Lane & Liang's (2011) interpretation, she is introspectively aware of her tactile sensation, but misrepresents the sensation as her niece's sensation. Lane & Liang take this to show that FB is not aware of this sensation *as her own* sensation although she is introspectively aware of it. This would directly contradict Self-Awareness, which holds that introspective awareness of a sensation implies awareness of that sensation as one's own.

Although Lane & Liang nominally challenge the Immunity Thesis, they actually spell out their critique this way: "Our main thesis is: awareness that mental states are instantiated does not entail awareness that said states are instantiated in self." (2011: 83) Or, putting it negatively, they challenge the claim that "[e]very mental state is, from the first-person point of view, *represented as* experienced by the one who is introspecting the state" (ibid.: 87). If we follow these passages, they are actually criticizing Self-Awareness, rather than Immunity. Their point, then, is that FB is aware of a sensation, but is not aware of that sensation as her own.

I must at this point reiterate my reservations about Lane & Liang's interpretation of the case. According to the conservative interpretation, FB merely misrepresents *where* she (FB) feels the touch, not *who* feels the touch. On this interpretation, the case is perfectly compatible with Self-Awareness. Putting aside the question whether Lane & Liang's interpretation gets this

particular case right, it may sensibly be questioned whether their suggested interpretation is conceptually coherent to begin with. In what sense of 'ownership' is it possible for a sensation which one is experiencing to not be one's own? Certainly, it is possible that the causal source of the tactile experience is not located within one's own body. But that is explicitly not what Lane & Liang have in mind. Without a further explanation of the notion of ownership they have in mind (that notion of ownership, in which they take FB to disown the sensation), it is hard to even grant their interpretation for the sake of argument.

Let us turn to Marcel's critique from anarchic hand syndrome. Taking another look, we find again that he must be having Self-Awareness in mind, rather than Immunity. Just like Lane & Liang, Marcel nominally challenges Immunity, but actually describes a thesis very close to Self-Awareness in spelling out his target: "If one is aware through internal proprioceptive awareness of an action, of a posture, or of a sensation, one might think that it is impossible to be [phenomenally] mistaken about *whose* it is."[91] (Marcel 2003: 80) Marcel's argument then is that in anarchic hand syndrome the patient is first-personally aware of an action, but is not aware of that action as his own. If this were correct, anarchic hand syndrome would refute Self-Awareness. But as I have argued above (§ 3.2), subjects are not first-personally aware of the hand's action *qua* action. What they are aware of first-personally is merely the hand's movement. Hence, anarchic hand syndrome is not a counterexample to Self-Awareness either. My main point at this moment is, however, that it is Self-Awareness rather than Immunity which really is targeted by both Lane & Liang's as well as Marcel's criticism.

It is thought insertion and made impulses which prove most convincing as counterexamples to Self-Awareness. In these cases, it seems, subjects are

[91] Note how this is, so to speak, a subject-neutral version of Self-Awareness. Marcel does not claim that introspected states necessarily appear as one's own, but, more generally, that their appearance cannot be misleading with respect to the question whose states they are. (For the Argument from Subject-Neutral Immunity see § 5.2.)

introspectively aware of thoughts or intentions without being aware of these thoughts or intentions as their own.

Defending Self-Awareness

A simple way to defend Self-Awareness is to construe it as making a claim about thought ownership rather than thought authorship. Remember, I introduced Self-Awareness as the claim which holds that introspective awareness of a mental state entails awareness of that state *as one's own*. Put this way, the claim is, at least for some mental events, ambiguous with respect to the notion of ownership. For instance, talk of a thought's being one's own is ambiguous between the thought occurring within one's mind (thought ownership) and the thought being brought about by oneself (thought authorship). Hence, with respect to thoughts, Self-Awareness can be construed as claiming, either, that introspective awareness of a thought entails a sense of ownership for that thought, or that introspective awareness of a thought entails a sense of authorship for that thought. Since it is assumed that subjects of inserted thoughts do not lack a sense of ownership for inserted thoughts (in the technical sense specified), thought insertion challenges only that interpretation of Self-Awareness on which introspection entails a sense of thought authorship.

Concerning the two different interpretations, see also an exchange between Stephens & Graham and Gibbs.[92] Stephens & Graham argue that thought insertion does not undermine what they call the *inseparability thesis*. The inseparability thesis says more or less the same as Self-Awareness, namely that introspection is not separable from the sense of subjectivity. Here, the notion of subjectivity is again as ambiguous as the general notion of a mental state 'being my own'. Stephens & Graham understand the thesis as saying that introspection implies a sense of thought ownership (in their terminology, the question is whether the thought is experienced as occurring within the boundary of the self or within the subject's psychological history). Consequently, they argue that thought insertion does not challenge

[92] Stephens & Graham 1994, Gibbs 2000a, Stephens 2000, and Gibbs 2000b.

the inseparability thesis. Gibbs, in contrast, construes the inseparability thesis as saying that introspection (also) implies a sense of mental agency. Consequently, he argues that the inseparability thesis is undermined by thought insertion.

All that being said, the easily defendable variant of Self-Awareness, which holds that introspection implies a sense of thought ownership, is not the interesting one. After all, the immunity of ownership-ascriptions is not disputed, the immunity of authorship ascriptions is. As far as the discussion of the pathological cases is concerned, the version of Self-Awareness that pertains to the sense of authorship, rather than the sense of ownership, is the interesting one. Thus, I will from here on discuss the authorship/agency variant of Self-Awareness.

Another defense of Self-Awareness appeals to the distinction between phenomenal self-awareness and doxastic self-ascription. It could be objected that the cases of alienation primarily show that subjects *believe* certain states not to be their own, while Self-Awareness is most naturally construed as a thesis regarding the phenomenal experience. It could then be argued that alienated subjects actually do phenomenally experience the thoughts, intentions and sensations as their own, and only fail to self-ascribe them due to other factors, such as delusional beliefs, which override this phenomenal awareness.

It is hard to tell whether this response is empirically adequate. Marcel's description of anarchic hand syndrome certainly speaks against it. According to Marcel, subjects primarily lack the feeling of agency: they may even acknowledge that it is their action, but maintain that it doesn't feel like theirs (cf. 2003: 79f.). So, to say the least, the cases of alienation put a lot of pressure on Self-Awareness.

More importantly, even if Self-Awareness can be defended this way, the cases remain a challenge for the explanation of Immunity in terms of identification-freedom. For, they would still show that Self-Awareness does not guarantee identification-freedom; the transition from experiencing a state as one's own would not, as it were, automatically lead to the self-

ascription of that state. It is assumed that subjects of inserted thoughts do make a mistake in denying authorship and in attributing authorship to an external entity. If Self-Awareness can be defended as a solely phenomenal claim, this just goes to show that the mistake must come in on the doxastic level. That is to say, even if subjects are phenomenally aware of the thoughts as their own, this awareness must be somehow overridden so that they end up believing, pace their phenomenal awareness, that the thoughts are not their own. The phenomenal sense of authorship might, then, be but one factor in a process of self-ascription which is overall liable to error through misidentification (see e.g. the idea of a multifactorial weighting process in Synofzik et al. 2008).

Epistemic Immunity and the Refined Critique

Now, that the explanatory background assumptions of the epistemic approach are spelled out and that we have already gotten some idea of how the pathological cases serve to challenge this overall view, let us take another look at the three arguments raised in the previous chapter.

The Argument from Subject-Neutral Immunity essentially hinges on the question whether we should construe Immunity as a subject-neutral claim. Subject-Neutral Immunity, to repeat, is not restricted to self-ascriptions, but would apply equally to external attributions. Although I have shown in the previous chapter that this is not the kind of claim Shoemaker and others have in mind, it actually is the version of Immunity that would best fit their theoretical framework. The idea is quite simple: If Immunity is understood as a result of introspection being identification-free, one should not expect that the Immunity Thesis holds for self-ascriptions only, but one should expect that it holds for all introspection-based judgments. In particular, one should expect that it holds also for external attributions (even if vacuously so, that is even if it only holds because introspected mental states are never externally attributed). In the light of these considerations, it makes sense that Marcel spells out the target of his critique in subject-neutral terms: "If one is aware through internal proprioceptive awareness of an action, of a posture, or of a sensation, [...] it is impossible to be mistaken about *whose* it

is." (2003: 80) As I have argued above, Subject-Neutral Immunity is defeated by the external attributions in the pathological cases. The challenge to the epistemic framework then is this: restricting the scope of the thesis to self-ascriptions risks being *ad hoc*, but without such a restriction the thesis is open to counterexamples.

A similar point can be made with respect to the Argument from Disidentification. In this argument, the central question is whether Immunity should be construed so as to rule out error through disidentification or not. I said that a neutral answer to this question can be given by looking at the theoretical background. One way of defending Epistemic Immunity would be to argue that disowning claims, construed as disidentifications, do not fall within its scope. But given the explanation of Immunity in terms of identification-freedom, this strategy is not convincing. At the core of the epistemic approach is the idea that being aware of a mental state means being aware of that state as one's own. Again, we can think of the critics to really have this idea in mind. Witness another way in which Marcel spells out the target of his critique: "The thesis is essentially that any knowledge about properties that is gained in certain ways cannot be known to apply to an individual without *ipso facto* knowing that the properties apply to oneself." (Marcel 2003: 81) But if this idea were correct, not only should it be impossible to positively ascribe a mental state to oneself that is not one's own.[93] It should also be impossible that one disowns an introspected state which actually is one's own. Hence, given the explanation of Immunity in terms of Self-Awareness and identification-freedom, excluding disidentifications from the scope of Immunity would, again, seem *ad hoc*. If Immunity is a matter of identification-freedom, disidentifications should be immune in the same way as self-ascriptions.

But couldn't proponents of the epistemic approach accept that disowning claims fall within the scope of Epistemic Immunity, but deny that the disowning claims are in error through misidentification? An attempt to do

[93] Actually, Self-Awareness does not explain why this kind of mistake should be impossible (see § 7.2).

so might take its lead from the argument that in cases of deviant causal chains there is no misidentification, but rather illusion. However, this strategy cannot be adapted to the case of disowning judgments. In cases of deviant causal chains, the argument goes, there simply is no question for the subject who is the owner of the introspected mental state. This is why, on the epistemic view, the self-ascription is mistaken, but nonetheless identification-free and hence not in error through misidentification. In the pathological cases, in contrast, things are reversed. Although the states in fact are the subjects' own, they think that they are not. It is the fact that subjects *assume* a deviant causal origin, which makes their judgments identification-dependent and which makes the resulting error an error through misidentification. (See also Evans's discussion of the headphone case, 1982: 184–188.)

The first two arguments of the refined critique are both related to the third, indirect challenge, the Argument from Identification-Dependence. Assuming, for the moment, that the pathological cases satisfy the introspection constraint, the cases show that introspection-based judgments are not identification-free. The Argument from Identification-Dependence still applies even if proponents of Epistemic Immunity exclude external attributions and disowning claims from the scope of the thesis. The argument is that, if external attributions and disownings are identification-dependent, so are self-ascriptions proper. This is closely related to the two other arguments in that it elucidates how the restriction to self-ascriptions would be *ad hoc*, given the explanatory framework of identification-freedom. The argument is based essentially on an assumption which, as far as I know, is shared by all proponents of the epistemic approach, the assumption that identification-dependence entails liability to error through misidentification.

But couldn't proponents of Epistemic Immunity reply as follows? The fact that external attributions and disownings are identification-dependent does not show that self-ascriptions proper (i.e. judgments of the form 'F(*i*)') are identification-dependent. The former are based on categorically different grounds and hence do not imply anything for the latter. This assumption

would help to rebut all three arguments. First, it would directly reject the central premise of the Argument from Identification-Dependence. Second (and third), it could serve to motivate the exclusion of external attributions and disowning claims from the scope of Immunity, which is necessary to reject the Argument from Subject-Neutral Immunity and the Argument from Disidentification. The idea would be that Epistemic Immunity must be restricted to self-ascriptions precisely because only these, but not external attributions and disownings, are based on identification-free grounds. However, the problem with this reply is, again, that the crucial assumption is not very convincing and sort of *ad hoc*. Why should we assume that external attributions and disownings are based on different grounds than self-ascriptions proper?

To defend this assumption, proponents of Epistemic Immunity would have to first tell us more on how to individuate grounds. What I mean by the individuation of grounds is this: It is not beliefs per se that are in error through misidentification or immune to such error, but judgments, i.e. beliefs based on certain grounds (see § 2.2). For a belief that is based on certain grounds to be immune to error through misidentification accordingly means that it is not possible for this type of belief to be in error through misidentification *when based on this type of grounds*. The question then is, how finely do we individuate the grounds? To illustrate the question, compare the two cases:

(1) I judge that Forrest is running based on seeing Tom run.
(2) I judge that Forrest is running based on seeing Forrest run.

Obviously, there must be some difference between the grounds of my judgment in (1) and the grounds of my judgment in (2). But that cannot be the kind of difference that plays a role for the individuation of grounds in immunity claims, for otherwise any judgment could be immune. To wit, my judgment that Forrest is running as based on seeing Forrest run would be immune because it is not possible for my belief that Forrest is running to be in error through misidentification *when based on seeing Forrest run*. So,

my grounds in cases (1) and (2) must be something like my visual impression as of seeing someone run whom I take to be Forrest.

This does not give us an answer to the question how to individuate grounds, it merely illustrates what the challenge is. I do not know how proponents of Epistemic Immunity would answer the question of individuation, but one thing should be clear: on the epistemic approach, grounds must be individuated from a subjective perspective. For instance, in proprioceptive cross-wiring cases subjects believe to be receiving information from their own body, when in fact they are receiving information from somebody else's body. To defend Immunity regarding proprioception-based self-ascriptions against cross-wiring cases, it will not do to say that cross-wiring involves different grounds than normal (i.e. not cross-wired) cases. Of course, from the outside we can easily distinguish the two kinds of cases, but the question is whether judgments are immune to error through misidentification from a subjective perspective. The question is for instance, when I have a visual impression as of standing on Oberbaumbrücke, can I be sure that it is me who is standing on Oberbaumbrücke?

Given a criterion for the individuation of grounds, proponents of Epistemic Immunity would next have to show that the pathological judgments we find in thought insertion are in fact based on different grounds than normal self-ascriptions. At this point, we not only run into controversial empirical questions regarding the genesis of the delusional beliefs (see e.g. the discussion of endorsement vs. explanationist approaches in § 3.1). A proponent of this reply would further have to show that 'normal' attributions of mental states are not based in the same way on the experience of the state in question or a higher order explanation. It is beyond the scope of this work to pursue this matter further. While I am far from having shown that Epistemic Immunity cannot be defended in this way, the above considerations should suffice to make clear that the defense is not trivial. Summing up, it can be stated that the Arguments from Identification-Dependence, from Subject-Neutral Immunity, and from Disidentification present a serious challenge to the epistemic approach. Since I am myself

pursuing the ontological approach to Immunity, I will leave the defense of Epistemic Immunity to someone else.

The Normative Approach

Before I move on to the ontological approach, let me mention a variant of the epistemic approach which is more readily defendable against the arguments raised in the previous chapter. I have assumed that the (traditional) epistemic approach construes the central claims (Self-Awareness, identification-freedom, and Immunity) as *descriptive* claims. I now consider a variant of the epistemic approach which construes these claims as normative claims. Hence, I call this the *normative approach* to Immunity.[94] Let me illustrate what that means.

On the epistemic approach, Self-Awareness is the claim that, as matter of fact (or even as a matter of necessity), if one is introspectively aware of a mental state, one is aware of that state as one's own. The approach is epistemic in so far as it understands error through misidentification as a phenomenon that has to do with the judgment's justification. It is descriptive in so far as it makes a descriptive claim about what the justification of introspection-based self-ascriptions actually looks like. Similarly, Immunity is the claim that introspection-based self-ascriptions of mental states cannot be epistemically based on a false identification component.

The normative approach, in contrast, does not make claims about how subjects actually justify introspection-based self-ascriptions, but rather about how subjects *should* justify these self-ascriptions, or how they *would* if they were *rational*. Self-Awareness, on this approach, is the claim that, when one is introspectively aware of a mental state, then one *is justified* to self-ascribe that state, or one *rationally should* self-ascribe that state. The idea that a self-ascription is identification-free could be read, on this approach, as saying that there is no gap between a subject's being justified in believing that a property is instantiated and the subject's being justified in believing that this property is instantiated in herself. Similarly, the

[94] Thanks to James Pryor for suggesting this approach to me.

