

SUBJECT AND PERSON
AN ESSAY ON SELF-REFERENCE AND PERSONAL IDENTITY

Roland Spitz, M.A., B.Phil.

St.John's College, Oxford

TRINITY TERM 1992

D.PHIL. THESIS

ABSTRACT

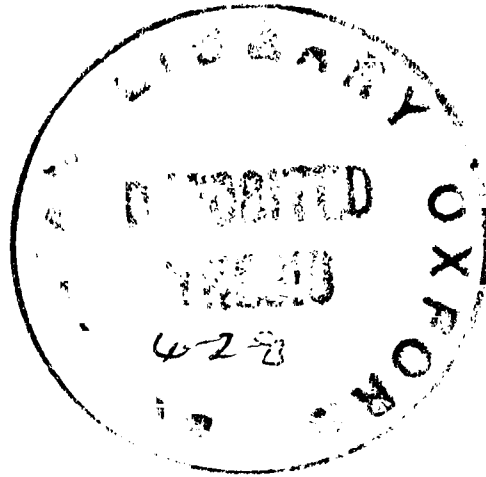
The main aim of this thesis is to investigate the relations between the topics of self-reference and personal identity and to demonstrate that, and how, certain misconceived views concerning self-reference have influenced the interpretation of various 'thought-experiments' cited in recent discussions of personal identity.

The problem posed by personal identity, I argue, is not so much an *ontological* problem concerning the special nature of a particular kind of substances (i.e. persons), but consists rather in the fact that *I* (and all other persons) are tempted to take *our* identity to be governed by a *criterion* of identity *different* from the criterion governing the identity of all other kinds of concrete particulars (namely *spatio-temporal* continuity). I argue that this temptation stems from the possibility of adopting a *first-person point of view* towards personal identity.

This possibility is shown to be intimately connected with a person's ability to refer to *himself-as-subject* (i.e. *without the application of any criteria of his identity*). This particular kind of self-reference displays certain *epistemic idiosyncracies* which make us susceptible to certain systematic illusions with regard to the nature of persons and of personal identity over time. Following Kant, I call these misconceptions '*paralogistic illusions*', and I try to show that, and how, various thought-experiments envisaging a person's putative '*body-change*' draw their persuasive strength from these illusions. Kant's distinction between a purely '*logical*' and an ontological, or *substantial*, understanding of the 'identity of the subject' is shown to be central to a dissolution of the illusion as well as to the solution of various philosophical conundrums concerning (particular kinds of) memory and imagination.

The view of persons and their identity that emerges from the discussion differs sharply from both a 'Simple View' and a 'Reductionist View' of personal identity. Persons, it will appear, are a functionally definable sub-class of *animals*, and -- given the fulfilment of certain background-conditions -- a person's numerical identity over time consists in the *spatio-temporal continuity* of his *living body*.

**SUBJECT AND PERSON
AN ESSAY
ON
SELF-REFERENCE AND PERSONAL IDENTITY**



**Roland Spitz, M.A., B.Phil.
St. John's College, Oxford
TRINITY TERM 1992
D.PHIL. THESIS AND ABSTRACT**

**SUBJECT AND PERSON
AN ESSAY
ON
SELF-REFERENCE AND PERSONAL IDENTITY**

ACKNOWLEDGEMENTS

In writing this D.Phil.-Thesis, I have had the benefit of having my ideas criticised by various people on several continents, and I am afraid the final result compares somewhat poorly with the global setting of the enterprise.

Some of the ideas propounded in this thesis go back to the time of my studies at Munich University. With regard to these I am particularly indebted to Dr. H.J. Falk, who first aroused my interest in the topic, and to Professor Dieter Henrich.

Parts of this thesis were handed in as a B.Phil.-Thesis in Trinity Term 1989. I should like to thank Professor Derek Parfit, under whose supervision and criticism it slowly developed. Although it will appear from the text that I disagree with him in all substantial respects, I learnt a lot from him, and I have fond memories of our discussions in All Souls College. I should also like to thank Dr. Snowdon, who made very helpful comments on a provisional draft of Chapter 3. I am particularly indebted to my teacher at St. John's College, Dr. Peter Hacker. Without his encouragement in the face of adversity, I probably would not have finished this thesis.

I should also like to thank Professor Barry Stroud, who during the time I spent at the University of California at Berkeley, perhaps inadvertently, turned me into something like a Wittgensteinian. His taciturn wit was a constant source of joy to me in an otherwise unintelligible cultural environment.

Special thanks go to Dr. Quassim Cassam who supervised this thesis in its later stages. His probing scepticism was a healthy counterweight to my impatience. Most of all, however, I should like to thank my dear friends Hanjo Glock and Luke Alvarez, who both read and commented on the entire script. Many a time they prevented me from making silly mistakes.

Finally I should like to acknowledge my debt to the German Academic Exchange Service for financing my studies at Oxford University, and to the Regents of the University of California for making it possible for me to spend a year at Berkeley.

When I returned to my B.Phil.-Thesis after a lapse of over two years, I had to realize that its subject-matter had suddenly become a 'hot topic' in contemporary discussions. Some of the claims I had made in my B.Phil.-Thesis had in the meantime been made in different ways by other people. I have left them, whenever it seemed necessary to preserving the logical flow of the argument.

I wish I had produced a better piece of work. But the topic of this thesis still seems to me one of the most intractable problems of philosophy, and I doubt that I could have improved this thesis substantially by spending even more time on it. I console myself with the thought that even the greatest philosophers seem to have found it very difficult to write clearly on this topic.

TABLE OF CONTENTS

1. INTRODUCTION ...	(1)
2. THE PROBLEM OF PERSONAL IDENTITY ...	(6)
2.1. <i>The Notion of a Criterion of Identity ...</i>	<i>(7)</i>
2.1.1. <i>Identification, Sortal Concepts, and Criteria of Identity...</i>	<i>(8)</i>
2.1.2. <i>'Conceptualism' versus 'Realism' ...</i>	<i>(14)</i>
2.1.3. <i>Criteria of Identity and Evidence for Identity ...</i>	<i>(21)</i>
2.1.4. <i>Numerical Identity, Qualitative Identity, and Spatio-Temporal Continuity...</i>	<i>(28)</i>
2.1.5. <i>What is a Criterion of Personal Identity...</i>	<i>(34)</i>
2.2. <i>Personal Identity and the First-Person Point of View...</i>	<i>(36)</i>
3. THE EPISTEMOLOGY OF SELF-REFERENCE....	(43)
3.1. <i>Some Characteristic Features of Self-Reference...</i>	<i>(43)</i>
3.2. <i>The Distinction between Semantic and Intentional Self-Reference...</i>	<i>(46)</i>
3.3. <i>Does 'I' Have a Sense? ...</i>	<i>(53)</i>
3.4. <i>The Matter of Identification in Self-Reference....</i>	<i>(62)</i>
4. THE ONTOLOGY OF SELF-REFERENCE....	(70)
4.1. <i>The Simplicity and Particularity of the Subject...</i>	<i>(70)</i>
4.2. <i>The 'Logical Identity' of the Subject ...</i>	<i>(81)</i>
4.3. <i>Experiential Memory and Personal Identity...</i>	<i>(93)</i>
4.4. <i>Subjective Imagination and Nagel's 'Objective Self'</i>	<i>(98)</i>
5. SELF-REFERENCE AND PERSONAL IDENTITY...	(106)
5.1. <i>Thought-Experiments...</i>	<i>(107)</i>
5.2. <i>Inductive Evidence, Criteria, and Background-Conditions...</i>	<i>(116)</i>
5.3. <i>Brain-Transfers and Personal Identity...</i>	<i>(119)</i>
5.4. <i>'Body-Changes' and Psychological Continuity...</i>	<i>(123)</i>
5.4.1. <i>Locke, Quasi-Memory, and Psychological Continuity...</i>	<i>(123)</i>
5.4.2. <i>'Branch-Line Cases' and the 'Reduplication Argument' ...</i>	<i>(129)</i>
5.4.3. <i>Psychological Continuity as a Criterion ...</i>	<i>(145)</i>
5.5. <i>'Body-Switches' and the First-Person Perspective...</i>	<i>(148)</i>
5.6. <i>The Simple View...</i>	<i>(157)</i>
5.7. <i>Parfit's Reductionism...</i>	<i>(165)</i>
5.8. <i>Persons and Their Bodies...</i>	<i>(171)</i>
5.8.1. <i>Logical Limits of Embodiment...</i>	<i>(172)</i>
5.8.2. <i>Persons, Animals, and Their Bodies ...</i>	<i>(177)</i>
6. OUTLOOK...	(187)
BIBLIOGRAPHY ...	(191)

I. INTRODUCTION

The entity I refer to by 'I' is an entity I believe to be one and the same throughout my whole life. This truism expresses the close relation which obtains between the issues of self-reference and personal identity in its most basic and uncontroversial form. The question of what, if anything, constitutes the identity of a person over time seems to be intimately related to the question of what, if anything, a person refers to by uttering (or thinking) 'I'. Thus it is perhaps no coincidence that both of these questions have experienced a nearly simultaneous revival in the discussion of the last two decades of contemporary philosophy, a development taking its point of departure from Sidney Shoemaker's pioneering work *Self-Knowledge and Self-Identity*.