Immunity Thesis, on this approach, could be read to say that if one is introspectively aware of a mental state one cannot rationally wonder whose state it is, or one cannot be justified or rational in doubting that it is one's own. Since I do not know of anyone explicitly taking this approach I shall only make two very brief remarks.[95]

First, discussing the pathological cases as putative counterexamples to Normative Immunity hinges on the controversial question whether the delusional beliefs in question can be considered as in some sense rational. Many theorists hold that delusions such as thought insertion are best construed as rational reactions to extremely unusual experiences. If this is correct, the pathologies seem to show that the normative Immunity Thesis is false: given these unusual experiences, it can be rational to wonder whose thought one is introspectively aware of. If, in contrast, one assumes that the delusional beliefs in the pathological cases are simply irrational, the cases do not have a direct bearing on normative Immunity (see e.g. Coliva 2002b: 43). But even then, and this is the second point, it could be argued that they have an indirect bearing on Normative Immunity. For, since the scenarios envisioned in thought insertion and the like arguably constitute conceptual possibilities, it is not irrational per se to wonder, in the spirit of radical Cartesian doubt, whether one is the thinker of an introspected thought. Hence, Normative Immunity does not appear to be a very promising claim and I will not further discuss it.

6.2 Ontological Immunity

I now propose a different approach to Immunity which remains untouched by the pathological cases. The ontological approach takes as its starting point the idea that for a judgment to be in error through misidentification is for the target object (the object to which a property is in fact ascribed) to be distinct from the source object (the object from which the predication information derives). For a judgment to be immune to this error hence

[95] James Pryor, in reviewing one of my papers, expressed some affinity to construing Immunity this way.

means that the property which a subject believes to be instantiated is ascribed to the right person, where 'the right person' has to be understood as 'the person who actually instantiates that property (or seems to instantiate that property, in cases of mispredication)'.

Applying this idea to First-Person Immunity yields the following. If one self-ascribes a state of which one is introspectively aware, then the object from which the predication information derives must be identical to the object to which the predicate is applied. Given that the target object is always the self (Immunity is a claim about *de se* self-ascriptions), we can further simplify. If one self-ascribes a state of which one is introspectively aware, then that state is one's own state.[96]

(Immunity) introspection-based self-ascription (F) → ownership (F)

Applying that notion of Immunity to the cases under discussion yields the following: if one self-ascribes a thought based on being introspectively aware of that thought, it must be one's own thought; if one self-ascribes an intention based on being introspectively aware of that intention, it must be one's own intention; if one self-ascribes a sensation based on being introspectively aware of that sensation, it must be one's own sensation.

Putting it this way brings out very clearly why the pathological cases do not even begin to challenge Ontological Immunity. What the pathological cases arguably show is that it is possible to be introspectively aware of a mental state without self-ascribing it. That is to say, they show that the antecedent of Ontological Immunity can be false. But that doesn't threaten the truth of the implication.

---

[96] Note that in cases of mispredication the self-ascribed state is not the same as the introspected state. The term 'that state' in the consequent must then be understood to refer to the introspected state. To illustrate, if I judge that I am hungry when really I am tired, my self-ascription of hunger does not imply that the hunger is my own hunger (for there is no hunger), but that that the tiredness on which the judgment is based is my own tiredness (which I mistook for hunger). For the sake of perspicuity, I will ignore the possibility of mispredication in what follows.

Note also that Ontological Immunity is related to Self-Awareness only in the sense that Self-Awareness may explain why the antecedent of Ontological Immunity is often true: we do (normally) self-ascribe mental states of which we are introspectively aware. But Self-Awareness does not even begin to explain why the implication holds, for it says nothing about whose states the introspected states actually are. Self-Awareness is perfectly compatible with Ontological Immunity being false, i.e. with my being introspectively aware of and self-ascribing somebody else's mental states. Conversely, Ontological Immunity is perfectly compatible with Self-Awareness being false, i.e. with my being introspectively aware of mental states without self-ascribing them (the falsity of the antecedent does not threaten the implication).

So, the explanation of Ontological Immunity cannot be given in terms of Self-Awareness. Rather, Immunity is explained by the fact that I cannot be introspectively aware of somebody else's mental states, but can be introspectively aware only of my own states. I dub this assumption the *introspection-ownership link*: If one is introspectively aware of a mental state then that state is one's own.[97] It licenses the following conditional: If one self-ascribes a mental state of which one is introspectively aware, one self-ascribes a mental state that is one's own. And that is exactly what Immunity states.

[97] The term 'ownership' is a bit problematic as it has been used in this debate to denote a number of different properties. In lack of a better term, let me stipulate that 'ownership of a mental state' shall denote roughly and ambiguously the same as the notion of a mental state's being one's own. For now, I am leaving this notion intentionally ambiguous and open to interpretation. In particular, note that the general notion of mental ownership is open to be interpreted in both senses in which a thought can be one's own: thought-ownership and thought-authorship. Hence, the general notion of mental ownership is not to be confused with the more specific notion of thought-ownership.

Similar explanations have been given by Campbell (1999a), Coliva (2002a), and Romdenh-Romluc (2013)[98]. Witness the following passage:

> A self-ascription will be IEM if it is based on some way of finding out about my own states and properties that only allows me to find out about myself. Thus a self-ascription of intention will be IEM if it is based on a way of knowing about my intentions that only provides me with knowledge of my own intentions. Since it seems that I can only be first-personally aware of my own intentions, self-ascriptions of intention based on this form of awareness are traditionally claimed to be IEM. (Romdenh-Romluc 2013: 497)

Note that Romdenh-Romluc does not seem to be aware of the fact that her approach diverges substantially from the dominant epistemic approach. She is correct that introspection-based self-ascriptions of intentions are "traditionally claimed to be IEM". But this claim is not traditionally supported by the idea that introspection provides access only to one's own intentions.

Coliva, in contrast, appeals to both explanations (2002a: 28f.). Although she points to the introspection-ownership link, she also tries to explain Immunity in terms of identification-freedom. More precisely, she offers the explanation in terms of the introspection-ownership link only as a supplement to explain why introspection-based self-ascriptions (in contrast to proprioception-based self-ascriptions) are *logically* (rather than *de facto*) immune.

Before I move on to the discussion of the introspection-ownership link, let me address a possible worry about the ontological definition of error through misidentification. It seems there is an obvious counterexample to the definition in cases in which a judgment is based on *several* misidentifications which happen to cancel each other out in the end. Consider the following case: I see Lili running, but mistake her for Clara and thereby come to believe that Clara is in a hurry. I further believe that Clara is the

[98] For a further very brief remark in that direction see also Hogan & Martin 2001: 207.

head of the philosophy department and thus conclude that the head of the philosophy department is in a hurry. However, as a matter of fact, not Clara but Lili is the head of the philosophy department. I thus end up having an (accidentally) true belief. Importantly, the belief is not only true, but source and target do not diverge: my predication 'is in a hurry' is based on seeing Lili running and it is ascribed to Lili. The simple definition of the ontological approach would imply that the judgment is not in error through misidentification. But intuitively, it is in error through misidentification. Not just once, but twice.

A simple fix for such cases is to say that a judgment is in error through misidentification either when source and target diverge, or when the judgment is based on another judgment that is in error through misidentification. Cases of multiple misidentifications which cancel each other out in the end can be dealt with by maintaining that the final judgment (here: 'the head of department is in a hurry') is based on another judgment (here: 'Clara is in a hurry') which is clearly in error through misidentification in virtue of a divergence of source and target.

The Introspection-Ownership Link

Let us take a closer look at the introspection-ownership link (henceforth: *Introspection-Ownership*). The idea underlying Introspection-Ownership is this: If, *per impossibile*, I was introspectively aware of somebody else's mental state, then *ipso facto* it would be my mental state. The claim is most plausible for purely phenomenal states. If, *per impossibile*, I was introspectively aware of somebody else's pain then I would *ipso facto* be in pain. Hence, it would be my own pain of which I was aware. As Coliva puts it, "it is a matter of conceptual truth that each mental state one is introspectively aware of is one's own" since "being introspectively aware of a certain mental state is a *criterion* of what has to count as *one's own* conscious mental state" (2002a: 29).

An idea that is very similar to Introspection-Ownership is suggested by Campbell as an explanation of Immunity.

> In the case of the first person, I suggested that what makes an experience an experience of X's is the possibility of self-ascription of it by X. If X is able to self-ascribe the experience, that constitutes the experience being an experience of X's. So when I ascribe an experience to myself I cannot be wrong in thinking that the experience is mine; for on this approach, the very fact of my self-ascribing the experience is enough to constitute its being an experience of mine. (Campbell 1999a: 97)

So, rather than an introspection-ownership link Campbell suggests a self-ascription-ownership link. Campbell's main goal here is to reject an explanation of Immunity in terms of the meaning of the first-person pronoun. He does not discuss and explain the self-ascription-ownership link in much detail. But I think it is safe to assume that it amounts pretty much to the same idea as Introspection-Ownership. Certainly, the possibility of self-ascription, which Campbell takes to guarantee ownership, must be realized by introspection, for otherwise the view would be open to obvious counterexamples. Suppose, with a bit of science-fiction, that I monitor a mental experience on a cerebroscope. This allows me to self-ascribe that experience, when I believe that it is my brain which I am monitoring. However, self-ascribing a mental experience which I monitor on a cerebroscope does not guarantee that it really is my experience; I could be mistaken about whose brain I am monitoring. Thus, brain-monitoring and other third-personal kinds of access to mental experiences certainly are not among the possibilities of self-ascribing an experience that are constitutive for the experience's actually being mine.

Three aspects of Introspection-Ownership require more discussion: its modal strength, the notion of ownership, and the kinds of states for which it holds. I believe that all three aspects are interrelated in the sense that, depending on what kind of state and what kind of ownership is in question the link holds with varying degrees of modal force. Introspection-Ownership is most persuasive with respect to phenomenal states (cf. Coliva 2002a) or mental experiences (cf. Campbell 1999a). Accordingly, Coliva and Campbell construe their respective claims as holding with conceptual

necessity. They claim that introspective awareness is a conceptually sufficient condition for ownership (Coliva) or that the possibility of self-ascription is constitutive of ownership (Campbell). Illustrating her claim with the examples of pain and belief, Coliva generalizes the point without further argument to the introspective awareness of mental states in general. It will become clear shortly why I find such generalizations problematic. Campbell makes a general claim about mental experiences. He is also aware of the fact that some experiences, such as thoughts, may allow for different kinds of self-ascriptions. That is to say, there are different senses in which a mental state or experience can be one's own (here: thought-ownership vs. thought-authorship). Accordingly, we have to distinguish the claim that introspection of a thought guarantees thought-ownership from the claim that introspection of a thought guarantees thought-authorship.

Although I am mainly interested in introspection-based self-ascriptions of mental states, I want to further suggest that my proposed explanation of Immunity generalizes perfectly to proprioception-based self-ascriptions of bodily states. It is the link between proprioception and ownership which explains the immunity of bodily self-ascriptions. If one is proprioceptively aware of a body state it is (normally) one's own state. However, it seems that proprioceptive awareness of one's legs being crossed does not guarantee with conceptual necessity nor constitute the fact that those legs are one's own.

Now, I said that Introspection-Ownership holds with varying degrees of modal force, depending on what kind of state is in question and, if applicable, what kind of ownership is in question. What this really means is that we should distinguish a number of different links and that these different links come in varying degrees of modal force.[99] For instance, while introspective awareness of a thought arguably implies thought-ownership (in the

[99] Still, in what follows, I intend the general notion of Introspection-Ownership to comprise all these different introspection-ownership links as well as proprioception-ownership links. Very generally, the idea of Introspection-Ownership is that there is a link between first-personal awareness and ownership.

technical sense) with conceptual necessity, it implies thought-authorship only *de facto*, given that there is no telepathy etc. Spelling out a comprehensive classification is beyond the scope of this thesis and not necessary for the overall argument. However, to make the claim more tangible I shall suggest how to treat some of the cases that play a central role in the debate on Immunity.

1. With conceptual necessity: phenomenal states of which one is introspectively aware are one's own (see the example regarding pain above). Hence, the corresponding self-ascriptions are logically immune.
2. With conceptual necessity: propositional attitudes such as thoughts and intentions of which one is introspectively aware are one's own in the sense that they occur within one's own mind or within one's own consciousness (see the technical notion of thought-ownership). Hence, the corresponding self-ascriptions are logically immune.
3. With nomological necessity: past experiences which one episodically remembers are one's own past experiences. Quasi-remembering of past experiences that are not one's own is conceptually, but not nomologically possible. If this is correct, memory-based self-ascriptions of past experiences are nomologically immune.
4. Barring cross-wiring scenarios: propositional attitudes such as thoughts and intentions of which one is introspectively aware are one's own in the sense that one is the causal origin of these attitudes. It is conceivable, but not nomologically possible, that somebody telepathically inserts a thought or intention into somebody else's mind. It seems nomologically possible, but doesn't happen very often, that a thought or intention is produced via direct electrical stimulation in another subject's mind. If this is correct, introspection-based self-ascriptions of authorship are *de facto* immune given that the propositional attitudes in question do not have causally deviant origins.
5. Barring cross-wiring scenarios: bodily states that one perceives proprioceptively are one's own. It is conceptually possible that two

subjects are wired up in a way such that one person proprioceptively perceives the other person's bodily states. Hence, proprioception-based self-ascriptions of bodily states are *de facto* immune given they do not have causally deviant origins.

6. Given normal conditions: external perception delivers information about one's own relation to other objects. For instance, when one has the visual impression as of standing on Oberbaumbrücke, it is usually oneself who is standing on Oberbaumbrücke. With the aid of virtual reality goggles, illusions can be created in which what one is seeing as if in front of oneself is really in front of somebody else (see e.g. Petkova & Ehrsson 2008). Hence, exteroception-based judgments about one's relation to other objects are *de facto* immune given normal conditions.

I am not committed to any particular evaluation of these cases and the overall argument does not depend on them. I offer them to illustrate how different introspection-ownership links may be of different strengths and to highlight how the suggested explanation of Immunity dovetails with the widely accepted idea that different kinds of judgments enjoy a different kind of immunity.

Ontological Immunity and the Refined Critique

I claimed in chapter 5 that Ontological Immunity can be defended against the Argument from Subject-Neutral Immunity and the Argument from Disidentification by excluding from the scope of the thesis external attributions and disowning claims (construed as disidentifications). Importantly, I can grant that the external attributions and disidentifications in the pathological cases involve ontological error through misidentification. I claim that they do not challenge Ontological Immunity because they fail the self-ascription constraint. The crucial question is whether this scope restriction can be properly motivated. I argued that the epistemic approach does not fare very well in this respect. I now show that the required restriction follows naturally from the explanatory framework and the background assumptions of the ontological approach.

On the ontological approach, Immunity is explained essentially as a consequence of Introspection-Ownership. When one self-ascribes a state of which one is introspectively aware, one cannot, as a matter of fact, ascribe the state to the wrong subject. Importantly, the ontological approach is not committed to any claims regarding the question how we come to believe that an introspected state is one's own or not. It is therefore fully compatible with a subject going wrong in ascribing an introspected mental state to another subject, or going wrong in disowning the state. In other words, it is completely compatible with the introspection-ownership link that certain mental states of which one is introspectively aware, are not recognized as one's own. This is why, given the explanation of Immunity in terms of Introspection-Ownership, we would not expect Immunity to hold for external attributions or for disowning claims.

Let me put the same point somewhat differently. With respect to introspection-based ascriptions of mental states, ontological error through misidentification can principally come in two forms: self-ascribing an introspected mental state that is actually not one's own and disowning (or ascribing to an external entity) an introspected mental state that actually is one's own. Introspection-Ownership only rules out the former case. This is why, given the ontological approach, it is natural to construe Immunity as not ruling out misidentifications in disowning claims and external attributions.

It should be clear, also, that the ontological approach not only provides the resources to defend Immunity against the Argument from Subject-Neutral Immunity and the Argument from Disidentification, but that it is also not affected by the Argument from Identification-Dependence. For the ontological approach does not rely on any assumption regarding the identification-freedom of introspection-based self-ascriptions. More than that, Ontological Immunity is fully compatible with the idea that introspection-based ascriptions of mental states are identification-dependent. Even if we were to go wrong most of the time in disowning or externally attributing introspected mental states, this would not challenge Ontological Immunity.

Before moving on, let me briefly address the worry that my ontological approach trivializes Immunity.[100] It may seem that Ontological Immunity is a tautology of the form: whenever one attributes a property to *x* based on information that derives from *x*, one cannot be in error through misidentification. That would be tautological indeed, but that is not the form of Ontological Immunity. The source constraint in my version of Immunity does not restrict the scope to judgments that are based on information coming from the subject herself. Rather, it restricts the scope to judgments that are based on introspection. So, there is an additional premise which plays a crucial role in my explanation of Immunity, namely the premise that introspection gives access solely to one's own mental states. And the present discussion is witness to the fact that this is not a trivial or tautological issue, but a debatable claim (see especially the discussion of the craniopagus twins' case at the end of this chapter).

Introspection-Ownership and Alienation

Finally, let me quickly sketch how the suggested explanation handles the putative counterexamples. The question in this section is not whether Immunity can be defended against the putative counterexamples. I have argued above that it can. The present question is whether my proposed *explanation* of Immunity works for these cases as well, or rather, whether my proposed explanation of Immunity is compatible with the pathological cases. For, it may seem that the pathological cases challenge Ontological Immunity indirectly in showing that subjects are introspectively aware of mental states which are not their own.

Generally, to assess the introspection-ownership link, we have to ask two questions, an epistemic question and an ontological question. The first question is whether the state is introspectively known to be instantiated (or, more generally, by first-personal awareness); the second question is whether the state is the subject's own. A counterexample against Introspection-Ownership would have to be a case that involves first-personal

[100] Thanks to Gottfried Vosgerau for pressing me on this.

awareness of a state that is not the subject's own. With respect to the pathological cases, the first question has already been discussed above. To quickly rehearse, in FB's case of somatoparaphrenia, the ascription of the sensation is clearly based on first-personal awareness. Further, I briefly presented the controversy on the question whether thought insertion involves first-personal awareness (§ 3.1). I will, again, grant for the sake of argument that subjects have first-personal awareness of authorship in thought insertion, and I will grant analogously that subjects have first-personal awareness of agency in made impulses. Finally, I argued that in anarchic hand syndrome there is no first-personal action awareness. This leaves us with three pathological cases in which the antecedent of the introspection-ownership link is true.

What about the ontological question then? Are the inserted thoughts, the made impulses and the alien sensations the subjects' own? The obvious answer seems to be: yes, certainly. I have assumed all along that subjects are deeply mistaken in disowning the thoughts, intentions, and sensations in question. And indeed, that is how I would answer the ontological question. In FB's case of somatoparaphrenia, it is beyond doubt that the sensation really is FB's sensation. (This is implicitly acknowledged also by Lane & Liang.) As for inserted thoughts, I have discussed in great detail what it means to be the author of a thought. This discussion now also proves relevant to the explanation of Immunity. The question is whether introspection guarantees authorship. Given the Causal Analysis, the answer is: yes. Virtually all thoughts causally originate in the appropriate way within the subject. In all normal circumstances, when we are introspectively aware of a thought, this thought is the causal result of cognitive processes in our brain. However, the cases of truly inserted thoughts have shown that this is certainly not a conceptual necessity. If some form of telepathy were possible, subjects could introspectively experience thoughts that are not theirs. Even more, the link does not even constitute a nomological necessity. With direct electric stimulation it is actually possible to insert motor intentions in subjects. So the introspection-authorship link has a modal force that is restricted to regular cases, i.e. cases which do not involve

telepathy, brain stimulation and the like. Analogously, we can say that in the phenomenon of made impulses, it is the subject himself who is carrying out and controlling the action, even if he lacks a sense of control and a sense of agency. But again, the introspection-ownership link for intentions does not hold with conceptual necessity. Given this view on what it means to be the author of a thought or agent of an action, the pathological cases do not spell any trouble for the introspection-ownership link.

Yet, it may be that not everyone shares this view. In my discussion of authorship I show that some theorists might be prepared to swallow that subjects actually are not the authors of inserted thoughts. Similarly, theorists may say that if subjects do not feel in control of the action in made impulses, then indeed they are not the agents of these actions.[101] This view would imply that there can be thoughts that do not have an author and that there can be actions that do not have an agent. Let me dub this the *no-agency assumption*. On this assumption, the pathological cases do seem to undermine Introspection-Ownership: subjects would be first-personally aware of thoughts or actions which are not their own in the sense of authorship or agency, respectively. In the remainder of this section I discuss the implications of the no-agency assumption for my view.

First, let me quickly explain why even under the no-agency assumption the pathological cases do not threaten Immunity. Here, we need to distinguish again between the two different ways of construing the disowning judgment (external attributions fall outside the scope of the thesis anyway and need no further discussion here). If a subject is introspectively aware of an unowned mental state and disowns that state in the sense of a *negative self-ascription* ('I am not-F'), then the judgment simply is correct: *ex hypothesis* the subject isn't the thinker of the thought or the agent of the action. If a subject is introspectively aware of an unowned mental state and disowns

[101] More precisely, the idea would be that in made impulses the intentions are not the subjects' own and that therefore the resulting actions are not the subjects' own either.

that state in the sense of a *disidentification* ('It's not me who is F'), then the judgment is either correct or falls outside the scope of the thesis.[102]

It may seem, however, that the no-agency assumption opens the door for a different kind of counterexample against Immunity. Consider a case just like thought insertion or made impulses but without the disowning or external attribution belief. That is to say, consider a case in which a subject experiences a thought that is in fact not her own, but of which she believes that it is her own. In such a case, a subject would self-ascribe a mental state that is not her own. But again, while the self-ascription in such a case would be false, it would not be in error through misidentification. The subject would not go wrong with respect to the question who is the author of the thought or the agent of the action. Rather, the subject would go wrong in ascribing authorship or agency in the first place. Her mistake would be an error through mispredication.[103]

Second, I turn to the fact that the no-agency assumption seems to refute Introspection-Ownership and thereby challenges the introspection-ownership explanation of Immunity. First of all, note that even if cases like thought insertion and made impulses were to serve as counterexamples to Introspection-Ownership this really isn't that big of a deal for my view. I never claimed that the introspection-authorship link for thoughts and the introspection-agency link for actions holds with conceptual necessity. Rather, I have granted all along that there are conceptually and nomologically possible counterexamples to these links. To wit, I assumed that there

---

[102] Insofar as the disidentification is about somebody other than the subject, e.g. in implying an existential claim (intuitively: 'someone is F, it's not me'), it does not fall within the scope of the thesis in failing the self-ascription constraint. Insofar as it is about the subject herself (intuitively: 'I am not the author'), it is correct. I am presupposing here that even under the assumption that the subject is not the author, the subject still is the source of the predication.

[103] Putting it in terms of source and target: Certainly the subject is the target of the ascription; arguably she is also the source (even if the subject is not the author). Hence, there is no divergence of source and target.

can be truly inserted thoughts of which a subject can be introspectively aware without being the author. The same holds for actions. The assumption that thought insertion and made impulses involve unowned mental states would just add another strange counterexample.

More importantly, however, the pathological cases are not cases in which a subject is introspectively aware of *somebody else's* mental state, but are rather cases in which a subject is introspectively aware of an *unowned* mental state. This allows us to introduce a simple amendment in defense of Introspection-Ownership. Remember, the initial motivation for Introspection-Ownership as an explanation of Immunity was to say that if one is first-personally aware of a state, then that state cannot be somebody else's state. Taking into account the possibility of unowned mental states, the introspection-ownership link has to be amended as follows.

> *Introspection-Ownership**: If one is first-personally aware of a state then that state is either one's own or unowned (but it certainly is not somebody else's).

Introspection-Ownership* stays true to the original idea, but is not vulnerable to cases of unowned mental states.

Introspection-Ownership* allows us to accommodate the no-agency assumption in the introspection-ownership explanation of Immunity as follows. My initial explanation of Immunity said that introspection guarantees ownership. Obviously, this explanation was given under the assumption that there cannot be unowned mental states. Given the no-agency assumption, the explanation needs a minor amendment which corresponds to the amendment in Introspection-Ownership*. Again, the fundamental idea remains the same: introspection-based self-ascriptions are immune to error through misidentification because one cannot introspect another person's mental states. When a mental state is self-ascribed based on introspection, the state must either be one's own (this is the old idea of the introspection-ownership link) or it is nobody's (this is the new idea, brought in by the no-agency assumption). If the state is indeed one's own, the self-ascription of it is correct. If the state is unowned, the self-ascription

is mistaken, but it is not in error through misidentification. For, the subject did not go wrong in figuring out who is the author, agent, or owner of the mental state, but rather went wrong in claiming that there is an author, agent, or owner to begin with. Hence, also on the no-agency assumption the introspection-ownership link can explain the immunity of introspection-based self-ascriptions.

Craniopagus Twins

Recently, Langland-Hassan (forthcoming) has suggested a very different kind of empirical counterexample against Immunity which particularly aims to challenge Introspection-Ownership. His counterexample is much like a real life case of cross-wiring, namely a case of twins who are conjoined at the heads and brains. It seems that, due to their unusual anatomy, each of the twins has introspective access to the other twin's mental states. That is to say, each twin is introspectively aware of what is going on in the other twin's mind. The particular instance discussed in detail by Langland-Hassan is introspective awareness of a visual experience: one of the twins, Krista, can tell with her eyes closed that her sister, Tatiana, is looking at a toy pony. Of course, there is no misidentification involved in this particular instance. But Langland-Hassan imagines a case in which one twin is introspectively aware of her sister's visual experience and, mistaking it for her own, self-ascribes it. In this case, it seems, we have an introspection-based self-ascription of a mental state that is in error through misidentification. Langland-Hassan is not so much interested in whether this *actually* sometimes happens to the twins. Rather, his point is, the twins' unusual connection shows that such error is conceptually and nomologically possible.

Langland-Hassan's critique of Immunity is special in at least three respects. First, in contrast to other critics of Immunity, Langland-Hassan pays attention to the fact that a counterexample to Immunity has to be an introspection-based self-ascription. In contrast to the cases of alienation, there is hence no question that the twins' case falls within the scope of the Immunity Thesis. The judgment 'I am having a visual experience as of a toy

pony', based on being introspectively aware of such an experience, satisfies both the introspection constraint and the self-ascription constraint more clearly than any of the four cases of alienation. Unlike the disowning claims, the judgment is a self-ascription proper and it is based on introspection if anything is. (What is not quite as clear is whether the judgment really is in error through misidentification, a question I will turn to shortly.)

Second, the case is special with respect to the kind of mental state that is self-ascribed. The case, so to speak, involves cross-wiring of visual perception. However, deviating from the traditional cross-wiring debate, Langland-Hassan aims to challenge the self-ascription of the (phenomenal) visual experience rather than the self-ascription of the (factive) perceptual state (see my distinction in § 2.5). Langland-Hassan argues that Krista is in error through misidentification not merely in judging that she herself is *visually perceiving* the pony (in the sense of light falling into *her* eyes), but that she is in error through misidentification in judging that the visual experience as of a pony occurs in *her* mind. Langland-Hassan explicitly states that in the case he imagines, Krista does not only *believe* that she is having a visual experience, rather she is introspectively aware of the experience in the sense that she is having the experience. That is to say, Krista's awareness of the visual experience is phenomenally indistinguishable from her regular own visual experiences, yet the experience is supposed to occur in Tatiana's mind only, not in Krista's. This is a particularly strong claim. Typically, the debate on perceptual cross-wiring asks whether one can be in error through misidentification in self-ascribing a factive perceptual state or in self-ascribing a relation to an external object (see e.g. Chen 2009: 29, Evans 1982: 184). To make his case, Langland-Hassan has to argue that although Krista is introspectively aware of the experience and phenomenally experiences it, it is not Krista's experience in the sense that it does not occur within her mind.

Third, unlike typical discussions of cross-wiring cases, Langland-Hassan's discussion does not concern the question whether Krista's self-ascription of Tatiana's visual experience is identification-free. Most discussions of comparable cross-wiring cases revolve around the question whether the

kind of mistake that the cross-wired subject makes is an error through misidentification or rather an error that is comparable to a hallucination or illusion. The typical, Evansian line of argument is that the visual experience is experienced by Krista *as her own*, hence her self-ascription is identification-free, and hence the mistake she makes is not a misidentification, but rather an illusion (cf. Evans 1982: 182-190). As I pointed out above, this is an issue only for the epistemic approach to Immunity. Langland-Hassan discusses none of that. Rather, he discusses the question whether it is Krista's, Tatiana's, or both' visual experience which Krista is introspectively aware of. More than that, he explicitly discusses and challenges the assumption that introspection of a mental state implies ownership of that state. This is exactly the assumption the ontological approach to Immunity is based on. It is Langland-Hassan's challenge of Introspection-Ownership that I want to discuss now.

Langland-Hassan's aim is to show that it is logically and nomologically possible to be introspectively aware of a visual experience that is not one's own (in the sense that it does not occur within one's mind). His argument is based on two assumptions. First, he imagines the twins' case as a 'one token scenario' (forthcoming: 11f.). That is to say, the way in which the twins are connected is not simply that there is a copy *c* in Krista's brain of Tatiana's visual experience *v*. If there were two token events and Krista were introspectively aware of *c* rather than of *v* the case would not involve misidentification. Second, Langland-Hassan assumes what he calls a "distinct existences view of introspection" (ibid.: § 4.1). This is the assumption that "introspection involves a state or process that is ontologically distinct from the states of which it makes one aware" (ibid.: 13). He explicitly does not claim that these assumptions are true in the actual case of the twins. His point is that they present logical and nomological possibilities. I will grant both assumptions for the sake of argument. Now, to refute the introspection-ownership link, Langland-Hassan has to argue for one final point. He has to establish (again, as a logical and nomological possibility) that the one token mental event, the visual experience of the toy pony, is not shared among the twins, but is solely in Tatiana's mind.

In a first step, Langland-Hassan ridicules the introspection-ownership link as implausibly strong and amounting to nothing less than magic. The way he puts it, Introspection-Ownership claims that

> [...] introspection has the power to transform a mental state that otherwise would not have been a part of one's mind into a proper part *of oneself*. In the case of the twins, the idea would be that Tatiana's visual experience *v* becomes a proper part of Krista at the moment Krista becomes introspectively aware of it. This is not unlike having the power to turn objects to gold with a touch of one's hand. For that reason, I will call the following principle 'Midas Touch' [...]: Subject *S*'s becoming introspectively aware of *m* suffices for *m*'s occurrence within *S*'s mind. (ibid.: 15f.)

Although I assume that Langland-Hassan does not intend this mockery to carry any argumentative weight, I nonetheless want to make two quick comments. First, whether Introspection-Ownership is a strong claim depends on one's view on introspection. I am not going to argue here about what is the right view on introspection, I merely want to point out that there may be views on introspection, views that Langland-Hassan has put to the side, but which he has not refuted, on which Introspection-Ownership is all but a strong claim. Admittedly, on perceptual theories of introspection it may seem like an odd idea to say that introspectively perceiving a mental state suffices to make that state one's own. But on other views on introspection which might say, for instance, that being introspectively aware of a pain simply means (phenomenally) being in pain, it may seem quite natural and not magical at all to assume that introspective awareness suffices for ownership.

Second, and more generally, the idea behind Introspection-Ownership is not that the act of introspection changes the ownership of the state in question. That is to say, it does not turn a mental state which previously was not one's own into one's own state. Rather, the idea is that only states which are one's own to begin with (states which occur within one's mind) are accessible to introspection. To set Langland-Hassan's picture straight: the correct analogy is not that anything that Midas touches turns into gold; the

analogy should rather be that everything he in fact touches is made from gold. And if he is locked in a room in which everything is made from gold, this isn't such a magical scenario after all.[104]

Let us now turn to Langland-Hassan's actual argument against the introspection-ownership link, which contains a number of steps. To make my response, it is not necessary to discuss the steps in detail. What matters primarily is their dialectical role. Here is a rough overview. On the one side, Langland-Hassan presents an alternative to Introspection-Ownership, a different sufficiency condition for ownership:

> *Brain Based*: if mental state *m* is realized in *S*'s brain and has at least some causal and inferential interaction with *S*'s other mental states, then *m* occurs in *S*'s mind. (forthcoming: 16)

On the other side, Langland-Hassan attacks the idea of strong integration, which he takes to be the most plausible motivation for Introspection-Ownership. Roughly, the view he attacks is this: introspection of a mental state guarantees that the state is strongly integrated with one's other mental states, and strong integration of that state implies that the state is one's own (cf. ibid.: § 5).

Finally, Langland-Hassan briefly turns to a conceptual motivation of Introspection-Ownership.[105] It is the idea that phenomenally having a mental experience simply is what it means for that experience to be one's

---

[104] A similar idea is expressed by Hogan & Martin: "Even if introspection were a kind of perception, introspectively based judgments still might be immune to error through misidentification. For instance, for some physiological reason, introspection might provide information only about the introspector, that is, the one who forms the introspectively based judgment, 'I am E'. Similarly, in the case of visual perception, a person may be physically restrained so that he can see only himself." (2001: 207)

[105] For the sake of completeness, note that Langland-Hassan addresses and rejects yet a further motivation for the introspection-ownership link, one that draws on moral obligation. I am not going to discuss that since it doesn't strike me as a very convincing motivation to begin with.

own. Experiencing a pain, for instance, simply is what it means to be in pain (even if the pain is realized in someone else's brain). And being in pain is what it means for that pain to be one's own (or: to occur within one's mind). Let me call this the *conceptual argument* for Introspection-Ownership. Langland-Hassan puts it off as flat-footed and question-begging (cf. ibid.: 22).