At the time of writing this book, Shoemaker thought that both the problem of what a person's identity over time consists in and the problem of how a person refers to himself allow of simple and straightforward solutions. And he claimed that the problem of personal identity

'arises mainly from philosophical views about, and philosophical perplexities about, the nature of self-knowledge, and that its solution consists mainly in the resolution of these perplexities and the correction of mistakes involved in these views'.¹

This fairly optimistic attitude is not shared by many contemporary philosophers any more. Too complex have the problems of self-reference turned out to be, central and of foremost importance to many issues in epistemology and ontology, and, it seems, firmly resisting any rash 'dissolutions' or 'reductions'. At the same time, the issue of *diachronic* personal identity (which is the only aspect of the problem I shall be concerned with in what follows) has in a way emancipated itself from the issue of self-reference, which has traditionally been approached mainly from a logical or epistemological point of view. Personal identity has emerged as a subject-matter of its own right, to be treated with its own methods (i.e. thought-experiments or puzzle-cases) different from the ones customarily applied in discussions of self-consciousness. The discussion on personal identity has thus opened up a whole new field of inquiry and has -- especially in Parfit's and some of Williams' work² -- arrived at indeed novel and unexpected results.

¹ Shoemaker, *Self-Knowledge and Self-Identity*, Preface, p. VIII.

² Cf. Parfit, *Reasons and Persons*, Part Three 'Personal Identity', and his earlier essay 'Personal Identity'; and Williams, 'The Self and the Future'.

Because of this development, in which issues which in earlier times had not been kept duly separate were now clearly distinguished, the close theoretical relations between the two problem-fields seem to me to have been somewhat neglected; and that despite the fact that theories of personal identity as well as theories of self-reference often enough surreptitiously draw on, and implicitly make use of, these relations. A further consequence of this development, it seems to me, has been that the question of what constitutes a person's numerical identity over time has come to be viewed as a basically *ontological* problem, i.e. a problem arising from the *special nature* of a particular kind of beings (namely persons).

The present essay, by bringing the conceptual relations between the issue of personal identity and the issue of self-consciousness out into the open, attempts to reverse this process and thus, in a way, revive Shoemaker's original insight. That is to say, it will try to show that Shoemaker was right in suggesting that the problems posed by personal identity are not so much problems arising from the *special* nature of a particular *kind* of beings (i.e. persons), but rather problems arising from certain systematic *misconceptions* about what constitutes their numerical identity over time that members of this kind are typically prone to in philosophical reflection. Shoemaker pointed out that the questions 'What is a person?' and 'What does a person's numerical identity over time consist in?' are often understood as the questions 'What am I?' and 'What does *my* identity over time consist in?'.³ I shall argue that it is this possibility of adopting a *first-person point of view* to matters of personal identity that makes personal identity a more difficult, and more interesting, topic of philosophical enquiry than the identity of 'merely' material objects. What is at issue in discussions of personal identity is (potentially) *my* identity. And it is the simple fact that I can always take a first-person stance to imaginary cases of personal identity which, I shall argue, makes us prone to certain illusions concerning what our (i.e. 'personal') identity consists in -- illusions which in turn, I shall try to show, stem mainly from a misconceived understanding of (a particular kind of) *self-reference*. It will be one of the main tasks of this essay to show that, and *how*, these misconceptions have influenced and, at times, distorted the interpretation of various thought-experiments concerning a person's numerical identity over time.

The investigation will start out (*Chapter 2*) by trying to get a firm grip on what

³Cf. Shoemaker, *Self-knowledge and Self-Identity*, p.9

the supposed ‘problem of personal identity over time’ really amounts to. Often the problem is stated by asking what a person’s numerical identity over time ‘consists in’. This, however, I shall argue, is a thoroughly obscure formulation of the problem. The characterization of the problem as the search for a *criterion* of identity for persons, though *prima facie* not much more perspicuous, will prove to be more fruitful, and the bulk of Chapter 2 will be dedicated to developing a clear notion of a criterion of personal identity.

Persons are a kind of *particulars* or *substances* (where by ‘person’ I shall provisionally understand ‘whatever kind of particular in the objective order of things *I* am -- be it a *res cogitans*, or a Martian, or a human being’). However, the problem of personal identity, it will emerge, does not consist in the special nature of this particular kind of particulars, but rather in the fact that I (and you, and all other creatures which are in the relevant respect similar to me) are tempted to take *our* identity to be governed by a *criterion* of identity *different* from the criterion governing the identity of all other kinds of particulars (namely *spatio-temporal* continuity). Persons, we are tempted to think, are not essentially individuated through their spatial location, and their numerical identity over time does not require any kind of spatial continuity. This temptation, I shall argue, stems from the already mentioned possibility of adopting a *first-person* point of view towards personal identity (it is *one’s own* identity over time that is in question), -- a possibility which in turn gives rise to the idea of a *special access* that everybody has to his own identity. I shall urge that this idea of a special access is based on an illusion, and that the solution to, or dissolution of, the problem of personal identity consists not so much in an elaborate ‘theory’ which provides a necessary and sufficient criterion of personal identity, but in our gaining insight into the origin and hidden workings of this illusion.

The idea of a person’s special access to his own identity will be shown to be intimately connected with a certain capacity which is distinctive of persons and may well be said to be a necessary requirement of personhood, namely the capacity to refer to oneself *without the application of any criteria of one’s identity whatsoever*. Drawing on Wittgenstein’s distinction between a use of ‘I’ as object and a use of ‘I’ as subject, I shall call this capacity the capacity to refer to *oneself-as-subject*. Chapter 3 will be dedicated to an investigation of the various epistemological idiosyncracies displayed by this particular kind of self-reference. I shall argue that these idiosyncracies are liable to give rise to various kinds of misconceptions regarding the nature of this kind of self-

reference, which in turn make us susceptible to certain systematic illusions with regard to the nature of persons and of personal identity over time. The view that a person's reference to himself-as-subject is necessarily a form of self-*identification* will emerge as the main source of these illusions.

Following Kant, I shall call this particular kind of misconception a '*paralogism*' or a '*paralogistic illusion*', and in *Chapter 4* I shall try to probe more deeply into its nature and origins. I start out by examining Kant's own struggling attempts to explain the true nature and delusory character of a person's self-reference in the 'Paralogisms'-chapter of the *Critique*. However, whereas Kant's main aim in those passages was to rebut certain unjustified ('transcendent') claims to knowledge of his 'rational psychologist' predecessors, I intend to show that Kant's reflections contain insights which can also be critically applied to the topic of 'empirical' personal identity. In particular, I shall suggest that they can help us to evaluate the theoretical validity of the '*first-person perspective*' so often invoked in thought-experiments concerning personal identity.

What lies at the heart of the paralogistic illusion is, I shall claim, the hypostazation of a '*special referent*' of 'I' when used as-subject (whether it be called a self, ego, or subject. This hypostazation of a special referent -- although it may well be, as Kant alleged, a 'necessary and inevitable illusion'⁴ -- is misconceived. The 'pure point of view of consciousness', it will be shown, cannot by itself determine any criteria of identity of the particular referred to by 'I'. *What* I am (i.e. what *kind* of particular), and hence what constitutes my identity over time, cannot be determined purely from the first-person point of view. Or, to put it differently, there are no criteria of identity for Subjects or Selves or Egos as such. Since a person's reference to himself-as-subject does not draw on any criteria of his identity, it cannot be taken to establish any such criteria. The first-person perspective does not by itself *identify* a given particular -- although, I shall suggest, it *presupposes the objective identifiability* of this particular as a member of a given kind and may well be taken to establish its *existence*. Its criterion of identity, however, has to be 'filled in' from the objective, or third-person, perspective.

Kant's distinction between a purely '*logical*' and an ontological, or *substantial*, sense of the identity of the subject will be shown to be of central importance in this context. I shall also try to demonstrate that it can help us in coming to grips with various philosophical conundrums concerning (particular kinds of) memory and imagination.

⁴Kant, *Critique of Pure Reason*, B354.