In the remainder of this section I will discuss the dialectics of the arguments laid out above. I am neither going to discuss whether Brain Based is plausible as a sufficient criterion for ownership, nor whether Langland-Hassan successfully rejects the strong-integration view. Rather, I grant these points to him and discuss what this shows for Introspection-Ownership. In particular, I argue that the conclusion drawn by Langland-Hassan is not warranted by his arguments and that therefore his critique of Immunity and Introspection-Ownership is not persuasive.

The contested question is whether the visual experience *v* is Krista's or not. Granting all his points, Langland-Hassan has refuted some motivations for thinking that it is Krista's and he has shown that *v* is Tatiana's. But, crucially, he has not established that it is not *also* Krista's. Yet, he thinks he has:

> I have given independent reasons for why an [*sic*] certain introspected state one "enjoys" might not occur in one's own mind (viz., it is not in one's brain, and is not well integrated with one's other psychological states) […]. (ibid.: 22)

His claim to have offered independent reasons against Introspection-Ownership is puzzling. Brain Based was offered as a *sufficient* criterion for ownership, explicitly not as a *necessary* criterion. Hence, the fact that *v* is not realized in Krista's brain (as Langland-Hassan argues) does not tell us anything about whether *v* occurs in Krista's mind or not. The remark about integration leaves me equally puzzled. Langland-Hassan discusses integration as a sufficient criterion for ownership, not as a necessary criterion, and he explicitly argues that it is possible for one's own mental states to not be integrated with one's other mental states (cf. ibid.: § 3). Hence, the claim

that *v* is not well integrated with Krista's other mental states again does not show that *v* is not occurring in Krista's mind.

One way to argue for his claim would be to assume that the disjunction of Integration and Brain Based is a *ceteris paribus* necessary condition for ownership: If a mental state occurs within a subject's mind, it must be realized in the subject's brain or strongly integrated with the subject's other mental states.[106] Langland-Hassan could now argue that *v* is neither realized in Krista's brain nor strongly integrated with Krista's other mental states and that this is a *ceteris paribus* reason to think that *v* does not occur in Krista's mind. But it isn't clear to me how he is going to argue for this necessary condition. Moreover, even if he can make a case for a *ceteris paribus* condition, it simply does not seem to apply to the twins' case. The twins' brains have a quite unusual connection and hence a *ceteris paribus* condition concerning the connection between brains and minds simply cannot be applied to this case without further argument.

Another way for him to argue would be to say that he has refuted *all* reasonable ways of motivating the view that *v* is Krista's experience, and that, therefore, we have reason to think that *v* is not Krista's. But I do not think he has shown this either. Remember the conceptual argument, which in fact I take to be the most plausible motivation for assuming that *v* occurs in Krista's mind. Langland-Hassan does not even attempt to refute this argument, but simply brushes it aside as question-begging. Now, one could try to turn this around and argue that Langland-Hassan's own claim is question-begging. But I will not engage in burden of proof tennis here. I claimed that *v* occurs (also) in Krista's mind, Langland-Hassan claims that it doesn't, and neither side seems to have a conclusive argument. I will end this discussion with a final attempt to motivate my view.

We have seen throughout this thesis that the question what it means for a mental state to be one's own is not a trivial one. In particular, there may be different notions of ownership, depending on context, on the kind of mental state in question, and possibly on other factors. Whether Introspection-

[106] Langland-Hassan suggested this view in personal communication.

Ownership holds, and whether Immunity holds, depends on the kind of notion of ownership that one has in mind. The view I am proposing comes equipped with two notions of ownership regarding perceptual states. We have, on the one hand, the phenomenal notion according to which a visual experience is a subject's own iff the subject *has* that experience. We have, on the other hand, the causal notion according to which a visual experience is the subject's own iff it casually originated in the subject's perceptual system. If Krista were to self-ascribe the experience in the phenomenal sense, she would be correct. Obviously, that is not the self-ascription Langland-Hassan has in mind. If she were to self-ascribe the experience in the causal sense, she would be in error through misidentification. This is the notion of ownership of a perceptual state that figures in traditional discussions of cross-wiring. But that is also not the notion Langland-Hassan has in mind, for he wants to make a stronger case. He wants to say that, even though Krista fully undergoes the experience in a phenomenal sense, the experience is not her own in the sense that it does not occur within her mind. It seems to me that, to make this case, he has to introduce a further notion of ownership, one that strikes me as somewhat superfluous or artificial.

Let me put the same critique in a slightly different way. Langland-Hassan has to assume a distinction between phenomenally *having* a mental experience and *owning* a mental experience. But such a distinction seems artificial insofar as the twins' case does not require any such distinction, or, to say the least, Langland-Hassan has not shown that it does. When I say that the case does not require the distinction I mean that the case can be described perfectly well without such a distinction.[107] It has not actually been *shown* that Krista is introspectively aware of an experience that is not hers, but it has rather been *assumed*. But if the distinction is introduced for the sole purpose of challenging Immunity and Introspection-Ownership the

[107] Compare this to disowning claims in thought insertion. To make sense of these claims we clearly need two notions in which a thought can be one's own, namely as occurring within one's mind vs. having been brought about by oneself.

whole critique runs risk of begging the question. Hence, I conclude that the twins' case does not amount to a persuasive refutation of Immunity or Introspection-Ownership.

6.3 Summary

In this chapter, I spelled out in more detail the differences between the epistemic and the ontological approach to Immunity. My main goal in doing so was to give a neutral answer to the question whether the Immunity Thesis should be taken to rule out error through misidentification in external attributions and error through disidentification in disowning claims. I argued that, on the epistemic approach, it makes sense to construe Immunity as claiming that these errors are not possible. On the ontological approach, in contrast, Immunity is most naturally construed as a thesis that does not apply to external attributions and disidentifications. This means that only Epistemic Immunity is subject to the refined critique from the pathological cases of alienation and that Ontological Immunity can naturally be defended against this critique. My claims about how to construe Immunity on each of the two views is neutral in the sense that they are motivated by reference to the explanatory background assumptions of the two views. In a last step, I defended Ontological Immunity and Introspection-Ownership against Langland-Hassan's challenge from the case of craniopagus twins.

7. Corollaries

I have presented two fundamentally different views on Immunity. The epistemic approach understands error through misidentification as a judgment's being epistemically based on a mistaken identification and consequently ties immunity to identification-freedom. The ontological approach understands error through misidentification as the divergence of source and target and consequently understands Immunity as the impossibility of such divergence. Neither of the two approaches is in principle superior to the other. That is to say, there is not the one correct way of construing the Immunity Thesis. Epistemic Immunity is better suited to capture the idea that, when a judgment is immune, it does not make sense to ask whether one is sure that it is *a* who is F. Ontological Immunity, in contrast, captures better the idea that one cannot be wrong about it being *a* that is F.

Epistemic Immunity is based on a certain idea of how we come to know that the mental states we are introspectively aware of are our own. It assumes, with Self-Awareness, that introspective awareness of a state implies awareness of that state as one's own. Hence, Epistemic Immunity is based on a claim about whose states the introspected states *appear to be*. Ontological Immunity, in contrast, is based on the idea that introspective awareness of a state implies ownership of that state. Hence, Ontological Immunity is based on a claim about whose states introspected states *actually are*. Explaining Epistemic Immunity means explaining *how one knows* (in an identification-free way) that the states one introspects are one's own. Explaining Ontological Immunity, in contrast, means explaining why the states one introspects *actually are* one's own.

Clearly, the epistemic approach is the dominant approach. Of the few writers who have actually made suggestions along the lines of the ontological approach, most seem not to be aware of (or are simply ignoring) the fact that they are bringing into play a fundamentally different approach (Campbell 1999a, Prosser 2012, Romdenh-Romluc 2013). A notable exception is Coliva (2002a) who distinguishes the ontological and the

epistemic explanation in offering the ontological explanation as an amendment to the explanation in terms of identification-freedom.

I argued that Ontological Immunity can readily be defended against the pathological cases. The explanation in terms of the introspection-ownership link makes plausible why the Immunity Thesis has to be restricted to self-ascriptions in such a way that external attributions and disidentifications do not fall within the scope of the thesis. The explanation of Epistemic Immunity in terms of identification-freedom, in contrast, suggests that Immunity should hold for all introspection-based ascriptions equally, be they prototypical self-ascriptions, external attributions or disidentifications. This makes Epistemic Immunity vulnerable to the Argument from Subject-Neutral Immunity and the Argument from Disidentification. Further, Epistemic Immunity is open to the Argument from Identification-Dependence which claims that introspection-based judgments simply aren't identification-free and therefore must be liable to error through misidentification.

The main question of this thesis was whether thought insertion and similar cases of alienation undermine Immunity. I argued in the previous chapter that the pathological cases do undermine Epistemic Immunity, but not Ontological Immunity. In so far as one is looking for the most promising version of Immunity, one should therefore follow the ontological approach. The more general question was whether introspection-based self-ascriptions of mental states are immune to error through misidentification. The previous chapter yielded that introspection-based self-ascriptions are not immune to error through *epistemic* misidentification, but are (in varying degrees) immune to *ontological* misidentification. In the light of this more general question, I now explore a number of further interesting differences between the epistemic and ontological approach. Some of these differences I take to provide further support for the idea that the ontological approach is the preferable one. But mainly, my aim is not to argue for one of the two views, but to spell them out in more detail and explore their implications.

7.1 Cross-Wiring

Let's take another look at cross-wiring in the light of the more nuanced description of the epistemic approach.

Does Cross-Wiring involve Error through Misidentification?

In introducing the idea of cross-wiring (see § 2.5), I have already pointed out that there is a controversy surrounding the question whether cross-wiring cases involve error through misidentification. On the ontological approach the answer is as clear as could be: yes. In cases of cross-wiring the predication information, e.g. the information that legs are crossed, comes from a subject that is distinct from the subject who self-ascribes that predicate. In other words, there is a divergence of source and target and the judgment is therefore in error through misidentification. On the epistemic approach, things are less clear. The discussion of cross-wiring cases will further illuminate the epistemic approach and prepare the presentation of additional interesting corollaries.

There are really two camps among proponents of the epistemic approach, the distinction of which is not always particularly clear. I will call them the *internalist* and the *externalist* camp. Roughly, internalists hold that a judgment is in error through misidentification iff the subject's actual and rationally accessible justification contains a false identity assumption. Externalists, in contrast, hold that a judgment is in error through misidentification iff rejecting an identity assumption would undermine the subject's justification for her original judgment. The internalist view is professed for instance by Evans (1982) and Coliva (2006), the externalist view by Shoemaker (1968) and Pryor (1999).

Although there is a debate between the two camps regarding cross-wiring cases (particularly concerning quasi-memory), there is virtually no systematic description of the differences. One notable exception is Coliva (2006). She distinguishes between error through misidentification relative to a subject's own rational grounds and error through misidentification relative to background presuppositions. This can be translated into my distinction

as follows: members of the internalist camp take the notion of error through misidentification to pertain only to misidentifications within one's own rational grounds. They construe Immunity as a claim that does not rule out misidentifications within a judgment's background presuppositions. Members of the externalist camp, in contrast, take the notion of error through misidentification to pertain to misidentifications both in one's own rational grounds and in the background presuppositions. They construe Immunity as a claim that rules out both misidentifications within one's own rational grounds as well as within the background presuppositions.

The epistemic definitions of error through misidentification are neutral with respect to the difference between externalists and internalists. That is to say, both camps can argue for their views by appealing to both definitions of immunity to error through misidentification (i.e. in terms of identification-freedom and in terms of retreat to existential generalization). Let us first look at the internalist camp. Here is what Evans famously says about cross-wiring cases.

> [T]he possibility of a deviant causal chain, linking the subject's brain appropriately with someone else's body, in such a way that he is in fact registering information from that other body [...] merely shows the possibility of an error; it does not show that ordinary judgements of the kind in question are identification-dependent.
>
> In the first place, we cannot think of the kinaesthetic and proprioceptive system as gaining *knowledge* of truths about the condition of a body which leaves the question of the identity of the body open. If the subject does not know that *he* has his legs bent (say) on this basis (because he is in the situation described), then he does not know *anything* on this basis. (To judge that *someone* has his legs bent would be a wild shot in the dark.) (1982: 221)

Evans maintains that the states are given to the subject as her own (be they cross-wired or not), that the corresponding judgments are identification-free, and that therefore the kind of error that we find in this case cannot be an error through misidentification. He further claims that it is impossible to challenge the judgment in a way that leaves open a retreat to existential

generalization. Either the subject knows that she herself is F, or she doesn't know anything at all. On the internalist view, the error we find in cross-wiring cases is not a misidentification, but comparable rather to an illusion or an hallucination.

Externalists, in contrast, can argue as follows. The fact that we can defeat the self-ascription by challenging an identity assumption simply is what it means for that self-ascription to be based on that identity assumption. Hence, the judgment is identification-dependent and if that identification turns out to be false, as is the case in cross-wiring, then the judgment is in error through misidentification. Basically the same idea has also been put in terms of retreat to existential generalization. It is possible to undermine an introspection-based self-ascription in a way that leaves intact grounds for an existential claim precisely by challenging the identity assumption in question (see e.g. Pryor 1999: 295; cf. also Smith 2006: 279). The idea is this. If my judgment 'my legs are crossed' is challenged by pointing out to me that my proprioceptive perceptions do not (or may not) derive from my body, then I lose my grounds for believing that *my* legs are crossed, but I retain grounds for the claim that *someone's* legs are crossed. (A question that would have to be discussed in more detail is whether this existential claim is really just based on the original grounds or whether it is partially based on the defeater itself.)

The same dispute can also be described in terms of Coliva's distinction between a subject's own rational grounds and the background presuppositions. Introspection-based self-ascriptions may be based on certain identity assumptions (here, the assumption that my proprioceptive perceptions derive from my own body) which turn out to be false in cases of cross-wiring. The point is that these identity assumptions do not figure in a subject's own rational grounds, but only in the background presuppositions. Hence, cross-wiring shows that introspection-based self-ascriptions are vulnerable to error through misidentification in the background presuppositions, but it can be maintained that the judgments are immune to error through misidentification in the subject's own rational grounds (cf. Coliva 2006).

Cross-Wiring and Identification-Freedom

Although Shoemaker and Evans come to different conclusions regarding the question whether cross-wiring involves error through misidentification or not, they both seem to sustain the assumption that identification-freedom entails immunity to error through misidentification. Evans takes memory-, perception- and proprioception-based self-ascriptions to be identification-free and therefore holds that the kind of mistake we find in cases of cross-wiring is not a misidentification. Shoemaker conversely argues that cross-wiring does involve error through misidentification, and that this reveals that the corresponding self-ascriptions are identification-dependent. Pryor, in contrast, challenges the very idea that identification-freedom implies immunity. He maintains the assumption that introspection-based self-ascriptions are identification-free, but takes cross-wiring cases to show nonetheless that they are vulnerable to error through misidentification (in particular, his argument is about memory-based self-ascriptions). As far I know, he is the only proponent of the epistemic approach to explicitly challenge the implication from identification-freedom to immunity.

Let us first look at his argument for the claim that memory-based self-ascriptions are identification-free. It hinges essentially on his notion of identification-dependence. According to the Evansian notion, a judgment is identification-dependent iff it rests on an identity assumption. According to Pryor's notion, in contrast, a judgment is identification-dependent iff *its justification* rests on *justification for an identity assumption*. Now, memory-based self-ascriptions, Pryor argues, are not identification-dependent in that sense, because, to be justified in self-ascribing a remembered past event, one need not have justification for the identity assumption that one's memories derive solely from one's own past. This he takes to show that the justification of memory-based self-ascriptions does not rest on justification for the identity assumption.

At the same time, Pryor argues that in cases of quasi-memory, memory-based self-ascriptions are in error through misidentification. Here, Pryor

appeals to the definition of error through misidentification in terms of existential generalization. His idea is that memory-based self-ascriptions could be challenged by telling the subject "that some of his memories are quasi-memories of events in someone else's past life, and that none of his memories as of being F derive from actual events in his own life." (1999: 295) In this case, Pryor argues, the subject loses her justification that she was F, but retains justification for believing that someone or other was F. And, according to the definition in terms of retreat to existential generalization, this shows that memory-based judgments are vulnerable to error through misidentification. Actually, Pryor's argument is much more sophisticated, but it is not necessary to go into the details here. The crucial question regarding this argument is whether the subject's justification for the existential claim that someone or other was F is really based on the original grounds, and not rather based partly on the defeater itself (which basically tells the subject that she might be cross-wired) (cf. Smith 2006: 279f.). I will not go into this. Rather, I will discuss Pryor's idea of how the two claims go together, that is his idea that memory-based self-ascriptions are identification-free yet liable to error through misidentification at the same time.

It may seem strange that Pryor considers memory-based judgments identification-free. After all, doesn't he assume that it is possible to partially defeat what we may call an identification-component of the judgment's justification, namely the assumption that my memories derive from my own past events? And doesn't that show that the initial judgment was identification-dependent after all? No, Pryor argues. The memory-based judgment is identification-free in the sense that holding it didn't require the subject to also hold justification for the identity assumption which was then brought into question by the defeater. Memory-based judgments are

> identification-free because, although the belief is *vulnerable* to misidentification, the circumstances in which such misidentification would arise are so pathological and rare that it's not a requirement, for you to be justified in believing that *a* is F, that you rule those possibilities out. Hence, you can have identification-free justification for

> a belief, even in cases where the belief still *rests on* certain identity assumptions, and so is still *vulnerable to* misidentification. (1999: 291f.)

So Pryor's idea is this: A belief can be vulnerable to error through misidentification because the belief rests on an identity assumption that can turn out to be false. For the belief to rest on an identity assumption I take to mean here, that if the identity assumption turns out to be false, this fact would undermine the belief. Yet the belief can be identification-free at the same time in the sense that the identity assumption does not figure in the justification of the belief. That is to say, in order to be justified in holding the belief, the subject does not need to hold justification for the identity assumption.

Now, here is a question to Pryor's view. Let us grant that being justified in self-ascribing remembered events *does not require* justification for the identity assumption. The question I would like to ask: are we, as a matter of fact, normally justified in believing the identity assumption? Again, I am granting that such justification is not *required* for the self-ascription of the past event, but I am asking whether we *actually have* such justification. This poses a dilemma for Pryor's view. If Pryor were to grant that *I have justification for believing the identity assumption*, the self-ascription of the past event would no longer be identification-free. For, whatever it is that would justify me in believing the identity assumption, it surely seems that my justification for my self-ascription would also rest on this justification. So, to secure the point that memory-based self-ascriptions are identification-free, Pryor has to hold that *I do not have justification for the identity assumption*. But this is a very strong and very skeptical claim! I think that, in normal situations I do have justification for believing that my memories derive from my own past. Of course, I may not be able to justify this belief in the face of radical Cartesian doubt or in the epistemic threat of being cross-wired. But if that were the challenge, I would neither be able to justify my memory based self-ascription. Hence, there is a tension in Pryor's idea that I do not have justification for the identity assumption, but I do have

justification for the self-ascription, where the self-ascription epistemically rests on the identity assumption.

Did I somehow miss Pryor's point in the quote above, that for my self-ascription to be justified (where this self-ascription rests on an identity assumption) I do not need justification for the identity assumption? I do not think so. Rather, I have exposed a tension in this idea. The reason, given by Pryor, why my self-ascription can be justified without justification for the identity assumption was that the circumstances in which the identity assumption would be false are so rare that I do not need to be able to rule them out. But if the circumstances in which the identity assumption is false are so rare, I reply, then neither do I need to be able to rule out these circumstances to have justification for the identity assumption itself. And if I do have justification for the identity assumption, then certainly my justification for the self-ascription (which rests on the identity assumption!) rests on my justification for the identity assumption.

I will leave it at this point. Certainly, my critique does not amount to a conclusive refutation. However, I believe to have exposed a tension in Pryor's claim that identification-free judgments can be liable to error through epistemic misidentification. My aim was to reinforce the idea, presented earlier, that on the epistemic approach, immunity to error through misidentification and identification-freedom go hand in hand.

Ontologically Deviant Conditions and Epistemically Deviant Conditions

I now come back to a crucial step in the externalists' argument that cross-wired subjects are in error through misidentification. Externalists argue that self-ascriptions can be challenged in a way that leaves intact grounds for an existential claim, namely by raising as a live epistemic possibility that the subject herself may be cross-wired. Only if cross-wiring is considered as a possible scenario does it make sense for a subject to retreat from the self-ascription of a past event to the existential claim that someone or other experienced that past event.

Let me say that a case involves *ontologically deviant conditions* when it actually is a case of cross-wiring and that a case involves *epistemically deviant conditions* when the subject considers cross-wiring as a live possibility. The debate about cross-wiring is mainly concerned with ontologically deviant conditions and the question whether the self-ascriptions in these cases are in error through misidentification. Epistemically deviant conditions are brought into play only in arguing that the ontologically deviant cases really imply an error through misidentification. Now, the interesting thing to note is this. Most writers are in agreement that under epistemically normal conditions introspection-based self-ascriptions are identification-free, but that the self-ascription becomes identification-dependent if the conditions become epistemically unusual, that is to say, as soon as the subject considers cross-wiring as a possibility. For, once a subject considers the possibility of cross-wiring, her subsequent self-ascriptions depend on the assumption that she is in fact not cross-wired, for instance the assumption that her episodic memories derive from her own past experiences. This has a number of interesting implications.

First, this implies that, on the epistemic approach, a judgment's being identification-free or identification-dependent is a context-dependent property. Very generally, identification-freedom depends on the judgment's justification and since justification changes with the context, so does the judgment's property of being identification-free. Note that I am not making the well-registered point that identification-freedom is context dependent in the sense that it depends on the grounds of the judgment. I am here saying that one and the same type of judgment, say the self-ascription of a past event based on one's remembering that event, is identification-free in some contexts, but identification-dependent in other contexts. In epistemically normal conditions the judgment is identification-free; in epistemically deviant conditions the judgment is identification-dependent.

Second, resulting from the context-dependence of identification-freedom, on those views that closely tie identification-freedom to immunity to error through misidentification, also immunity is context-dependence. According to some writers, especially members of the internalist camp, the kind of

mistake we find in typical cross-wiring cases (ontologically deviant but epistemically normal conditions) is not an error through misidentification. However, these writers at the same time agree that epistemically deviant conditions lead to identification-dependent judgments (see e.g. Evans 1982: 185 and 189f., Coliva 2006: 421). Judgments made under epistemically normal conditions are identification-free and therefore immune; judgments made under epistemically deviant conditions are identification-dependent and therefore liable to error through misidentification. That is to say, whether an introspection-based self-ascription in ontologically normal conditions is immune or not depends essentially on the question whether the subject considers cross-wiring cases as possible. Similarly, whether a self-ascription in a cross-wiring scenario is in error through misidentification or not depends on whether the subject considers cross-wiring cases as possible. If a cross-wired subject does not consider cross-wiring and hence self-ascribes the cross-wired state (epistemically normal but ontologically deviant conditions) the kind of mistake will not be an error through misidentification; if, however, a cross-wired subject does consider the possibility of cross-wiring but nonetheless self-ascribes the cross-wired state (epistemically and ontologically deviant conditions) then she will be in error through misidentification.

The only writer who seems to be at least somewhat aware of this context-dependence is Coliva (2006). In making the distinction between a subject's own rational grounds and the background presuppositions, Coliva explicitly admits that by considering the possibility of cross-wiring, an identity assumption can be moved from the background presuppositions to the rational grounds (cf. 2006: 421). Thus, the question whether a misidentification is located within the subject's own grounds or within the presuppositions turns out to be a context-dependent matter. This again implies that Coliva's preferred notion of error through misidentification, error through misidentification relative to the subject's own rational grounds, is context-dependent and that a judgment's being immune or liable to such error depends on the question whether the subject considers cross-wiring as a possibility or not. While I do not find this implication very attractive, it

might not be entirely unmotivated. The fact that error through misidentification is always relative to a judgment's grounds can be spelled out this way. Judgments made under epistemically deviant conditions, on this view, are simply based on different grounds than judgments made under epistemically normal conditions.

However, this context-dependence adds to the notoriously unclear question what it is for a judgment to be identification-free. On Coliva's approach, what corresponds to this notorious question is the question how to draw the line between a subject's own grounds and the background presuppositions. All she says is that the rational grounds are those beliefs a subject would offer in defense of the judgment in question (cf. 2006: 416). But that is not satisfactory. For, which judgments a subject would offer naturally depends on the way in which the judgment is challenged. If a background presupposition gets challenged it would be natural to offer the background presupposition as a ground for the belief, which would then move that assumption from the presuppositions into the rational grounds. Perhaps the idea is that the rational grounds are those which the subject would offer without any particular challenge but when asked simply to justify the belief. But this is not very helpful either. The question what an adequate response would be to the question 'why do you think that p?' depends again on the epistemic context.

Third, the way proponents of the epistemic approach discuss cross-wiring cases supports my discussion of how the epistemic approach should deal with the pathological cases. After all, the pathological cases of alienation can be considered as the reverse scenario of cross-wiring. While typical cases of cross-wiring involve ontologically deviant and epistemically normal conditions, the pathological cases can be described as involving ontologically normal but epistemically deviant conditions. In other words, what patients believe to be the case corresponds to what philosophers imagine to be the case in cross-wiring. Of course, patients are not actually cross-wired, but they believe that they are. The fact that patients believe to be subject to cross-wiring is precisely what makes their ascriptions of thoughts and intentions identification-dependent. If they attribute the thought or

intention to an external agent their judgment is in error through misidentification; if they disown the thought their judgment is in error through disidentification.

Let me round up the discussion by contrasting these implications of the epistemic approach to the ontological approach. On the ontological approach, the questions whether self-ascriptions are in error through misidentification and whether they are immune to such error depend solely on the question whether the scenario is affected by ontologically deviant conditions or could be so affected, respectively. The question whether a subject considers the possibility of deviant conditions does not matter for the immunity of the judgment. This difference between the two approaches can be nicely illustrated by appeal to Bermúdez's explanations of why introspection- and proprioception-based self-ascriptions are immune. Bermúdez explains the immunity of introspection-based self-ascriptions in terms of the "mastery of what might be termed a simple theory of introspection" which "amounts to nothing more than some level of mastery of the *a priori* link between being introspectively aware of a thought and it being the case that one is thinking it" (2003b: 225). Crucially, on Bermúdez's view, it is not the *a priori* link itself which explains Immunity, but the subject's *knowledge* of that link. The same explanation is given with respect to proprioception-based self-ascriptions. After acknowledging the nomological possibility of proprioceptive cross-wiring, Bermúdez maintains:

> This possibility does not cast doubt upon the immunity to error through misidentification of somatic proprioception, given that, as things are, we are not wired up to other bodies and have no reason to think we might be. In fact, it is really the lack of any grounds for thinking that we might be wired up to other bodies that secures immunity to error through misidentification. (ibid.: 226)

So, again, it is *knowledge* of the proprioception-ownership link that secures Immunity rather than the link itself. The difference to introspection-based self-ascriptions is that in this case the "theory of proprioception [is] based on the *de facto* link between ownership and the objects of proprioception" rather than on an *a priori* link (ibid.).

I take it as an advantage of the ontological approach that it does not appeal to the subject's *knowledge* of the introspection-ownership link, but to *the link itself*. Knowledge of that link, as I have argued, is context-dependent and I do not find the context-dependent notion of immunity to error through misidentification attractive. To say the least, the context-dependent epistemic notion of immunity does not capture the idea that in introspection-based self-ascriptions one *cannot go wrong* about whose mental state it is.

7.2 Explaining the Core of Immunity

In the previous section I discussed how the explanatory framework of the epistemic approach deals with cross-wiring cases. The picture just developed yields a further surprise: without the assumption of an introspection-ownership link, the epistemic approach lacks the resources to explain the immunity of the most fundamental cases, self-ascriptions of introspected mental experiences.

For the moment, let us take as a datum that self-ascriptions of introspected mental experiences are immune to error through misidentification. When I speak of the self-ascription of introspected mental *experiences*, I have in mind for instance, the self-ascription of a visual experience based on having that experience, and the self-ascription of a pain based on being in pain. What I have in mind here is the core of Immunity. I am not speaking about the introspection-based self-ascriptions of causal or physical relations (such as: self-ascriptions of standing in a visual perceptual relation to an external object or being the person from whose body the pain causally originates). Taking this immunity as a datum means that, for the moment, we forget about possible counterexamples against this claim. We assume that, if any self-ascriptions are immune, then certainly these are.[108] The question then is, how do we explain their immunity?

[108] Smith, for instance assumes, that "an account of IEM should, at the very least, capture those self-ascriptions that are agreed by all to be central to our conception of ourselves as self-conscious subjects. Specifically, an account of IEM should imply

On the ontological approach, the explanation is straightforward. Suppose I judge that I am in pain based on my experiencing pain. To explain why this judgment is immune, we can first say what an error through misidentification would theoretically look like and then say why it cannot occur. In this case, my self-ascription of pain would be in error through misidentification if, *per impossibile*, I was experiencing and therefore self-ascribing a pain that was actually not my own. Introspection-Ownership, then, explains why it is not possible to experience a pain that is not one's own: with conceptual necessity, if I experience a pain it is my pain.

On the epistemic approach, in contrast, explaining the core of Immunity is not as straightforward. The crucial question is whether cross-wiring cases, and thereby the (impossible) scenario sketched above, involve error through misidentification or not. On the Evansian view, cross-wiring does not count as error through misidentification (see e.g. the undetectable headphones case, 1982: 184–188). In a sense, this view does explain why introspection-based self-ascriptions of mental states are immune. The explanation simply is that the kind of mistake that could occur, if it were possible to be introspectively aware of another person's mental experience, is not an error through misidentification. Note that, on this view, it does not matter whether the suggested case of experiencing another person's pain is possible or not. What matters is that, if such a case were possible, it would not involve a misidentification of who is in pain, but rather an illusion of pain. This smacks of a terminological sleight of hand. Rather than explaining why a certain error cannot occur, this view gives the error in question a different label.

Let me illustrate my complaint with an analogy. Everyone agrees that visually based self-ascriptions are not immune to error through misidentification. Canonical examples of visual error through misidentification are mirror cases. My judgment that I have a stain on my jacket, based on seeing

---

that the self-ascription of occurrent mental episodes (e.g., 'I have a headache') are IEM." (2006: 274). For the idea of taking Immunity as a datum see also Campbell (1999a: 91).

a reflection in a shop window, is liable to error through misidentification because it may be somebody else's reflection that I have seen. Note how the explanation of why visually based judgments are vulnerable to error through misidentification is that a certain scenario is possible, a scenario in which the property that I visually perceive is not actually mine, but somebody else's. My complaint is that we should expect an analogous explanation of the fact that introspection-based judgments are immune to that error. The explanation should have the form that certain scenarios are *not* possible, scenarios in which the property that I introspect is actually somebody else's. This is indeed the explanation given by the ontological approach. But an explanation of this form cannot be given in terms of identification-freedom. The only explanation afforded by the idea of identification-freedom is the one given above, i.e. the idea that when it comes to introspection, self-ascribing an experience that is not one's own cannot be considered a misidentification.

Of course, not all proponents of the epistemic approach are committed to this move. Shoemaker and Pryor accept that cross-wiring leads to error through misidentification. By the same token, they should agree that if, *per impossibile*, a subject were introspectively aware of another person's mental experience, she would be in error through misidentification when self-ascribing that state. Explaining the core of Immunity, on this assumption, means explaining why this scenario is not possible. But the idea of identification-freedom does not get us anywhere in explaining the impossibility of this scenario. For, the identification-freedom of introspection is perfectly compatible with the possibility of being introspectively aware of another person's mental experiences. More than that, if it were possible to be introspectively aware of another person's mental experiences, Self-Awareness and identification-freedom would imply that one would mistake those experiences for one's own. That is to say, Self-Awareness would imply that, in cases of introspective cross-wiring, one would make an error through misidentification.

The only option I see for proponents of the epistemic approach is to supplement their view with the assumption of an introspection-ownership

link. And in fact, this is precisely what Coliva (2002a) does. The problem with this move is that it is not clear what role the idea of identification-freedom is left to play. Appealing to the introspection-ownership link suffices to explain why, in introspection-based self-ascriptions, one cannot go wrong in ascribing the property to the wrong person. The idea of identification-freedom is not necessary for this explanation and does not add anything to it either. If one assumes that introspection-based judgments are identification-free one still needs to explain why it is only one's own states that are accessible to introspection. In that sense, assuming identification-freedom doesn't add anything to the explanation. If, in contrast, one does not assume identification-freedom, one can still explain the core of Immunity in terms of Introspection-Ownership. For, even if introspection-based self-ascriptions are based on an identification, Introspection-Ownership explains why that identification cannot ever be mistaken. In that sense, assuming identification-freedom is not necessary for the explanation.

Note how this point is closely connected to the refined critique of Immunity (see § 5). I just argued that Self-Awareness and identification-freedom cannot explain the immunity of self-ascriptions of occurrent mental experiences. Asking the other way around, what they can explain, leads back to the counterexamples. If introspection-based judgments are identification-free in the sense that introspective awareness guarantees the self-ascription of the state then it should neither be possible to go wrong in ascribing an introspected state to someone else nor should it be possible to go wrong in denying that it is one's own. The assumption of identification-freedom serves as a motivation for a version of Immunity that rules out error through misidentification in external attributions and error through disidentification. Such a version of Immunity could not be explained (or motivated) by Introspection-Ownership alone. But as I argued (§ 6.1), such a version of Immunity is open to counterexamples. Hence, the assumption of identification-freedom is not only superfluous for the explanation of the core of Immunity, the assumption, more than that, affords an 'explanation' and thereby motivation for a version of Immunity that is false.