Chapter 5 will apply the insights gained in chapters 3 and 4 to various thought-experiments which are aimed at proving that the identity of a person consists in something other than the persistence of his living body, such that that one numerically identical person could in principle ‘inhabit’ several different bodies throughout his life. It will become apparent that without exception these (Lockean or neo-Lockean) thought-experiments envisaging putative ‘*body-changes*’ are based on, and draw their persuasive strength from, certain of the paralogistic illusions discussed before. Although, it will appear, there are imaginary cases in which the question ‘Is person P_1 at t_1 identical with person P_2 at t_2 ?’ has become *conceptually undecidable*, there is no sense in which one *numerically identical* person (the referent of a person’s ‘I’-thoughts) can be said to have changed bodies. Persons are particulars, and what gives substance to a person’s particularity is the *spatio-temporal* continuity of a person’s *body*. Thus the identity of persons will turn out to be governed by exactly the same criterion as the identity of all other kinds of concrete particulars, namely spatio-temporal continuity.

From this I shall conclude that the concept of identity as applied to persons is not -- as has sometimes been thought -- more precise or more ‘indestructible’ than the concept of identity as applied to other worldly particulars (although, of course, the way in which it is applied is distinctively different). The capacity to refer to oneself-as-subject does not make persons (or their identity) *substantially* different from all other kinds of particulars. It does not endow them with a necessarily simple ‘ontological core’ (as has been thought by adherents of a so-called ‘*Simple View*’ of personal identity). Nor, however, are the identities of persons any more ‘fictitious’ than the identities of other things (as has been suggested by adherents of a ‘*Reductionist View*’ of personal identity). Rather, persons are a functionally definable kind of *animals*; they are animals with certain, vaguely specifiable, *capacities*. What, besides other things, sets them apart from the rest of the animal kingdom is not that they, as opposed to animals, could change bodies. What sets them apart is only their susceptibility to certain paralogistic illusions regarding what constitutes their numerical identity over time. And what makes them susceptible to these illusions is just the same as (part of) what makes them persons, namely their capacity to refer to themselves-as-subject, i.e. without application of any criteria of their identity.

2. THE PROBLEM OF PERSONAL IDENTITY

One of the main difficulties in dealing with the problem of personal identity is getting a clear conception of what exactly the problem is. Often it is put as the question of what a person's numerical identity over time 'consists in'. This, however, is a thoroughly obscure formulation of the question, which is evident from the fact that it is understood in different ways by different theorists of personal identity: Some of them (such as Chisholm, Swinburne, Shoemaker, Williams, Wiggins, with certain qualifications, and Parfit, at times) view it as a question to be answered in terms of a *conceptual* analysis of our concept of personal identity (or, put differently, in terms of an analysis of the *meaning* of statements such as 'P₁ at t₁ is numerically identical with P₂ at t₂'). Others (such as Mackie, Perry, and Nagel) consider it to be a *quasi-empirical* question to be answered, at least in part, in terms of empirical findings concerning what 'underlies' our, i.e. human beings', mental lives.

Thus *Mackie* has argued that the question what personal identity consists in can be answered by specifying the 'real essence of personal identity', by which he means 'whatever underlies and makes possible the unity of consciousness'¹. A person's numerical identity through time can then be said to consist in whatever is required for the persistence of this underlying real essence. What this real essence will finally turn out to be is, according to Mackie, an empirical question, to be answered by medical findings. According to the latest pronouncements of the scientific community, the best candidate is of course the human central nervous system, or even more specifically, the human brain. Similarly, *Nagel* has propounded the thesis that 'I am whatever persisting individual in the objective order underlies the subjective continuities of that mental life that I call mine'. This thesis, Nagel says, is to be understood not 'as an analysis of the concept of the self but as an empirical hypothesis about its true nature.'². And, finally, *Perry* has proposed to view personal identity as a relation holding between successive stages of *human* beings which explains certain very general truths about them (e.g. regularities of character, continuity of memories); and he proposes *having the same brain* as the most natural candidate for this relation³.

¹Mackie, *Problems from Locke*, p. 200.

² Nagel, *The View from Nowhere*, p.40.

³Cf. Perry, 'The Importance of Being Identical', esp. pp. 69-72.

What underlies these empirical (or quasi-empirical) approaches are Quinean doubts about the possibility of drawing a clear distinction between conceptual and factual analysis, and/or -- at least in Nagel's case -- an adherence to a version of the Kripke-Putnam account of reference and to the Kripkean notion of a posteriori, or metaphysical, necessity. I share neither the Quinean doubts nor the Kripke-Putnam views; nor am I so certain that notion of a posteriori, metaphysical, necessity is intelligible. However, I shall not try to justify these views in the present context. I shall try to show below that a clear account of what a conceptual analysis of problems of personal identity amounts to can be given. And I also think that the specifically philosophical problems which personal identity poses do not arise from ignorance of empirical facts about ourselves, but from conceptual confusions and a special kind of 'paralogistic' illusion. Therefore, in what follows, I shall understand the question of what personal identity 'consists in' as asking for a conceptual analysis; and I shall take questions concerning an entity's 'true nature' or 'real essence' as conceptual, not empirical questions. Nagel's and Mackie's 'empirical hypothesis' about the true nature of the self is something scientists have found out *about* personal identity (namely that it is a person's brain, rather than his heart or his left thumb, that 'causally underlies' certain mental continuities). But it is not -- in the sense I shall understand it -- 'constitutive' of personal identity.

2.1. The Notion of a Criterion of Numerical Identity

Another way of understanding the question what personal identity consists in is to take it as a request for the *criterion (or criteria) of personal identity*. This manner of exposition, however, is only marginally better than the previous one, since the notion of a criterion is one of the most iridescent and controversial notions of contemporary philosophy. Its introduction into the recent discussion on personal identity has caused many misunderstandings, and has at times been a serious obstacle to any progress in coming to understand what is essential about the issue. However, since I think that the notion of a criterion is, if used properly, a powerful philosophical tool, and since I shall therefore want to make use of this notion, I shall try in the following to explain and, where necessary, criticize the use that various theorists of personal identity have made of it. And I shall then proceed to explain and justify my own use of this notion.

2.1.1. Identification, Sortal Concepts, and Criteria of Identity

In its modern use the notion of a criterion of identity goes back to Frege, who introduced the notion (although not the term) in attempting to determine the identities of *abstract* objects such as numbers, sets, and directions⁴. Other philosophers, such as Davidson, have discussed the criteria of identity of events⁵. In what follows, however, I shall be concerned entirely with the criteria of identities of *concrete* (and that is, perceptible and persistent) individuals such as animals, plants, and artefacts. The term ‘criterion of identity’ in its application to *concrete* individuals (which I shall call *particulars*) is ambiguous in at least two different respects. Firstly, there are different kinds of identity-question that a criterion of identity may be used to answer. Secondly, there are diverging views as to the logical status (‘evidential’ or ‘constitutive’) of so-called criteria of identity.

Let us start with the first ambiguity: A criterion of identity may be used to give grounds for identifying a particular as of a certain kind ϕ ⁶. It is thus used to answer the ‘What is it?’ or *identification-question*. Then, a criterion of identity may also be used to give grounds for identifying a certain particular ϕ as the same ϕ as a certain ϕ given at the same time (synchronic identity). If used in this way, a criterion of identity gives an answer to a ‘Which one?’ or *counting-question*. Finally, a criterion of identity may be invoked to give grounds for identifying a given particular of a given kind ϕ as the same ϕ as a certain ϕ given at an earlier or later time (diachronic identity). In this case, a criterion of identity answers a ‘Is it one and the same ϕ ?’ or *reidentification-question*.

A criterion of identity invoked in answering an *identification-question* may be understood as a criterion of *application* of a given concept. Thus, to give a very simple example, one might say that our criteria for saying that somebody is a boy are: that he is a human being, male, and biologically immature. We determine whether a given concept applies to a given entity by checking whether it fulfills the necessary and sufficient conditions (the Fregean ‘Merkmale’ -- if there are such) for the application of that concept. Understood in this way, citing the criteria for the application of a concept is

⁴Cf. Frege, *The Foundations of Arithmetics*, §§62ff..

⁵Cf. Davidson, ‘The Individuation of Events’.

⁶For the following distinctions, cf. R. Swinburne, ‘Persons and Personal Identity’, pp. 221-223, and Wiggins, *Identity and Spatio-Temporal Continuity*, pp.35f..

nothing more, and nothing less, than giving a good, old-fashioned analysis of the concept in question. Criteria may be invoked in explaining the meaning of a given concept-word (in terms of the necessary and sufficient conditions of its application), and at the same time they may be invoked in justifying one's claim that a given entity falls under the given concept by citing the (criterial) grounds one has for that claim.

An analysis along these lines, although it might be formally correct, could be very misleading as an analysis of what is involved in answering a counting or reidentification question. For -- as was probably first noted by Locke⁷ and in recent years developed by Wiggins -- the criteria of distinctness and reidentification invoked in answering such question differ depending on the *kinds* of particulars in question⁸. What constitutes the synchronic and diachronic identity is not the same for all particulars of all kinds, but is typically determined by *what kind* of thing a given particular is. Therefore, it would seem, no general answer can be given to the question how we judge the truth or falsehood of an identity-statement with respect to *all* kinds of particulars. Rather, questions of synchronic and diachronic identity of a given particular *x* are to be settled by determining what kind of thing the particular *x* is (i.e. by answering the identification-question), since the concepts which may be invoked in answering an identification-question are at the same time the ones that provide, or specify, criteria of (synchronic and diachronic) identity for the entities that fall under them.

The distinction between genuinely *sortal* (or *substance kind*) concepts and merely *characterizing* concepts was first clearly drawn by Strawson in *Individuals*⁹. There Strawson pointed out that, although both sortal and characterizing universals

⁷Cf. Locke, *Essay Concerning Human Understanding*, II, ch. xxvii, §9.

⁸This has been stated as the thesis that identity is relative, or sortal-relative (cf. Geach, *Reference and Generality*, pp.174ff., and esp. p.216). Wiggins, in *Sameness and Substance*, pp. 15-44 repudiates the thesis, whilst trying to preserve the basic insight underlying it. I will bypass the question whether identity is a 'relative' (Geach) or 'absolute' (Wiggins) notion, since nothing much seems to hang on it as far as personal identity is concerned.

⁹ Cf. *Individuals*, pp. 168/9, and 202-213. More precisely, Strawson claimed that although both 'sortal and characterizing universals' may be said to supply principles of grouping, and even of counting, characterizing universals, if at all, 'supply such principles only for particulars already distinguished, or distinguishable, in accordance with some antecedent principle or method' (p. 168). Cf. also Geach's very similar distinction between 'substantival' and 'adjectival' general terms, in *Reference and Generality*, pp.63f..

provide principles of grouping, only sortal universals provide principles for *distinguishing* the particulars they collect. That is to say, only sortal concepts provide criteria for (synchronic) distinctness and (diachronic) reidentification for the entities that fall under them. (Hence, e.g., ‘table’ is a sortal term, but ‘brown’ is not). This distinction between sortal and characterizing concepts was subsequently developed by David Wiggins in his book *Identity and Spatio-Temporal Continuity*, and then substantially revised and extended in his *Sameness and Substance*. Wiggins’ theory of identity and individuation (which may be understood as an attempt to clarify, and where possible, justify the ancient doctrine of substance and accident) is by far the most sophisticated in the literature on the topic. Although it would lead too far in the present context to scrutinize the terminological niceties of Wiggins’ theory and to examine the manifold theoretical motivations that underlie them, a brief look at some very general results of Wiggins’ theory of identity can help us in clarifying our own conception of the notion of a criterion of identity.

Wiggins starts out his theory of identity by further sharpening Strawson’s notion of a sortal concept. He does that by stipulating various logical requirements that a concept has to fulfill in order to count as a genuinely sortal concept, i.e. to count as a concept that can ‘cover’ an identity-statement of the form ‘x is the same ϕ as y’. Put in a nutshell, the final outcome of this discussion is that the notions of a sortal concept and of substance-identity are mutually interdependent. As Wiggins himself puts it, ‘...a genuinely sortal predicate must stand for a concept that implicitly or explicitly determines identity, persistence and existence conditions for members of its extension.’¹⁰

This, Wiggins argues, explains what renders identity-questions concerning continuants as definite and determinate as they generally seem to be. For

‘...what an entity is determines what shall count as the continuance and persistence of this entity, and [...] what determines the continuance and persistence of this entity determines what it is to *be* that entity.’¹¹

Wiggins goes even one step further when he concludes that

‘an account of what x sortally is and the specification of the principle of individ

¹⁰Wiggins, *Sameness and Substance*, p. 62.

¹¹Op.cit., p. 60.

uation of x are two aspects of one and the same thing;¹².

This conclusion allows Wiggins to link up his theory with the traditional understanding of the Aristotelian ‘What is it?’ or identification-question as asking for the *essential nature*, or kind-membership, of a given substance, and the familiar search for the *principium individuationis* for substances of a certain kind (which finds expression in what we called the counting and reidentification question). The logical dependency between the answers given to the three questions is mutual. What kind of thing an entity is typically determines its principle of individuation, or ‘identity [and] persistence condition’¹³ (which is to say both its criterion of numerical identity at a given time and at different times). And, vice versa, the principles that govern the answers to the counting and reidentification questions with respect to entities of a given kind ϕ determine (at least partly) what kind of entities ϕ s are. Thus Wiggins can claim that an answer to the Aristotelian identification-question, which specifies a given particular as of a certain kind, can be understood as ‘an account of what constitutes identity’¹⁴ for particulars of that kind. Wiggins also pays tribute to the Aristotelian tradition by further distinguishing *substance concepts* as a subclass of sortal concepts. If a sortal concept f is such that if ‘ x is no longer f ’ entails ‘ x is no longer’, then f is a substance concept¹⁵. As opposed to other sortal concepts which may only specify a ‘nominal essence’ (such as the concept of a *tailor*), a substance concept specifies a particular’s ‘real essence’. An entity’s ‘real essence’ in turn determines what all the members of the relevant substance kind have to have in common in order to be members of the given kind, and what any individual member of the kind cannot do without.

¹²Op.cit., p.60.

¹³Cf. op.cit., pp. 60, 70, 141, 151.

¹⁴Op.cit., p.53; cf. the discussion below, sect.2.1.3..

¹⁵Op.cit., p.64. It has to be added that this is only one aspect of Wiggins’ definition of what he means by a substance concept. Later he adds the further requirement that, in order to be a substance concept, a concept has to determine ‘either a principle of *activity*, a principle of *functioning* or a principle of *operation*’ for members of its extension.’ (p. 70). I shall here have to ignore this (admittedly crucial) aspect of Wiggins’ theory. I shall return to it briefly below cf. sect. 5.4..

One major problem with this account seems to me that most of what we naturally consider sortal concepts do not really fulfill the strict requirements imposed on them by Wiggins. If we take any of our concepts of natural kinds (which surely must count as the prime candidates for sortal concepts) we will have to recognize that we can hardly specify any of the clear-cut principles of individuation invoked by Wiggins. (And it may be no accident that nowhere in his book does Wiggins actually ever spell out any of his principles of individuation, or ‘persistence-conditions’.) All that even our best candidates for sortal concepts seem to be able to provide are, as Strawson put it in his article ‘Entity and Identity’ (although not with explicit reference to Wiggins),

‘a whole cluster of readiesses for variation *between* individual members of the kind and of expectations of typical continuities and typical modifications *in* individual members of the kind’¹⁶.

These expectations are, as it were, built into our sortal concepts. But they are by far not as strict and clear-cut as Wiggins makes them out to be. Strawson goes on to claim that this vagueness is inherent. He says:

‘It is impossible to specify fully and precisely these expectations and readiesses. If we made the attempt, we should find ourselves launching into a disquisition on typical histories and varieties of organism, non-organic natural object and artefact. A consequence is that when we attempt a reasonably brief, yet full and precise, statement of a criterion of identity for some substance kind, we either overshoot the mark or fall short of it’¹⁷.

This sounds like a fair analysis of the situation we actually find ourselves in. However, as it stands, it cannot count as (and perhaps was not meant to be) a decisive argument against Wiggins’ (arguably too stringent) account of the criteria of identity specified by our sortal concepts. For Wiggins could reply to this objection that this is just what was to be expected when analysing our *pre*-scientific notions of natural kinds. According to Wiggins’ terminology (which he borrows from Frege), our pre-scientific notions of natural kinds express only the ‘conceptions’ we presently have of the respective ‘concepts’ which specify the real essences of natural kinds. Only the ‘concepts’, not the ‘conceptions’, give determinate persistence-conditions for all members of a given natural kind. To repeat the example given by Wiggins’ himself to illustrate this claim: whereas the *concept* of a horse is *what it is to be a horse*, and thus can be said to

¹⁶Strawson, ‘Entity and Identity’, pp. 210f..

¹⁷Strawson, ‘Entity and Identity’, p. 211.

specify a horse's real essence, our *conception* of the concept *horse* is the, very primitive, *theory* we have about what it is to be a horse¹⁸. Thus Wiggins would reply to the (Strawsonian) vagueness-objection that our present natural kind concepts cannot yet specify clear-cut persistence conditions for the entities that fall under them because they are not yet sharp enough clearly to demarcate, or single out, the 'real essences' of these entities. Our criteria of identity are not yet, as Wiggins puts it at one point, 'filled out in the direction of adequacy'¹⁹. This 'filling out', he seems to imply in several passages, is the job of science, not of philosophy.