7.3 The *De Se* Constraint

I introduced in § 2.3 what I call the *de se* constraint, the idea that the scope of the Immunity Thesis is restricted to self-ascriptions in the *de se* mode. In other words, Immunity applies only to I-thoughts, thoughts that would find their natural linguistic expression in terms of the first-person pronoun. As we have seen, the *de se* constraint does not play any interesting role in the discussion of the pathological cases. External attributions clearly fail the *de se* constraint and the controversial question whether disowning claims (construed either as negative self-ascriptions or as disidentifications) satisfy the self-ascription constraint is not illuminated by the fact that these claims are clearly made in a *de se* mode. So far, I have simply assumed the *de se* constraint to be implicitly contained within the self-ascription constraint. I now want to separate and contrast these two constraints in order to explore in more detail what role the *de se* constraint plays in the explanation of Immunity. In particular, I will explore whether it is possible to drop the *de se* constraint. This also means exploring what role the first-person concept plays in the explanation of immunity. The discussion simultaneously illustrates and deepens the ontological explanation of Immunity.

Let me begin by specifying the relation between the *de se* constraint and the self-ascription constraint (henceforth, I will understand the self-ascription constraint as not implying the *de se* constraint). The self-ascription constraint is satisfied whenever a property is in fact ascribed to oneself (we can ignore, for the moment, the question whether negative self-ascriptions of the form 'I am not-F' and disidentifications of the form 'I ≠ the F' count as self-ascriptions in this sense). Crucially, to satisfy the self-ascription constraint a judgment does not have to be made in a *de se* mode, it can be made using a demonstrative, definite description or proper name (or rather, as we are concerned with thoughts, the mental counterparts thereof). Whenever I ascribe a property to myself, be it intentionally (e.g. by thinking 'the most tired person in the office wants to have a coffee', see § 2.3) or accidentally (e.g. by thinking 'the murderer of Laius shall be exiled'), the self-ascription constraint is satisfied.

The *de se* constraint, in contrast, is satisfied when I ascribe a property (to myself) in the *de se* mode (i.e. in a mode that would find its natural linguistic expression using the first-person pronoun). I assumed above that an ascription in the *de se* mode cannot refer to anyone other than oneself (I will defend this assumption shortly). If this is correct, then there are no *de se* ascriptions that are not self-ascriptions, but there are self-ascriptions that are not *de se* ascriptions. This means that whenever the *de se* constraint is satisfied, the self-ascription constrained is satisfied as well, but not the other way around. More precisely, whenever the *de se* constraint is satisfied, the self-ascription constraint cannot fail to be satisfied solely because the self-concept refers to a person other than oneself. It might still fail to be satisfied because the judgment in question is not an ascription in the proper sense. This, I argued, is the case in disowning claims construed as disidentifications ('it's not me who is F'). Such judgments are clearly made in a *de se* mode, but they are not self-ascriptions in the relevant sense (that is to say, the ontological Immunity Thesis does not apply to these claims). For this reason, we cannot drop the self-ascription constraint and determine the scope of Immunity solely in terms of the *de se* (and introspection) constraint. However, might it be possible to drop the *de se* constraint?

De Re Immunity

On the ontological approach, the question whether a judgment is in error through misidentification depends on whether the predicate is ascribed to the right subject. It does not matter whether that ascription is identification-dependent and it does not matter whether the ascription is made in a *de se* mode or not. The reason why introspection-based self-ascriptions cannot be in error through misidentification, I argued, is that one's own property is ascribed to oneself. This suggests that whenever an introspection-based property is self-ascribed—be it in a *de se* mode or not—the property is one's own and hence cannot be in error through misidentification. Thus, it may seem that the self-ascription constraint (in addition to the introspection constraint) does all the essential work and that we can do

without the *de se* constraint. Dropping the *de se* constraint, it seems, would broaden the scope of Immunity. It would mean claiming that all introspection-based self-ascriptions are immune, be they in a *de se* mode or not. I will refer to non *de se* self-ascriptions as *de re self-ascriptions* and I will dub the claim that all introspection-based self-ascriptions are immune the *de re version* of Immunity.[109]

Surprisingly, Shoemaker (1970) makes a similar claim in revoking his (1968) assumption that only *de se* statements are immune.[110]

> I made the mistake of associating [immunity to error through misidentification] with the peculiarities of the first-person pronouns. But in fact present tense statements having the appropriate sorts of predicates are immune to error to [*sic*] misidentification with respect to any expressions that are "self-referring" in the sense of footnote 3, including names and definite descriptions. (1970: 270, fn. 5)

Shoemaker's claim is slightly weaker than the suggestion I just made. As self-referring expressions he has in mind expressions whose "reference is in fact to the speaker" *and which* "the speaker intends in using […] to refer to himself" (1970: 270, fn. 3). That is to say, Shoemaker broadens Immunity to hold for *some de re* self-ascriptions, namely for *intentional* self-ascriptions, but not for *accidental* self-ascriptions. My suggestion was even broader, namely to include *all* self-ascriptions, regardless of the subject's referential intentions. Regrettably, Shoemaker does not argue for or even explain his claim. I already suggested my own argument for *de re* Immunity above: when a subject self-ascribes an introspected property (*de se* or not *de se*)

---

[109] Following Vosgerau (2009b: 107), I am ignoring here the distinction between *de re* and *de dicto* thoughts. I intend the label '*de re* self-ascription' to apply to all non *de se* self-ascriptions, be they in a *de re* mode or in a *de dicto* mode. Note also that *de re* Immunity is not a version of Immunity that applies to *de re* self-ascriptions only, but one that applies to *de re* self-ascriptions *in addition to de se* self-ascriptions.

[110] Not many writers seem to have taken note of this and I know of no other writer discussing the possibility of dropping the *de se* constraint.

she self-ascribes one of her own properties and hence cannot be in error through misidentification. I now discuss a whole row of objections and defenses regarding *de re* Immunity.

Two Counterexamples: Accidental and Intentional de re Self-Ascriptions

Whether we can drop the *de se* constraint depends on the question whether (introspection-based) *de re* self-ascriptions are liable to error through misidentification. Now, when we think of potential misidentifications in *de re* self-ascriptions, two kinds of cases come to mind in which someone self-ascribes a property by using a name, description, or demonstrative: accidental self-ascriptions and intentional self-ascriptions.

In cases of *accidental self-ascriptions*, a subject ascribes a property by means of a name, description or demonstrative which, unknowingly to the subject, in fact refers to the subject herself (see e.g. Perry's (1979) sugar case[111]). We can slightly alter such cases so that they involve error through misidentification.

> GLASSES. I believe that you are the tallest person in the room and, seeing that you wear glasses, I judge that the tallest person in the room wears glasses. However, in fact, I am the tallest person in the room, so that, accidentally, I judge that I wear glasses.

This is a harmless case, and obviously not a counterexample to *de re* Immunity for the imagined case does not satisfy the introspection constraint. It may seem impossible that a person accidentally self-ascribes an introspected state, but actually it is not. Consider a variation of Perry's amnesic Lingens.

> DELUDED AMNESIAC. Rudolf Lingens is a deluded amnesiac. He is deluded in that he suffers from thought insertion and he is amnesic in that he does not know that he is Rudolf Lingens. On one occasion, he

[111] Seeing a trail of sugar in the supermarket, Perry thinks 'the shopper with the torn bag is making a mess', just to find out later that he is the shopper with the torn bag himself.

> believes of a thought which he is introspectively aware of that it is Rudolf Lingens's thought.[112]

In this case, Lingens accidentally attributes an introspected thought to himself. But this is not a counterexample to *de re* Immunity either, for the judgment is not in error through misidentification. After all, it is his own thought which he ascribes to himself. The two cases illustrate a general idea why we do not find introspection-based accidental *de re* self-ascriptions that are in error through misidentification. If knowledge of the property is based on introspection, it is one's own property. And when an own property is self-ascribed (accidentally or not), there is no error through misidentification.[113]

In cases of *intentional de re self-ascriptions*, a subject intentionally ascribes a property to herself by using a non *de se* singular concept (i.e. the mental counterpart to a name, description, or demonstrative). Here is a case that may seem to be in error through misidentification.

> COFFEE BREAK. Wanting to take a coffee break and believing that I am the most tired person in the office I judge that the most tired person in the office wants to take a coffee break. However, as a matter of fact, somebody else is the most tired person in the office.

But again, we do not actually get a *de re* self-ascription that is in error through misidentification. The crucial question is whom my descriptive concept 'the most tired person' refers to. If it refers, as I intend, to myself

[112] The original case of amnesic Lingens (without the delusion) is due to Perry 1977.

[113] We could turn DELUDED AMNESIAC into a counterexample against Immunity by assuming that Linigens is not only deluded and amnesic, but also subject to cross-wiring. In that case, the thought in question would not actually be Lingens's thought, but someone else's. But such a counterexample would merely exploit the limitations of the introspection-ownership link. Since the same type of case can be made with a *de se* self-ascription, it would not show anything particular about *de re* self-ascriptions. In other words, such a case would not be a particular counterexample against *de re* Immunity, it would not speak against the idea of dropping the *de se* constraint.

(in spite of me not fitting the description) there is no error through misidentification because I successfully ascribed the property to the right subject. If, in contrast, the description refers to the actually most tired person, there is error through misidentification alright, but the judgment is (in spite of my intentions) not a self-ascription. Rather, it is a case in which the subject intends *but fails* to make a *de re* self-ascription. Finally, suppose the description 'the most tired person' refers *both* to myself (via intention) *and* to the most tired person (via description). On this view, would the case be a *de re* self-ascription that is in error through misidentification? No. On the assumption that the judgment is ambiguous, one can reply as follows. Insofar as the judgment is a self-ascription, it is not in error through misidentification; and insofar as the judgment is in error through misidentification, it is not a self-ascription.[114] Thus, on none of the possible views regarding the reference of intended *de re* self-ascriptions do we get a *de re* self-ascription that is in error through misidentification.

The Indirect Argument

It looks as though we do not find an introspection-based *de re* self-ascription that is in error through misidentification, neither in cases of accidental self-ascription, nor in cases of intentional self-ascriptions.[115] However, we do not actually need such a case to refute *de re* Immunity. To prove the necessity of the *de se* constraint (over and above the introspection and self-ascription constraint), one does not have to find a *de re* self-ascription that *is in* error through misidentification. This, I just argued, is not possible. Rather, one only has to find a *de re* self-ascription that *is liable to* error through misidentification. Arguing against *de re* Immunity based on this insight is what I will call the *indirect argument against de re Immunity*.

[114] I will discuss this case in more detail below.

[115] More precisely, we do not find a judgment of this type in which the *de re* mode is essential to the case.

Now, the difference between a judgment's *being in* error through misidentification and a judgment's *being liable to* error through misidentification should be clear. A judgment is in error through misidentification when—depending on one's approach—that token judgment is based on a false identity assumption, or source and target of that token judgment diverge. For a judgment to be liable to error through misidentification, in contrast, means that it is possible that a judgment of that type (same type of belief based on the same type of grounds) is in error through misidentification. Thus, to show that *de re* self-ascriptions are liable to error through misidentification, we only need to show that judgments *of that type* can be in error through misidentification. And this can be shown for both kinds of *de re* self-ascriptions.

First, consider Lingens's accidental self-ascription. It is correct, in a sense, that Lingens's *de re* self-ascription cannot be in error through misidentification: whenever *Lingens himself* judges that an introspected thought is Lingens's thought, he cannot be in error through misidentification. However, the type of judgment that he makes is still liable to error through misidentification. The type of judgment is the belief 'This is Lingens's thought' based on introspective awareness of the thought in question. Now, whenever anybody other than Lingens makes this type of judgment, they are in error through misidentification. And this means that also Lingens's judgment, and accidental *de re* self-ascriptions in general, are liable to error through misidentification.

The same goes for intentional self-ascriptions. The type of judgment made in COFFEE BREAK is the belief 'the most tired person wants to take a coffee break' based on one's desire for a coffee break (and on one's belief that one is the most tired person). In some circumstances, when I am in fact the most tired person, I successfully self-ascribe the desire. In those circumstances the judgment satisfies the self-ascription constraint and therefore this type of judgment falls within the scope of *de re* Immunity.[116] However,

[116] Whether this judgment really satisfies the introspection constraint is debatable. I will come back to this question.

in other circumstances (when I am not the most tired person) judgments of this type are in error through misidentification (I am assuming for now that the description 'the most tired person' refers solely to the actually most tired person). Hence, also intentional *de re* self-ascriptions are liable to error through misidentification.

But wait! The *de re* ascriptions that are in error through misidentification are not self-ascriptions. How can they refute *de re* Immunity when they fail the self-ascription constraint and thus do not even fall within the scope of Immunity? It is correct that these cases are not *direct* counterexamples against *de re* Immunity in the sense of being *de re* self-ascriptions that *are in* error through misidentification. They are indirect counterexamples in the sense of showing that *de re* self-ascriptions *are liable to* error through misidentification. In other words, *de re* self-ascriptions are a type of judgment (simply: *de re* ascription) that is liable to error through misidentification, even if, *when* an introspection-based *de re* ascription is a *self*-ascription, it cannot be in error through misidentification. It may sound contradictory to say that *de re* self-ascriptions *cannot be in* error through misidentification *but still are liable to* error through misidentification. But that is just because not all judgments of the relevant type are self-ascriptions.

Let me illuminate this a bit more. A token judgment is immune to error through misidentifications when it is not possible to make a judgment of that type which is in error through misidentification. Immunity is a claim about types of judgments in the sense that for a token judgment to be immune means that all judgments of that type cannot be in error through misidentification. Underlying my argument is an assumption about the way we individuate types of judgments. The crucial point is that judgments cannot be individuated by appeal to the self-ascription constraint. That constraint merely limits the scope of the thesis: any token judgment that satisfies the scope criteria falls within the scope of the thesis. Types of judgments are individuated in terms of the content of the belief and the type of grounds the belief is based on. In particular, I am assuming that the judgment '*I* want a coffee' is of a different type than the judgment '*MS*

wants a coffee' (either because the involved beliefs have different contents or because the judgments involve different grounds – the latter judgment involves an identity assumption). The introspection-based belief 'I want a coffee' is immune: whenever anybody makes that judgment, it cannot be in error through misidentification. The introspection-based belief 'MS wants a coffee' is not immune, it would be in error through misidentification when someone other than myself makes that judgment. All judgments of the type 'I want a coffee' are self-ascriptions, but not all judgments of the type 'MS wants a coffee' are self-ascriptions.

My argument contradicts Shoemaker's claim that intentional *de re* self-ascriptions are immune to error through misidentification. More precisely, my argument contradicts not Shoemaker's claim, which is a claim on the level of language, but the equivalent to Shoemaker's claim on the level of thought. Since Shoemaker does not provide any support for his claim, I do not really know how to argue with him on this point. Let me at least discuss his example, if only to illustrate my claim once more.

> If someone says "De Gaulle intends to remove France from NATO," and is using "De Gaulle" to refer to himself, his statement is in the relevant sense immune to error through misidentification, regardless of whether he is right in thinking his name is "De Gaulle" and that he is the President of France. (1970: 270, fn. 5)

I am puzzled by this claim. If the speaker is a deluded subject who is not De Gaulle, doesn't he say something false of De Gaulle? And doesn't he thereby make an error through misidentification? One way in which one could make sense out of this claim is by assuming that in virtue of the speaker's *intention* of using the name to refer to himself, the speaker does in fact refer to himself. I will shortly argue that such a move cannot succeed on the level of thought. Another way to make sense of this claim is to maintain that the kind of mistake that a deluded subject would make would not count as error through misidentification. In virtue of the subject's intention to refer to himself, the judgment would be identification-free and hence any resulting mistake cannot be an identification mistake. However, this is quite counter-

intuitive and also contradicts Shoemaker's own definition of error through misidentification. For, the speaker knows of himself that he wants to remove France from NATO, but makes the mistake of asserting 'De Gaulle wants to remove France from NATO' because, and only because, he mistakenly thinks 'De Gaulle' refers to himself (cf. Shoemaker 1968: 557).

Rather than speculating on what Shoemaker may have had in mind here, let me use the case to illustrate my own view again. Note that, now, we are back to discussing the issue on the level of thought, not language. I claim that *de re* self-ascriptions are liable to error through misidentification. The idea is this. It is true that, whenever De Gaulle self-ascribes a property using his name, such as 'De Gaulle intends so-and-so', he cannot be in error through misidentification. However, this does not make the judgment immune to error through misidentification, for it is possible that this type of judgment is made under circumstances in which it is in error through misidentification. For instance, when a deluded subject intends to self-ascribe an intention by thinking 'De Gaulle intends so-and-so' the judgment is in error through misidentification. Hence, *de re* self-ascriptions, even if they are intentional self-ascriptions, are liable to error through misidentification.