Wiggins calls this position 'realist conceptualism', or 'conceptual realism'. The interest of this position to our topic of personal identity is that one of its most significant consequences is what Wiggins calls his 'animal attribute view'²⁰ of personal identity. According to this view the concept of a person is a *functional* specification of the natural kind concept animal in terms of an (in principle open-ended set) of capacities essential to personhood. The claim that the concept of a person is a functional specification entails that the concept of a person is not a genuine sortal concept, since it cannot be said to specify a principle of individuation for persons as such. Rather, a given person's criterion of identity is specified by the respective real essences characteristic of the natural kind it (essentially) belongs to. Thus there would be different criteria of identity for dolphin-persons, parrot-persons, etc..²¹. I shall discuss Wiggins' animal attribute view below (cf. sect. 5.8.2.). For the moment, I should like to focus on a more general problem inherent in Wiggins' 'conceptual realism'.

The topics thrown up by Wiggins' 'conceptual realism' are far-reaching. In particular, they touch upon the question if (and if so, in what sense) natural kind concepts -- as opposed to other sortal concepts -- are 'nomologically founded'²². I shall

¹⁸Op.cit., p.79, note 2. Wiggins' use of the notion 'real essence' must not be confused with Mackie's or Nagel's use of the expression (cf. above, pp. 10 ff.). For Wiggins a real essence is *conceptually* determined (cf. op.cit., p.87). Cf. the discussion below, sect. 2.1.2..

¹⁹Op.cit., p. 70.

²⁰Cf. op.cit., p.171 ff..

²¹Cf. op.cit., p. 172.

²²Cf. Wiggins, *Sameness and Substance*, p.77-86; cf. also Lowe, *Kinds of Being*, pp.138-164.

not attempt to discuss the matter in any depth. However, in what follows I *shall* try to throw some doubts on a claim which seems to be implicit in at least some of the things Wiggins says. This is the ('realist') claim that the criteria of identity specified by a given sortal concept (for whatever falls under it) are fixed independently of, or transcend, what we *mean* when we apply a given sortal term. The position I am inclined to adopt is more 'conceptualist' than the view endorsed by Wiggins. It requires that the criteria of identity specified by a given sortal concept (if only in the vague Strawsonian sense) are *conceptually* determined in a more narrow sense than the one implied by Wiggins. That is to say, if -- as Wiggins rightly claims -- these criteria are (partly) to determine our sortal concepts, they have to play *some* role in our application of these concepts. And this implies that they cannot be determined by 'real essences' which may in our linguistic practice be inaccessible to us.

2.1.2. 'Conceptualism' versus 'Realism'

I think we can agree with Wiggins that not all of the general terms that are true of an object are logically on a par. That is to say, we may agree that a select few of the terms which are true of an object play a quite special role in determining a particular's identity, and that the characteristics specified by these terms are, in a sense to be explained, *essential* to the particular's being what it is. And in this we may agree with Wiggins independently of our point of view as to whether the fact that there *are* such concepts is simply to be accepted as a brute fact characterizing our present conceptual scheme; or whether, alternatively, it has to be explained in terms of some absolute conception of reality, which is conceived as independent of any conceptual scheme, but nevertheless can be assumed to display the fundamental substance-accident structure (as has recently been urged by Michael Ayers²³). However, if we really take seriously Wiggins' thesis that, with respect to sortal concepts in general, and substance concepts in particular, an account of what is specified by the concept (the 'real essence') is just a specification of the criteria of identity of the entities falling under it (and thence grasping the concept is simply grasping the principles of individuation or conditions of persistence of the entities falling under it), then the notion of a real (as opposed to a merely nominal) essence seems to lose at least some of its explanatory value. This can be brought out in the following way.

²³Cf. Ayers, *Locke*, vol. II, esp. pp. 1-15, and 110-131.

We saw above that Wiggins defines a substance concept as a concept that is such that, if it is true of an object at all, it is necessarily true of an object throughout *its* entire existence. This is certainly an elucidating way of drawing a distinction between different kinds of sortal concepts. It shows why, e.g., 'horse' is a substance concept, whereas (e.g.) 'tailor' is not. For, if a given entity is a horse, it is necessarily a horse throughout *its* entire existence, but if an entity is a tailor, it need not necessarily remain a tailor. In other words, a particular cannot cease to be a horse without ceasing to exist *at all*. But it can cease to be a tailor whilst still remaining in existence. However, what Wiggins' definition of a substance-concept *cannot* show is that the set of properties specified by a particular's real essence (i.e. the properties that a particular has *necessarily* and without which it would not be what it is) could somehow transcend *what we mean* when we apply a given substance concept. For, whether a given concept is, if true of it at all, necessarily true of an object throughout *its* entire existence, cannot be established independently of *what we would count* as its continued existence, and, equally, *what we would count* as its perishing. Whether a given concept is, if true at all, necessarily true of an object throughout its entire existence, will finally depend on whether, if the concept did *not* apply any more, we would *count* the object as having *gone out of existence*, or rather having undergone a (possibly very fundamental) *change*. And what we would say in such a case (i.e. whether we would count the object as having gone out of existence, or, alternatively, whether we would count it as having just changed, but survived the change), is finally a matter of *what criteria of identity* we generally apply to entities of the given kind. In other words, the circumstances that would count as a substance's perishing, and the circumstances that would count as its continued existence, must be something already built into a given sortal concept -- in the sense that it must be possible, at least in principle (if perhaps not in practice), to cite these circumstances in determining the truth or falsehood of a judgement of the form ' x_1 is the same ϕ as x_2 '. That is to say, it must be possible, at least in principle (if perhaps not in practice), to *apply* the criteria of identity specified by a given substance concept ϕ in order to determine the truth or falsehood of a judgement of the form ' x_1 is the same ϕ as x_2 '. Hence the notion of a real essence cannot be said to specify a principle of *individuation* over and above, and independently of, the criteria for *distinctness and reidentification* which are built into the given substance concept, and which constitute what it is to be a member of the kind specified by the concept. If so, however, it would seem that Wiggins' notion of a real essence specifying criteria of identity for members

of a given natural kind, cannot fulfil the function it is supposed to fulfil. It cannot *guide* us in determining the truth or falsehood of an identity-judgement. If -- *per impossibile* - it did specify a principle of individuation, it would not play any role in our identity-discourse about substances of the kind specified by the concept ϕ ²⁴.

[Perhaps this criticism can gain in clarity by being put into a wider perspective: I think that part of what makes Wiggins' account so difficult to understand, and what may have induced him to overestimate the explanatory power of his notion of a 'real essence', is his adoption of the obscure (Fregean) terminological instrumentarium distinguishing concept-terms, conceptions, and concepts. Of course there is nothing wrong with saying that a concept-term refers to, or stands for a concept. Nor is there anything wrong with saying that a concept, by specifying 'what it is' to fall under it, determines what falls under it, and what not. These claims are just not very elucidatory by themselves. The crucial question here is what this ability of concepts to determine what falls under them, and what not (i.e. their ability to determine reality 'save for a Yes or No'²⁵) is grounded in. Wiggins sometimes seems to come close to the suggestion that concepts are some kind of 'Platonic objects', which can determine what falls under them completely independently of what people who grasp them would *count as* falling under them. It is of course this conception of a 'self-interpreting rule' that was attacked, and (with, I think, good reasons) discarded in Wittgenstein's rule-following considerations in the *Philosophical Investigations*:

Just as a rule determines what counts as its correct (and incorrect) application - but what rule a thinker is following is determined by what he counts as its correct (and incorrect) application, a concept determines what falls under it and what not -- but what concept a thinker has is determined by what he would count as an entity's falling under it and what not. Just as a rule sets its own standard of correctness, a concept sets its own standard of instantiation. (In a sense, rules or concepts are nothing more, and nothing less than these standards). But although these standards have 'the hardness of the logical must'²⁶, they are are not independent of human practices. In applying a rule, or in grasping a concept, we are not in touch with eternal objects in a Fregean Third Realm (a relationship which, notoriously, would give rise to all kinds of sceptical problems). A rule is nothing *independent* of what human beings count as its correct (and incorrect) application. A concept is nothing independent of what human beings count as its instant-

²⁴There are several other considerations that Wiggins adduces in favour of his 'modest essentialism' (cf. *Sameness and Substance*, p.104) which -- because of their complexity -- I cannot try to do justice here. These considerations concern mainly issues of modal logic, and in particular the idea of *de re*-necessities (cf. op.cit., pp. 104-117, and pp. 131-140).

²⁵Cf. Wittgenstein, *Philosophical Grammar*, p.213

²⁶ Wittgenstein, *Philosophical Investigations*, §437.

iation (and non-instantiation). And human beings manifest what they would count as a correct, and incorrect application of a concept-term by *explaining* its meaning, i.e. by explaining what they mean by it. Thus in order for a language-user to mean something by using a certain concept-term he has to be able to explain what he means. What he means cannot be something concealed from himself. After all, Wittgenstein says, 'nothing is hidden'²⁷ here. Of course these explanations of meaning will come to an end at some stage. They will 'reach bedrock'. But this bedrock cannot be deeper than the thinker can dig. In grasping a concept a thinker sets the standard of what is to count as its instantiation, and non-instantiation. The standard is his. It cannot be opaque to him²⁸.

Wiggins' distinction between the concept *horse* (which specifies a horse's real essence) and the conception we have of this concept, it seems to me, is in danger of opening up a gap between a concept and what is to count as its correct application, and thus invites a criticism of the cited Wittgensteinian kind.]

The same point can perhaps be made in a slightly different way. As I understand Wiggins' talk of the 'scientifically palpable real essence[s]'²⁹ allegedly specified by natural kind concepts, these essences are supposed to function as ideal standards that scientific knowledge has to aspire to. I mentioned above that Wiggins distinguishes between the (Fregean) reference of the concept-term 'horse', namely the concept *horse* (i.e. *what it is to be a horse*) and the *conception* we have of that concept (i.e. a theory of what a horse is, or 'what it is to be a horse'³⁰). At present, Wiggins seems to imply, we (or at least most of us) only have a conception of a horse. We do not yet know what it is to be a horse; we have not yet grasped the real essence specified by the concept-term 'horse'. And thence we cannot yet specify any clear-cut 'persistence-conditions' for horses. We have to wait for science to fill in our concept 'in the direction of adequacy' (cf. above). This strikes me as a curious thing to say. What could Wiggins' talk of 'adequacy' mean in this context? What, to put the question differently, would be our criterion of adequacy? How could we ever tell that our conception of a concept matches the real essence specified by the concept?

²⁷Wittgenstein, *Philosophical Investigations*, §435.

²⁸Cf. Pears, *The False Prison* vol. 2, ch.17, pp. 460-501, and Baker/Hacker, *Wittgenstein. Rules, Grammar, and Necessity*, pp.81-107.

²⁹Wiggins, *Sameness and Substance*, p. 87.

³⁰Cf. Wiggins, *Sameness and Substance*, p.79, note 2; a more elaborate exposition of this distinction is to be found in Wiggins, 'The Person as Object of Science, as Subject of Experience, and as Locus of Value', pp.56-63.

Wiggins seems to think that it is the *explanatory power* of our respective conceptions of concepts that could provide such a criterion of adequacy. And he backs up this thesis by a combination of Kripke's views on the relation between essence and identity and Putnam's account of natural kinds, according to which the extension of natural kind terms (and correspondingly the real essences of the members of a natural kind) is *fixed* by ostensive definition in a name-like fashion³¹. According to Putnam's account the determination of natural kind terms presupposes the existence of nomological principles, known or unknown, that will collect together all the members of a kind around an arbitrary good representative of the kind (what he calls a 'stereotype'). That is to say, according to Putnam, a particular *x* is a member of a given natural kind if, and only if, given one or more exemplary members of the kind, the most explanatory and comprehensive true theoretical description of the kind exemplified by these paradigms would group *x* alongside these paradigms.

It would lead too far in the present context to discuss this conception in detail. We may agree with Putnam in that we acquire natural kind concepts by some sort of ostensive definition ('Any animal similar to this is a fish'). But this analysis by itself does not commit us to the view that these ostensive definitions somehow *fix the reference* of the natural kind concept (and thus the 'real essences' of the members of the natural kinds) once and for all. We initially define the concept of a fish as 'anything similar to this' -- although it is worth pointing out that we do not use any particular fish as a *sample* for deciding whether anything counts as a fish (as e.g. with colour-terms). When we learn more *about* fish, we may start to distinguish between different kinds of fish -- for purposes of scientific classification, or for other, more mundane, purposes. In doing so we may find out that whales are very different in their internal constitution from other fish. And then we sharpen our concept of a fish to exclude whales. This conceptual innovation is of course based on scientific knowledge acquired through empirical investigations. But it does not show that our first concept was somehow *wrong*. Nor does it show that anybody who uses a natural kind term is thereby committed to the view that there is some underlying true scientific theory that unites all the members of the kind. We start out by discriminating and grouping as we find it useful

³¹Cf. Kripke, *Naming and Necessity*, Lecture III, and Putnam, 'Is Semantics Possible?'.

to do so, and as our interests require³², without having any clear-cut criteria or persistence conditions for members of the kinds distinguished. Of course, with the development of natural science we acquire the capacity to develop a systematic taxonomy of orders, families, genera and species, and the corresponding principles for deciding whether to count a given plant or animal as belonging to a given species. And with the development of such a systematic taxonomy we at the same time develop more clear-cut criteria of synchronic and diachronic identity for the members of the various species. They in turn become part of our, now more specific, concepts, and they thus guide us in determining the truth or falsehood of identity-claims concerning members of a given species. But they guide us in becoming part of what Wiggins would call a 'nominal essence' (i.e. by becoming part of the criteria for the application of the concept in question), not by fixing the extension of the concept via the notion of a real essence.

Thus I am inclined to conclude that the criteria of identity specified (if only in the weak Strawsonian sense) by a given sortal concept have to be understood as *conceptually determined* in a more narrow sense than the one envisaged by Wiggins. They have to be determined by *logical rules*, i.e. by *conventions*, not by inductive correlation discovered in experience.

It is crucial to note in this context that it is not as though this 'conceptualist' approach to questions of identity advocated here is simply not interested in the question of what an entity's identity 'really consists in', or what 'constitutes' an entity's numerical identity over time. This would be a misunderstanding. Since, as Wittgenstein once put it, 'essence is expressed in grammar'³³, a criterion of identity can also be said to constitute (at least part of) the essence of what the expression is properly applied to. Equally, the conditions of synchronic and diachronic identity specified by a given sortal concept ϕ (i.e. the criteria we use in determining the truth or falsehood of identity-claims of the form ' x_1 is the same ϕ as x_2 ') are constitutive of what it is to be *one* ϕ at a given time, and what it is to be *one and the same* ϕ over time. Stating the criteria we use in determining matters of identity of ϕ s is an answer (and indeed, in my understanding, the only possible answer) to the (ontological) questions 'What does the numerical identity over time of ϕ s consist in?' or 'What *makes* x_1 at t_1 numerically identical

³²Cf. Strawson, 'Entity and Identity', pp. 197 ff.

³³ Wittgenstein, *Philosophical Investigations*, §371.

with x_2 at t_2 ?'. For, as I argued above, a philosophical analysis of what identity for entities of a given kind ϕ 'consists in' can only be a conceptual analysis, i.e. an analysis of *what we mean* when we say x_1 is the same ϕ as x_2 . And what we mean when we predicate identity of entities of a given kind is brought out by the criteria we would apply in order to answer questions of identity with respect to members of that kind.

Fortunately, the controversy between the 'conceptualist' account of a criterion of identity advanced here and a conception that understands itself as more 'realist' or 'ontological' (although I think this would be a *misunderstanding*) is not so much about *which* criteria are actually applied in determining questions of identity with respect to a given kind, but more about the theoretical status which is to be attributed to them and about the direction justification should take. Whereas at least some of the theorists who view themselves as realists seem to imply that in finding out whether an entity of a given kind x_1 is identical with an entity x_2 we *apply* the criteria we do in fact apply *because* we know them to be at the same time the 'ontological' criteria of these entities and thence to constitute their identities (and are thus, and only thus, justified in the *application* of these criteria), the adherent of a 'conceptualist' conception of criteria of identity would maintain that it is just the other way round: we accept certain criteria as *constituting* an entity's identity simply *because* these are the 'epistemological' criteria we use (or at least would use, if we could) as evidence in determining the identity of such an entity. Whereas the 'realists' would claim that we have *these* criteria of distinctness and reidentification for ϕ s because we have *these* criteria of synchronic and diachronic identity, their 'conceptualist' opponents would hold that we have *these* criteria of synchronic and diachronic identity because we have *these* criteria of distinctness and reidentification. And they would add that the choice of just these criteria need not, nor could it possibly, be justified in terms of the way the world is; there is no such thing as justifying a choice or convention in terms of reality. Conceptual (or 'grammatical') rules (such as the one expressed in saying that *these* criteria of identity are used to determine the identity of ϕ s) are not responsible to reality in terms of truth or falsehood. They are antecedent to truth and falsehood. They are neither true nor false, but determine only what makes sense, and what not. *That* these criteria are used in order to determine the identity of ϕ s partly defines the concept of a ϕ ; it partly defines what it is to *be* a ϕ . But concepts do not predetermine reality in any way. They only delimit a logical space which it is up to reality to fill in or not. And of course some concepts are more precise than others. Some can serve purposes that other concepts cannot. Thus

natural kind concepts typically specify criteria of identity of the entities that are members of the kind, and they typically can be used to answer the Aristotelian ‘What is it?’-question. Other concepts cannot do that. Thus natural kind concepts may be said to have a privileged, or at least special, position in our discourse about particulars, in so far as they are more prone to be revised and sharpened in the light of new empirical discoveries concerning members of its extension. But there is no such thing as a right concept, a concept that, in Kant’s words, ‘carves nature at its joints’.