Objection: De Re Modes of Reference Fail the Introspection Constraint

Finally, there is another, much more promising way of defending *de re* Immunity against the indirect argument and the cases I suggested, namely by appeal to the introspection constraint. I have presented cases in which knowledge of the property in question (wanting to take a coffee break or being the thinker of a thought) is plausibly based on introspection. But what about the mode of reference? Is the non *de se* mode of reference in these cases based on introspection?

In COFFEE BREAK, it obviously is not. Knowledge of the coffee desire is based on introspection, but my belief that I am the most tired person in the office could hardly be based on introspection. Another way of putting the objection quite simply: my judgment that the most tired person in the office wants a coffee is based on the identity assumption 'I am the most tired

person' which cannot be based on introspection. The point generalizes to other intentional *de re* self-ascriptions. If I intend to refer to myself using a name, I must implicitly hold the belief that I am the person to whom this name refers. If I intend to refer to myself using a description, I must implicitly hold the belief that this description applies to myself. If I intend to refer to myself using a visual or other external demonstrative, I must implicitly hold the belief that this demonstrative refers to myself. All these implicit beliefs cannot be based solely on introspection.[117] Hence, it seems that intentional *de re* self-ascriptions do not fall within the scope of Immunity to begin with and hence cannot be a counterexample to *de re* Immunity.

Things are less clear in the case of accidental self-ascriptions. In accidental *de re* self-ascriptions, the subject does not hold similar implicit beliefs. In the classic cases of accidental self-ascriptions, neither the mode of reference nor knowledge of the property are based on introspection (see judgments such as 'the shopper with the torn sack is making a mess' (cf. Perry 1979) or 'that person is a shabby pedagogue' (cf. Mach 1922)). We had to turn to pathological cases such as DELUDED AMNESIAC to get a case in which knowledge of the property is based on introspection. But in these cases, it is not clear at all what the mode of reference is based on. Is it possible to be *introspectively* aware of a thought, for instance, as Lingens's thought? I believe that many writers do not find this idea very plausible.

The discussion must at this point remain inconclusive. I content myself with having shown that the discussion of *de re* Immunity boils down to the question whether *de re* self-ascriptions can be fully based on introspection.

---

[117] Or can they? Perhaps there are introspection-based names, descriptions, or demonstratives that refer to oneself. Perhaps one can intend to refer to oneself with a name given to oneself in a mental act of baptizing, or with a mental demonstrative such as 'this subject of experience' or with the description 'the person whose thoughts I am aware of'. But when these introspection-based modes of reference are used, it is not clear that they can ever refer to anyone other than oneself , which they would have to do to figure in counterexamples.

If they can, we can come up with counterexamples against *de re* Immunity which show that the *de se* constraint is necessary. If *de re* self-ascriptions cannot be fully based on introspection, all putative counterexamples fail the introspection constraint and hence do not undermine *de re* Immunity. However, in that case the *de se* constraint is not *false* (in the sense of overly narrowing down the scope of Immunity), but rather redundant. That is to say, if there cannot be introspection-based *de re* self-ascriptions we do not *need* the *de se* constraint, in the sense that it does not make a difference to the scope of the thesis (given the introspection- and self-ascription constraint). But it doesn't do any harm either, i.e. it does not exclude any judgments from the scope of Immunity that would not also be excluded by the introspection and self-ascription constraints. Given this situation, either the *de se* constraint is necessary or it is redundant, but it certainly isn't wrong (in the sense of being overly limiting). Thus, it makes sense to simply stick with the *de se* version of Immunity.

Reference of the Self-Concept

I assumed throughout that the self-concept necessarily refers to the thinker and that therefore I-thoughts (thoughts in the *de se* mode) necessarily are about the thinker. Let me end by defending this view against a possible objection. Above we considered cases in which a singular non *de se* concept is used with the intention to refer to oneself. We now consider the reverse, i.e. the case in which the self-concept is used with the intention to refer to someone other than oneself.

Here is an argument by Coliva (2003) which attempts to show that the first-person can be used with the intention to refer to someone other than oneself. Crucially, the argument does not concern the self-*concept* but the first-person *pronoun*. She appeals to a case originally presented by Rovane:

> Suppose I am facing a mirror and I believe that I see my own reflection when I really see someone else's. On the basis of what I see reflected in the mirror I say, "There's an incredibly tasteless painting hanging on the wall directly behind me." Because I believe that I am the person reflected in the mirror, I take 'me' to refer to the person

> reflected in the mirror. (Rovane 1987: 153f; quoted from Coliva 2003)

One can ask two questions about this case: Whom does the speaker *intend* to refer to, and whom does she *actually* refer to? Coliva argues that intention (or, speaker reference) and actual linguistic reference come apart in this case. She argues, on the one hand, that the judgment "is affected by error through misidentification precisely because it is a statement about *me*—and not about the person reflected in the mirror" (2003: 424). She argues, on the other hand, that the speaker's "original referential *intentions* in using 'I' were *primarily* directed towards that person [reflected in the mirror], whom she mistakenly took to be herself." (ibid.: 427). She thus concludes that intention and actual linguistic reference of the first-person can come apart.

But how does this apply to I-*thoughts* and the self-*concept*? It is hard to make sense of a distinction between intended and actual reference in thought. In fact, it may be that in thought nothing corresponds to semantic reference and that intention is key in determining reference. On Coliva's assumption that the subject's *primary intentions* are to refer to the person in the mirror, one may be tempted to think, then, that the self-concept in the thought 'There is a painting behind me' actually refers to the person in the mirror rather than to the thinker herself.

I do not find this argument very plausible. Surely, the thinker intends to refer to the subject in the mirror, but the thinker clearly *also* intends to refer to herself. Obviously, the judgment 'There is a painting behind me' is based on the identity assumption 'I = that person (seen in the mirror)' and on the belief 'there is a painting behind that person'. Hence, the referential intention in the final judgment is sort of ambiguous: the thinker assumes that she *is* the person in the mirror, hence she wants to make a claim about both herself *and* the person in the mirror. Putting the same point a bit stronger: given the thinker's identity assumption, it does not even make sense to ask of the thinker which one of the two persons she intends to refer to. How do we resolve the case then?

One option would be to say that the first-person concept in the thought 'there is a painting behind me' refers to both the thinker and the person in the mirror. Is it plausible to assume that a singular thought refers to two objects? Surely this is possible. Mistaken identity beliefs of the form '*a* = *b*' are singular thoughts that refer to two different objects. If a singular thought of the form 'F(*a*)' is based on such an identity belief it can inherit, so to speak, the dual reference of that identity assumption. On that view, the first-person concept is guaranteed to refer to oneself, although it can at the same time refer also to someone else.

A more straightforward option would be to say that, strictly speaking, there simply is not *one thought* that corresponds to the sentence 'there is a painting behind me' as uttered with the intention of referring to the person in the mirror.[118] On the level of thought, what corresponds to that sentence, are really two thoughts: 'there is a painting behind me' and 'I = that person'. On this analysis, the first thought captures the way in which the sentence is about the speaker and the second thought captures the way in which the sentence is uttered with the intention of referring to the person in the mirror. On that view, there is no question that the self-concept always refers to the thinker. The thought 'there is a painting behind me' really is solely about the thinker. The thinker's intention to also refer to the person in the mirror is captured by the fact that the thought is based on and held simultaneously with the identity belief 'I = that person'.

A corresponding analysis can now be given of intentional *de re* self-ascriptions. Again, the idea is that, strictly speaking, one cannot simply think 'the most tired person wants a coffee' or 'De Gaulle intends so-and-so' with the intention to refer to oneself. In thought, what corresponds to the sentence 'the most tired person wants a coffee' as uttered with the intention of referring to oneself are really two thoughts: 'the most tired person wants a coffee' and 'I = the most tired person'. On the one hand, this analysis underwrites the indirect argument from intentional *de re* self-ascriptions. For, when intentional *de re* self-ascriptions are analyzed as a *de re* predica-

[118] Thanks to Gottfried Vosgerau for this suggestion.

tion that is accompanied by an identity assumption, it is clear that this identity assumption is liable to error through misidentification. On the other hand, this analysis underwrites also the objection from the introspection constraint. It brings out that the putative counterexamples to *de re* Immunity are based on identity assumptions that cannot be known introspectively.

Now, the general idea behind this approach is to do with the essentially indexical nature of the self-concept, that is its being closely tied to certain ways of gaining information and to certain ways in which it motivates action (cf. Perry 1979). Consider Kaplan's example of seeing the reflection of a person whose pants are on fire (cf. 1989: 533). The question whether this leads me to think 'his pants are on fire' or 'my pants are on fire' makes an essential difference to how I react to the situation. All and only *de se* thoughts have particularly direct behavioral implications. This motivates the idea that *de re* self-ascriptions that are intended to be about oneself are really hidden *de se* thoughts (or accompanied by implicit *de se* thoughts). For instance, thinking the thought 'De Gaulle intends to remove France from NATO' with the intention to refer to oneself really means thinking two thoughts simultaneously 'De Gaulle intends to remove France from NATO' and 'I = De Gaulle'. Without the identity belief, the judgment could not have the self-specific behavioral implications which it must have in order for us to say that the judgment is made with the intention to refer to oneself. Similarly, thinking an I-thought with the intention to refer to someone picked out in a *de re* mode must be analyzed as involving an implicit identity assumption. For instance, thinking 'there is a painting behind me' with the intention to refer to the person in the mirror really means thinking the two thoughts 'there is a painting behind me' and 'I = the person in the mirror'. This way of analyzing the ambiguous cases captures the essentially indexical nature of I-thoughts and is compatible with the idea that the self-concept cannot fail to refer to the thinker.

Summary and Upshot

I discussed at length the idea of *de re* Immunity, i.e. the idea of dropping the *de se* constraint. Whether that is possible depends on whether *de re* self-ascriptions can be fully based on introspection. Interestingly, it turned out that the pathological cases have some bearing on that question. For instance, a potential counterexample against *de re* Immunity depends on whether it is possible to be *introspectively* aware of a thought *as someone else's thought*. However, I did not take a stand on this question. For, even if the *de se* constraint *can* be dropped, this is only in the sense that it is redundant, not that it is overly limiting. Keeping the *de se* constraint thus doesn't do any harm and, to be on the safe side, I conclude we should simply do so.

Explaining Immunity means explaining why certain types of judgments cannot be in error through misidentification. On the ontological approach, the explanation is that, first, whenever the ascription of a mental state is based on introspection of that state, that mental state is the subject's own state, and second, any ascription of that state that is in a *de se* mode necessarily refers to the subject herself. Introspection-based *de se* ascriptions of mental states therefore cannot involve a divergence of source and target, which means that they are immune to error through misidentification. This way of appealing to the *de se* constraint dovetails with the idea that whether a judgment is in error through misidentification is solely determined by actual source and target, rather than by the subject's way of picking out the target or the subject's intended target. The role of the *de se* constraint merely is to guarantee the reference to the thinker.

The discussion further illuminates the question whether the meaning of the first-person plays a role in the explanation of Immunity. Campbell for instance argues that "the explanation of [First-Person Immunity] does not lie in an account of the meaning of the first person. It has to do rather with the idiosyncrasies of our ways of finding out about psychological states." (1999a: 91) I completely agree that Immunity is primarily secured by the introspection-ownership link, or, as Campbell puts it, by the idea that "we

could have a way of finding out about particular properties which was, as a matter of logic, confined to finding out about the properties of just one object." (ibid.: 93) However, I maintain that the meaning of the first-person also plays role in the explanation of Immunity. The role is not that the first-person secures Immunity in virtue of having a descriptive content (this is the idea Campbell is arguing against). Rather, the first-person plays a role in securing that any *de se* ascription is in fact a self-ascription.

8. Conclusion

The question of this work was whether the Immunity Thesis holds, in particular whether certain pathologies of alienation present counterexamples to the thesis. My result is not a simple yes or no. Rather, I have offered a conditional defense of Immunity. I agreed that pathological alienation presents a serious challenge for the dominant, epistemic approach to Immunity, which ties immunity to identification-freedom. I argued, however, that an alternative, ontological approach to Immunity is available, which is not affected by the cases of alienation. Hence, construed on the ontological approach, introspection-based self-ascriptions of mental states are immune to error through misidentification.

Three points played a major role in reaching this conclusion. First, of course, the distinction between the epistemic and the ontological approach to Immunity. Second, the idea that cases of alienation challenge Epistemic Immunity by showing in different ways that introspection-based judgments are identification-dependent. Third, the idea that the ontological approach allows us to distinguish between the self-ascriptions of different kinds of mental states and to relativize the modal strength of immunity claims accordingly. Here is how all this figures in the broader context.

I discussed four potential counterexamples to the Immunity Thesis: thought insertion, anarchic hand syndrome, made impulses, and somatoparaphrenia. Within these cases I distinguished between external attributions and disowning claims. In a first step, I defended Immunity against the counterexamples as follows. While external attributions are in error through misidentification, they do not fall within the scope of Immunity since they are not *de se* self-ascriptions. And while disowning claims, construed as negative self-ascriptions ('I am not-F'), arguably do fall within the scope of Immunity, they are not in error through misidentification, but rather in error through mispredication. Hence my first result is that, on this simple view, neither disowning claims nor external attributions undermine the Immunity Thesis.

In a second step, I developed a refined critique of Immunity which consists of three arguments. The Argument from Subject-Neutral Immunity urges that, given an explanation of Immunity in terms of identification-freedom, we should expect Immunity to hold not only for *de se* self-ascriptions, but also for external attributions (even if vacuously so). But construed as a thesis that holds also for external attributions, Immunity is open to refutation by the pathological cases. Second, the Argument from Disidentification urges us to construe disowning claims as having the form 'it is not me who is F', that is as being based on a falsely negated identity belief. The pathological cases therefore show that introspection-based judgments about mental states are liable to what I called error through disidentification. Finally, the Argument from Identification-Dependence takes the pathological cases to show that, contrary to the traditional view, introspection-based self-ascriptions of mental states simply aren't identification-free.

My reply to the refined critique is based on the distinction between the epistemic and the ontological take on error through misidentification. On the epistemic approach, a judgment is in error through misidentification when the justification involves a mistaken identification component and a judgment is immune to such error when it is identification-free. The pathological cases show that introspection-based judgments are not identification-free in the required sense (Argument from Identification-Dependence). Further, the explanatory background of Epistemic Immunity suggests to construe Immunity as a thesis, according to which error through misidentification in external attributions and error through disidentification is not possible. Hence, Epistemic Immunity is also susceptible to the Argument from Subject-Neutral Immunity and the Argument from Disidentification.

On the ontological approach, in contrast, a judgment is in error through misidentification iff the source of the predication information is distinct from the object to which the predicate is applied. The assumption of an introspection-ownership link explains why introspection-based self-ascriptions cannot be in error through misidentification in this sense: introspection gives access to only one's own mental states. Ontological

Immunity does not make any claims about identification-freedom and is most naturally construed as a thesis that does not make any claims about external attributions and about disidentification judgments. Hence, Ontological Immunity is not affected by the refined critique. Given the ontological approach, my answer to the main question is: yes, introspection-based self-ascriptions of mental states are immune to error through misidentification.

As a third result, I specified this claim. We need to distinguish between self-ascriptions of occurrent mental experiences and self-ascriptions of mental states more broadly construed, such as factive perceptual states and causal relations. Depending on the kind of mental state that is self-ascribed, Immunity holds with different degrees of modal strength. Self-ascriptions of a mental experience that are based on being introspectively aware of that experience, or, what I take to amount to the same thing, that are based on *having* that experience, are logically immune to error through misidentification. In contrast, self-ascriptions of causal or physical relations which are based on introspective awareness of a mental experience are just *de facto* immune. That is to say, roughly, that they are immune given that the mental experience does not have a causally deviant history.

The idea that different kinds of self-ascriptions enjoy different degrees of immunity is not particularly new. However, the ontological approach dovetails particularly well with this idea and affords a new explanation of it. Introspective awareness of a mental experience is linked to one's *having* that experience with conceptual necessity. Hence, the corresponding self-ascriptions are logically immune. In contrast, introspective awareness of a mental experience is linked to standing in a certain physical or causal relation only given that there is no cross-wiring (or other kinds of deviant causal chains). Hence, the corresponding self-ascriptions are only *de facto* immune.

The distinction between the two kinds of self-ascriptions also fits my discussion of thought insertion. I have argued at length for the idea that the notion of 'mineness' that we should apply to our analysis of thought

insertion (i.e. thought-authorship) is a causal notion. Hence, while the introspection-based self-ascription of a thought in the technical sense of thought-ownership is logically immune to error through misidentification, the introspection-based self-ascription of a thought in the technical sense of thought-authorship is only *de facto* immune. Given that the scenarios envisioned by subjects of thought insertion are conceptually possible, we can conceive of introspection-based self-ascriptions of authorship that are in error through misidentification. But this does not add a new challenge to the Immunity Thesis since, on its most plausible reading, it claims logical immunity only for introspection-based self-ascriptions of occurrent mental experiences.