To a large extent, I think, the divergences between the two opposing camps are based on mutual misunderstandings, a claim which may be supported by the observation that divergences concerning the theoretical status of criteria of identity hardly ever seem to result in diverging accounts of *what* criteria are actually applied in answering questions of identity. Wiggins has claimed that, properly understood, ‘realism’ and ‘conceptualism’ will collapse into ‘perfect consonance’³⁴. This may well be true. However, in order to understand *why* it is true, we will have to further examine the logical relation between the notion of a *criterion* of identity and the notion of *evidence* for the truth or falsehood of an identity-statement. The upshot of this discussion will be that a criterion of identity, if it is to be a *criterion* at all, must be *applicable* (if not in practice, at least *in principle*) as evidence in determining the truth or falsehood of an identity-statement.

2.1.3. *Criteria of Identity and Evidence for Identity*

I said above that, *prima facie*, the term ‘criterion of identity’ is ambiguous in at least two respects. Firstly, it is ambiguous between the notion of a criterion of application of a given sortal concept ϕ (the identification-question) and the criteria of synchronic and diachronic identity of entities (the counting and reidentification questions) which fall under the concept ϕ (although, as we have seen, these two issues are systematically related). Secondly, it might seem, it is ambiguous between the epistemological notion of a *way of telling* whether some state of affairs p obtains, and the ontological notion of what is *constitutive of* this state of affairs p obtaining (or, put differently, the semantic notion of what is constitutive of the *meaning* of a statement that p). With regard to questions of a particular’s numerical identity over *time*, this may be taken to imply that the notion of a criterion of identity is ambiguous between the *evidence* we use in judging whether a particular x_1 at t_1 is numerically identical with a particular x_2

³⁴ Wiggins, *Sameness and Substance*, p.141, and the passage on pp.141/1.

at t_2 , on the one hand, and *what this is evidence for* (namely the obtaining of identity), on the other.

There has been much confusion about this distinction in recent discussions of personal identity. Thus when, e.g., *Parfit* speaks of a criterion of identity then he conceives it ontologically as a *necessary and sufficient condition* of identity³⁵. A criterion of an entity's numerical identity is supposed to tell us what this entity's identity 'necessarily involves, or consists in'³⁶. It is, as Parfit also puts it, 'what makes'³⁷ an entity x_1 at t_1 numerically identical with an entity x_2 at t_2 . *Noonan* distinguishes between 'constitutive' (which he characterizes as 'metaphysical-cum-semantic') and 'evidential' criteria of personal identity³⁸. Similarly, *Wiggins* characterizes his search for the criterion of identity of entities of a given kind as the search for 'what *constitutes* identity' with respect to entities of this kind, and he explicitly contrasts this notion with that of any kind of evidential criterion. He says:

'Some philosophers have used criterion to mean "way of telling" or "conceptually determined way of telling". Where they have meant this, there is a contrast with my theory. The Aristotelian *what is it* question does both less and more than provide what counts as evidence for or against identity. It does less because it may not suggest any immediate tests at all. It does more because it provides that which *organizes* the tests or evidence [...]. That is an account of what *constitutes* identity.'³⁹

Equally, *Lowe* tries to clarify his use of the term 'criterion of identity' by stating that

'it is not a requirement of a criterion of identity in my sense that it should necessarily provide us with an effective means of coming to *know* whether or not a given identity statement, 'x is identical with y' is true: rather it should tell us, so to speak, *what it takes* for x and y to be the same or different...' ⁴⁰.

In a similar vein, Perry, Lewis, and Shoemaker have tried to distinguish constitutive from evidential criteria by presenting their analyses of what constitutes the numerical identity over time of a given kind of particulars in terms of the notion of a *unity-relation*, by which they mean the relation that holds between different temporal stages

³⁵ Cf. Parfit, *Reasons and Persons*, p. 202.

³⁶ Parfit, *Reasons and Persons*, p. 203.

³⁷ Ibid..

³⁸ Cf. Noonan, *Personal Identity*, p.2.

³⁹ Wiggins, *Sameness and Substance*, p.53.

⁴⁰ Lowe, *Kinds of Being*, p.16.

of a given particular of a given kind if, and only if, they are stages of one and the same particular⁴¹.

This first understanding of the term 'criterion of identity' may be illustrated by the example of inanimate *physical objects* (such as stones, chairs, buildings, etc.). Arguably, a 'merely' physical object x_1 at t_1 is identical with an object x_2 at t_2 if, and only if, x_1 is (to some degree) spatio-temporally continuous with x_2 . Spatio-temporal continuity is, in a loose sense, what numerical identity 'means' with respect to these kinds of physical objects. Or, in other words, the fact that spatio-temporal continuity is our criterion for determining the truth of a claim that x_1 at t_1 is the same ϕ as x_2 at t_2 is at least partly meaning-constitutive of such a claim. And in this sense it can be said to be the (logically) *constitutive* criterion of numerical identity for physical objects. It is 'what makes' different 'physical object-stages' stages of one and the same physical object. It is the *unity-relation* for physical object-identity over time.

So much for the first understanding of the term 'criterion of identity'. Alternatively, when *Chisholm*, and similarly *Swinburne*, speak of our having a criterion of identity for a certain kind of entity ϕ , what they mean by this is only that we typically regard a certain kind of state of affairs as *evidence* for the identities of ϕ s. A criterion of identity in this understanding is a generally good, but not decisive ground for predicated identity. As such it is *logically independent* of what it is a criterion for⁴², namely the obtaining of numerical identity. To give an example concerning personal identity:

⁴¹ Cf. Perry, 'The Problem of Personal Identity', passim; and Lewis 'Survival and Identity' passim; and also Shoemaker, in Shoemaker/Swinburne, *Personal Identity*, p. 74.). Shoemaker at least seems to conceive of the analysis of numerical identity in terms of a unity-relation as a **conceptual** analysis (cf. op.cit., p. 127), which is not clear as to Perry and Lewis.

In what follows I shall disregard the 'stage-terminology'. The Lewis-Perry-Shoemaker approach to questions of diachronic identity in terms of a unity-relation unifying different temporal stages of a numerically identical object may be justified in cases in which the entities united by the unity-relation are identificationally more basic than the entities into whose numerical identity they are united. However, this does not seem to be the case for persons. Person-stages are entirely artificial entities. In fact, it is not quite clear what they are, -- if not very short-lived persons.

⁴² Cf. Chisholm, *Person and Object*, p.112. Swinburne does not explicitly speak of criteria, but the distinction he makes between the meaning of a claim and the evidence we have for it parallels Chisholm's conception, cf. Swinburne/Shoemaker, *Personal Identity*, p.3.

qualitative identity of fingerprints is considered as a generally good, but *purely evidential* (or *inductive*), criterion of numerical identity of person, and it is used as such by various law enforcement agencies all over the world. The reason for this is that, as *experience* has taught us, identity of fingerprints and identity of person normally go together. 'Same fingerprints' normally implies 'same person'; which is the reason why checking fingerprints has become a convenient way of telling people apart, and tracing their identities through time. But identity of fingerprints is not a constitutive, and hence indefeasible, criterion of personal identity. We know that clever surgeons can alter a person's fingerprints, and we also have to account for the off-chance of two different persons having qualitatively identical fingerprints. Identity of fingerprint and identity of person can part company in two different directions. And the same can happen with any state of affairs that is commonly used as evidence in deciding matters of personal identity. And this, the argument continues, implies that the truth of an identity-statement concerning persons is always logically independent of the truth of any statement citing evidence for or against the obtaining of personal identity.