In thought insertion, anarchic hand syndrome, made impulses, and somatoparaphrenia subjects fail in one way or another to properly recognize their own mental states. These cases provide a stimulating background for the discussion of Immunity and thereby bring out certain details and restrictions of the thesis that may not have been recognized before. Nonetheless, they do not undermine what I argued to be the most plausible version of the Immunity Thesis, Ontological Immunity.

References

Anscombe, G. E. M. (1975): "The First Person." In: Samuel Guttenplan (ed.): *Mind and Language*. Oxford: Oxford University Press, 45–65.

Bayne, Tim & Pacherie, Elisabeth (2007): "Narrators and comparators: the architecture of agentive self-awareness." *Synthese* 159 (3), 475–491.

Bermúdez, José Louis (2000): *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Bermúdez, José Louis (2003a): "I-Thoughts and Explanation: Reply to Garrett." *The Philosophical Quarterly* 53 (212), 432–436.

Bermúdez, José Louis (2003b): "The Elusiveness Thesis, Immunity to Error through Misidentification, and Privileged Access." In: Brie Gertler (ed.): *Privileged Access*. Aldershot: Ashgate, 213–231.

Bortolotti, Lisa & Broome, Matthew (2009): "A role for ownership and authorship in the analysis of thought insertion." *Phenomenology and the Cognitive Sciences* 8, 205–224.

Bottini, Gabriella; Bisiach, Edoardo; Sterzi, Roberto; Vallar, Giuseppe (2002): "Feeling touches in someone else's hand." *Neuroreport* 13 (2), 249–252.

Campbell, John (1997): "Sense, Reference and Selective Attention." *Aristotelian Society Supplementary Volume* 71 (1), 55–74.

Campbell, John (1999a): "Immunity to Error through Misidentification and the Meaning of a Referring Term." *Philosophical Topics* 26 (1&2), 89–104.

Campbell, John (1999b): "Schizophrenia, the Space of Reasons, and Thinking as a Motor Process." *Monist* 82 (4), 609–625.

Campbell, John (2002): "The Ownership of Thoughts." *Philosophy, Psychiatry, & Psychology* 9 (1), 35–39.

Carruthers, Glenn (2012): "A metacognitive model of the sense of agency over thoughts." *Cognitive Neuropsychiatry* 17 (4), 291–314.

Chen, Cheryl K. (2011): "Bodily Awareness and Immunity to Error through Misidentification." *European Journal of Philosophy* 19 (1), 21–38.

Christofidou, Andrea (1995): "The Demand for Identification-Free Self-Reference." *Journal of Philosophy* 92, 223–234.

Coliva, Annalisa (2002a): "Thought Insertion and Immunity to Error Through Misidentification." *Philosophy, Psychiatry, & Psychology* 9 (1), 27–34.

Coliva, Annalisa (2002b): "On What There Really Is to Our Notion of Ownership of a Thought." *Philosophy, Psychiatry, & Psychology* 9 (1), 41–46.

Coliva, Annalisa (2003): "The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee." *Journal of Philosophy* 100 (8), 416–431.

Coliva, Annalisa (2006): "Error Through Misidentification: Some Varieties." *Journal of Philosophy* 103 (8), 403–425.

Daprati, E.; Franck, N.; Georgieff, N.; Proust, Joëlle; Pacherie, Elisabeth; Dalery, J.; Jeannerod, Marc (1997): "Looking for the agent: an investigation into consciousness of action and self-consciousness in schizophrenic patients." *Cognition* 65, 71–86.

Desmurget, Michel (2009): "Movement Intention After Parietal Cortex Stimulation in Humans." *Science* 324, 811-813.

Eilan, Naomi; Marcel, Anthony; Bermúdez, José Louis (1995): "Self-Consciousness and the Body: An Interdisciplinary Introduction." In: José Louis Bermúdez, Anthony Marcel & Naomi Eilan (eds.): *The Body and the Self*. Cambridge, MA: Bradford, 1–28.

Evans, Gareth (1982): *The Varieties of Reference*. Oxford: Oxford University Press.

Feinberg, Irwin (1978): "Efference Copy and Corollary Discharge." *Schizophrenia Bulletin* 4, 636–640.

Fernández, Jordi (2010): "Thought Insertion and Self-Knowledge." *Mind & Language* 25 (1), 66–88.

Fink, Thomas & Mao, Yong (1999): "Designing tie knots by random walks." *Nature* 398, 31–32.

Frankfurt, Harry G. (1988): "Identification and Externality. In: Harry G. Frankfurt: *The Importance of What we Care About*. New York, NY: Cambridge University Press, 58–68.

Frith, Chris (1992): *The Cognitive Neuropsychology of Schizophrenia*. Hillsdale, NJ: Lawrence Erlbaum.

Gallagher, Shaun (2000): "Self-Reference and Schizophrenia: A Cognitive Model of Immunity to Error through Misidentification." In: Dan Zahavi (ed.): *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-Experience*. Amsterdam: John Benjamins Publishing, 203–239.

Gallagher, Shaun (2004a): "Neurocognitive Models of Schizophrenia: A Neurophenomenological Critique." *Psychopathology* 37, 8–19.

Gallagher, Shaun (2004b): "Agency, ownership and alien control in schizophrenia." In: P. Bovet, Josef Parnas & Dan Zahavi (eds.): *The structure and development of self-consciousness: Interdisciplinary perspectives*. Amsterdam: John Benjamins Publishing, 89–104.

Gallagher, Shaun (2007a): "Sense of agency and higher-order cognition: Levels of explanation for schizophrenia." *Cognitive Semiotics* 0, 32–48.

Gallagher, Shaun (2007b): "The Natural Philosophy of Agency." *Philosophy Compass* 2, 1–11.

Gerrans, Philip (2001): "Authorship and Ownership of Thoughts." *Philosophy, Psychiatry, & Psychology* 8 (2), 231–237.

Gibbs, Paul J. (2000a): "Thought Insertion and the Inseparability Thesis." *Philosophy, Psychiatry, & Psychology* 7 (3), 195–202.

Gibbs, Paul J. (2000b): "The Limits of Subjectivity: A Response to the Commentary." *Philosophy, Psychiatry, & Psychology* 7 (3), 207–208.

Graham, George (2004): "Self-Ascription." In: Jennifer Radden (ed.): *Philosophy of Psychiatry*. Oxford: Oxford University Press, 89–105.

Graham, George & Stephens, G. Lynn (1994): "Mind and Mine." In: George Graham & G. Lynn Stephens (eds.): *Philosophical Psychopathology*. Cambridge, MA: Bradford, 92–109.

Hamilton, Andy (2007): "Memory and self-consciousness: immunity to error through misidentification." *Synthese* 171 (3), 409–417.

Hoerl, Christoph (2001): "On Thought Insertion." *Philosophy, Psychiatry, & Psychology* 8 (2), 189–200.

Hogan, Melinda & Martin, Raymond (2001): "Introspective misidentification." In: Andrew Brook & DeVidi Richard (eds.): *Self-reference and Self-awareness*. Amsterdam: John Benjamins Publishing, 205–213.

Horgan, Terence; Tienson, John; Graham, George (2003): "The Phenomenology of First-Person Agency." In: Sven Walter & Heinz-Dieter Heckmann (eds.): *Physicalism and mental causation*. Exeter: Imprint Academic, 323-340.

Howell, Robert J. (2007): "Immunity to Error and Subjectivity." *Canadian Journal of Philosophy* 37 (4), 581–604.

Jeannerod, Marc & Pacherie, Elisabeth (2004): "Agency, Simulation and Self-identification." *Mind & Language* 19 (2), 113–146.

Kaplan, David (1989): "Demonstratives." In: Joseph Almog, John Perry & Howard Wettstein (eds.): *Themes From Kaplan*. Oxford: Oxford University Press, 481–563.

Lafraire, Jérémie (2013): "Two Notions of (Mis)-Identification." *Philosophical Inquiries* 1 (2), 39–53.

Lane, Timothy & Liang, Caleb (2010): "Mental ownership and higher-order thought: Response to Rosenthal." *Analysis* 70 (3), 496–501.

Lane, Timothy & Liang, Caleb (2011): "Self-Consciousness and Immunity." *Journal of Philosophy* (108), 78–99.

Langland-Hassan, Peter (2008): "Fractured Phenomenologies: Thought Insertion, Inner Speech, and the Puzzle of Extraneity." *Mind & Language* 23 (4), 369–401.

Langland-Hassan, Peter (forthcoming): "Introspective Misidentification." *Philosophical Studies*. http://philpapers.org/archive/LANIM-3.pdf, retrieved March 4 2014.

Liang, Caleb & Lane, Timothy (2009): "Higher-order thought and pathological self: the case of somatoparaphrenia." *Analysis* 69 (4), 661–668.

Marcel, Anthony (2003): "The Sense of Agency: Awareness and Ownership of Action." In: Johannes Roessler & Naomi Eilan (eds.): *Agency and Self-awareness*. Oxford: Clarendon Press, 48–93.

Martin, Jean-Remy & Pacherie, Elisabeth (2013): "Out of nowhere: Thought insertion, ownership and context-integration." *Consciousness and Cognition* 22 (1), 111–122.

Mellor, C.S (1970): "First rank symptoms of schizophrenia." *British Journal of Psychiatry* 117, 15–23.

Mizumoto, M. & Ishikawa, M. (2005): "Immunity to Error through Misidentification and the Bodily Illusion Experiment." *Journal of Consciousness Studies* 12 (7), 3–19.

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

Mullins, Simon & Spence, Sean (2003): "Re-examining Thought Insertion." *British Journal of Psychiatry* 182, 293–298.

Musholt, Kristina (2011): "Self-consciousness and nonconceptual content." *Philosophical Studies* 163 (3), 649–672.

Pacherie, Elisabeth; Green, Melissa; Bayne, Tim (2006): "Phenomenology and delusions: Who put the 'alien' in alien control?" *Consciousness and Cognition* 15 (3), 566–577.

Peacocke, Christopher (2003): "Action: Awareness, Ownership, and Knowledge." In: Johannes Roessler & Naomi Eilan (eds.): *Agency and Self-awareness*. Oxford: Clarendon Press, 94–110.

Peacocke, Christopher (2007): "Mental action and self-awareness (I)." In: Brian P. McLaughlin & Jonathan Cohen (eds.): *Contemporary debates in the philosophy of mind*. Oxford: Blackwell, 358–376.

Perry, John (1977): "Frege on Demonstratives." *Philosophical Review* 86, 474–497.

Perry, John (1979): "The Problem of the Essential Indexical." *Noûs* 13, 3-12.

Perry, John (1998): "Myself and I." In: Marcelo Stamm (ed.): *Philosophie in Synthetischer Absicht*. Stuttgart: Klett-Cotta 83–103.

Petkova, Valeria I. & Ehrsson, H. Henrik (2008): "If I Were You: Perceptual Illusion of Body Swapping." *PLoS ONE* 3 (12), e3832.

Pickard, Hanna (2010): "Schizophrenia and the Epistemology of Self-Knowledge." *European Journal of Analytic Philosophy* 6 (1), 55–74.

Prosser, Simon (2012): "Sources of immunity to error through misidentification." In: Simon Prosser & François Recanati (eds.): *Immunity to*

Error Through Misidentification: New Essays. Cambridge, MA: Cambridge University Press, 158–179.

Pryor, James (1999): "Immunity to Error through Misidentification." *Philosophical Topics* 26 (1&2), 271–303.

Recanati, François (2007): *Perspectival Thought*. Oxford: Oxford University Press.

Recanati, François (2012a): "Immunity to error through misidentification: what it is and where it comes from." In: Simon Prosser & François Recanati (eds.): *Immunity to Error Through Misidentification: New Essays*. Cambridge, MA: Cambridge University Press, 180–201.

Recanati, François (2012b): *Mental Files*. Oxford: Oxford University Press.

Romdenh-Romluc, Komarine (2013): "First-Person Awareness of Intentions and Immunity to Error through Misidentification." *International Journal of Philosophical Studies* 21 (4), 493–514.

Rosenthal, David M. (2010): "Consciousness, the self and bodily location." *Analysis* 70 (2), 270–276.

Rosenthal, David M. (2012): "Awareness and identification of self." In: J. Liu & John Perry (eds.): *Consciousness and the Self*. NY: Cambridge University Press, 22-50.

Rovane, Carol (1987): "The Epistemology of First-person Reference." *Journal of Philosophy* 84, 147–167.

Seeger, Max (2013): "Commentary on Martin & Pacherie. Out of nowhere: Thought insertion, ownership and context-integration." *Consciousness and Cognition* 22, 261–263.

Seeger, Max (forthcoming): "Authorship of thoughts in thought insertion: What is it for a thought to be one's own?" *Philosophical Psychology*.

Shoemaker, Sydney (1968): "Self-Reference and Self-Awareness." *Journal of Philosophy* 65 (19), 555–567.

Shoemaker, Sydney (1970): "Persons and Their Pasts." *American Philosophical Quarterly* 7 (4), 269–285.

Shoemaker, Sydney (1996): *The first-person perspective and other essays*. Cambridge: Cambridge University Press.

Smith, Joel (2006): "Which Immunity to Error?" *Philosophical Studies* 130 (2), 273–283.

Smith, Joel (forthcoming): "Alienation and Self-Presentation." Consciousness Online 2013. http://consciousnessonline.files.wordpress.com/2013/02/smith-co5.pdf, retrieved March 2013.

Sollberger, Michael (forthcoming): "Making Sense of an Endorsement Model of Thought-Insertion." *Mind & Language*.

Sousa, Paulo & Swiney, Lauren (2011/2013): "Thought insertion: Abnormal sense of thought agency or thought endorsement?" *Phenomenology and the Cognitive Sciences* 12 (4), 637–654.

Stephens, G. Lynn (2000): "Thought Insertion and Subjectivity." *Philosophy, Psychiatry, & Psychology* 7 (3), 203–205.

Stephens, G. Lynn & Graham, George (1994): "Self-consciousness, mental agency, and the clinical psychopathology of thought insertion." *Philosophy, Psychiatry, & Psychology* 1, 1–10.

Stephens, G. Lynn & Graham, George (2000): *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*. Cambridge, MA: MIT Press.

Sugimori, Eriko; Asai, Tomohisa; Tanno, Yoshihiko (2011): "Sense of agency over thought: External misattribution of thought in a memory task and proneness to auditory hallucination." *Consciousness and Cognition* 10, 688–695.

Swiney, Lauren & Sousa, Paulo (2013): "When our thoughts are not our own: Investigating agency misattributions using the Mind-to-Mind paradigm." *Consciousness and Cognition* 22 (2), 589–602.

Synofzik, Matthis; Vosgerau, Gottfried; Newen, Albert (2008): "Beyond the comparator model: A multifactorial two-step account of agency." *Consciousness and Cognition* 17 (1), 219–239.

Vignemont, Frédérique de (2012): "Bodily Immunity to Error." In: Simon Prosser & François Recanati (eds.): *Immunity to Error Through Misidentification: New Essays*. Cambridge: Cambridge University Press.

Vosgerau, Gottfried (2009a): *Mental Representation and Self-Consciousness: From Basic Self-Representation to Self-Related Cognition*. Paderborn: mentis.

Vosgerau, Gottfried (2009b): "Stufen des Selbstbewusstseins: Eine Analyse von Ich-Gedanken." *Grazer Philosophische Studien* 78, 101–130.

Vosgerau, Gottfried & Newen, Albert (2007): "Thoughts, Motor Actions, and the Self." *Mind & Language* 22, 22–43.

Vosgerau, Gottfried & Voss, Martin (forthcoming): "Authorship and Control over Thoughts." *Mind & Language*.

Wegner, Daniel M. & Sparrow, Betsy (2004): "Authorship Processing." In: Michael Gazzaniga (ed.): *The Cognitive Neurosciences* (3rd Edition). Cambridge, MA: MIT Press, 1201–1209.

Wittgenstein, Ludwig (1969): *The Blue and Brown Books*. 2nd ed. Oxford: Blackwell.

Wright, Crispin (1998): "Self-Knowledge: The Wittgensteinian Legacy. In: Crispin Wright; Barry Smith; Cynthia Macdonald (eds.): *Knowing Our Own Minds*. Oxford: Clarendon Press, 13–45.

Wright, Crispin (2012): "Reflections on François Recanati's 'Immunity to error through misidentification: what it is and where it comes from'."

In: Simon Prosser & François Recanati (eds.): *Immunity to Error Through Misidentification: New Essays*. Cambridge*:* Cambridge University Press, 247–280.

Young, Garry (2006): "Kant and the Phenomenon of Inserted Thoughts." *Philosophical Psychology* 19 (6), 823–837.