Thus it might seem as if the cited ambiguity in the notion of a criterion could be resolved simply by distinguishing between purely constitutive and purely evidential criteria of identity. However, this view of the matter would be too optimistic. For what underlies the cited divergences in the use of the term 'criterion' in general, and 'criterion of identity' in particular is more than just terminological divergence. It is a cluster of profound controversies about a correct account of the relation between semantics, epistemology and ontology. Let us take a step back:

What unites the Parfit-Wiggins and the Chisholm-Swinburne approaches to questions of numerical identity over time is the shared preconception that we can in general draw the distinction between our *way of telling* whether an entity x_1 is identical with an entity x_2 , on the one hand, and what this identity *consists in*, on the other. They would (presumably) all agree that it is one thing to state the evidence we typically use in finding out about, or determining, the identity of an entity x_1 with an entity x_2 , and another thing altogether to state 'what it means to say' that x_1 is identical with x_2 , and thence 'what is the case' if x_1 is identical with x_2 . Stating the (ontologically conceived) criterion, i.e. the necessary and sufficient condition, of a given entity's identity over time may serve as an elucidation of what we mean by predicating identity over time for a given kind of entities ϕ -- not, perhaps, what naively and at first thought, but what on reflection and on second thoughts we take ourselves to mean. Thus, it may be held, what

it *means* to say that a 'merely' physical object x_1 is identical with an object x_2 is precisely that x_1 is *spatio-temporally continuous* with x_2 . x_1 is numerically identical with x_2 *if and only if* x_1 is spatio-temporally continuous with x_2 . As far as 'merely' physical objects are concerned, the meaning of identity-statements can thus be explained in terms of statements about spatio-temporal continuity. And in this sense spatio-temporal continuity can be said to be the constitutive, or ontological, criterion of numerical identity with regard to physical objects. It is our *principium individuationis* for physical objects. Of course, a criterion of identity, thus conceived, may also serve us as a means of finding out about, or ascertaining, the identities of physical objects. That is to say, a constitutive criterion of identity *may*, under certain favourable circumstances, be used as a *way of telling whether identity obtains*, i.e. it may be used as an evidential criterion of identity. For example, in order to find out whether a physical object x_1 is identical with a physical object x_2 we will try to *trace* a spatio-temporal path between the two. And although we are not generally in a position to do that (that is, although we cannot *generally*, or *in practice*, use spatio-temporal continuity as an evidential criterion), this is what we would do, if we could, and the result of our investigation would then settle the matter.

However, the constitutive criteria theorist continues, the fact that a constitutive criterion of identity may be used as evidence is in no way essential to its being what it is. Rather, it is only because it is a constitutive criterion (and thus a necessary and sufficient condition of what it is a criterion for) that its application as an evidential criterion can confer certainty on the identity-fact in question. In fact, it is held, this is just what distinguishes a constitutive from a purely evidential, or inductive, criterion. An inductive evidential criterion is in principle always defeasible by better evidence pointing to the contrary, a constitutive criterion of identity is not. If the constitutive criterion obtains, identity obtains, and vice versa. Apart from that, the meaning of an identity-statement is logically independent of the epistemic means we use in finding out about its truth. A criterion of identity is either necessary and sufficient, and then it is *meaning-constitutive* of the respective identity-statement (and hence constitutive of the obtaining of identity), or it is *defeasible*, and then it is purely evidential, or inductive⁴³.

⁴³ Cf. Chisholm, *Person and Object*, p.112. Cf. Swinburne's criticism of what he calls 'empiricist theories' of personal identity, in Shoemaker/Swinburne, *Personal Identity*, pp. 3-21.

This is precisely what is denied by the alternative *Wittgensteinian* conception of a criterion in general, and a criterion of identity in particular. In order to understand the Wittgensteinian notion of a criterion of *identity* we will have to take a step back and first examine Wittgenstein's notion of a criterion *in general*.

We mentioned above that when philosophers talk about criteria *tout court* they often mean a criterion of *application* of a given concept-term which can be given in terms of necessary and sufficient conditions (i.e. in an analytical definition). Wittgenstein's conception of a criterion, as it is used in the *Blue Book* and the *Philosophical Investigations*, is related to this notion, but it differs in at least one important respect: namely in the assumption that there can be logical relations between statements, which are in principle defeasible, and hence fall short of entailment. Put briefly, Wittgenstein conceived of a criterion as in principle *defeasible, but nevertheless meaning-constitutive*, and thence *non-inductive, evidence* in favour of the truth of a judgement⁴⁴.

[According to Wittgenstein, the *meaning* of a word is given by an explanation of its use. For some terms this is done by giving ostensive definitions (e.g. colour-terms), for others this may be done by citing necessary and sufficient conditions of their correct application (e.g. technical, scientific, mathematical, some legal terms.). However, the meaning of at least some kinds of expressions (in particular psychological predicates and dispositional predicates) is, Wittgenstein claimed, at least partly explained in terms of their *criteria*. Thus the concept of an ability can be explained by reference to a typical performance of the respective ability, and the concept of a mental state can be explained by reference to one of its characteristic manifestations or expressions. When we explain the meaning of an expression of the appropriate kind in terms of its criteria, we spell out the criterial relations that connect it with other expressions in our language. Criterial relations are determined by conventions, not by facts; they are not inductive correlations discovered in experience. That is to say, criterial relations between different expressions are fixed by what Wittgenstein calls 'grammatical rules', i.e. rules which determine the meaning of a term. Thus, to give an example, the statement that a person's displaying pain-behaviour (which includes the person *saying* that he is in pain) is a criterion for the person's being in pain, expresses a grammatical rule for the use of the predicate 'is in pain'. And similarly, the statement that a person's playing the piano is a criterion for that person's ability to do so, expresses a grammatical rule for the use of

⁴⁴ Wittgenstein's notion of a criterion has been the topic of a heated debate for many years. I shall ignore details, and instead focus on the common elements of various otherwise diverging interpretations. Cf. e.g. Albritton, 'On Wittgenstein's Use of the Term "Criterion"', Kenny, 'Criterion' in: Edwards, ed. *Encyclopedia of Philosophy*, vol.2, pp. 258-261, Hacker, *Insight and Illusion* Ch. XI, and McDowell, 'Criteria, Defeasibility, and Knowledge'.

the predicate 'can play the piano'. That a person's displaying pain-behaviour justifies the ascription of pain to that person is a feature of the grammar of 'is in pain'. That a person's playing the piano justifies the ascription of the ability to play the piano to this person is a feature of the grammar of 'can play the piano'. It determines its correct application and is thus at least partly *meaning-constitutive* of the expression 'can play the piano'; or, put directly (i.e. in ontological, not semantic, terms), it is partly constitutive of what the ability to play the piano 'consists in'.

Thus, in giving the criteria for one's claim that p one may not only give an explanation of its meaning, but at the same time *justify* one's claim that p, or, put differently, give *reasons* for p's being true. Checking whether the criteria for p (i.e. the states of affairs that constitute criteria for p) obtain is a (in fact the only) way of finding out if p obtains. That is to say, criteria can be viewed as *evidence* for the obtaining of a given state of affairs p. Criterial evidence is *non-inductive*. A criterion for p is, as Shoemaker put it, a 'logically necessarily good evidence'⁴⁵ for the obtaining of p. This is precisely what distinguishes criteria from *symptoms*, which Wittgenstein conceives as purely inductive and not meaning-constitutive evidence for the truth of a judgement. That the obtaining of one state of affairs p is a symptomatic evidence for the obtaining of another q is something we discover in experience. It thence presupposes that p and q are identifiable *independently* of each other. However, this is not the case for criterial evidence. There is no better way of finding out that somebody is in pain than by observing his behaviour, just as there is no better way of finding out that somebody can play the piano than by hearing him play. There is no 'more direct' access to another person's mental states and abilities than by observing their manifestations -- in the sense in which there is (at least in principle) a more direct access to a disease than via its outwardly observable symptoms.

Although he did not deny that there are expressions whose meaning can be explained in terms of necessary and sufficient conditions, Wittgenstein seems to have restricted his application of the concept of a criterion to those expressions that cannot be thus explained. That is to say, with the possible exception of what Wittgenstein sometimes calls 'defining criteria'⁴⁶, he seems to have conceived of criteria as in principle *defeasible*. In other words, that a criterion for a claim that p is at least partly meaning-constitutive of p, and that it is a logically good evidence for the obtaining of p, is not to say that it is either a necessary or a sufficient condition of its obtaining. Put differently, if p is a criterion for the obtaining of q, then p does not entail q, nor vice versa. A person may display pain-behaviour, but be pretending. And, vice versa, a

⁴⁵Shoemaker, *Self-Knowledge and Self-Identity*, p.3.

⁴⁶Cf, the discussion in Hacker, *Insight and Illusion*, p. 311.

person may be in pain, whilst not displaying any kind of pain-behaviour. Equally, a person's performance may be a fluke. And, vice versa, a person may have an ability without ever manifesting it. Nevertheless, Wittgenstein claims, it would not be rationally justified to deny the obtaining of *q* in the presence of *p* without being able to cite any defeating condition for *p*, i.e. any other good reasons to the effect that *q* does not hold (and the number of those is up to our imagination). And, vice versa, it would not be rationally justified to assert the obtaining of *q* without being able to cite any criterion *p* that speaks in its favour.

It is, of course, this combination of meaning-constitutiveness and defeasibility in Wittgenstein's notion of a criterion that has seemed mysterious to many interpreters. I shall not make an attempt to justify this conception on a general level. The question I should like to address in what follows is rather whether we can make sense of the possibility of defeasible criteria of *numerical identity*. This question is of particular relevance to the question of personal identity which has often been said to have multiple, mutually irreducible, criteria of identity which are nevertheless mutually defeasible in favour of each other⁴⁷. I shall argue that, although we have to allow for the possibility of various forms of evidence for identity which are meaning-constitutive of the respective identity-statements (and hence non-inductive) whilst at the same time neither necessary nor sufficient, these meaning-constitutive forms of evidence cannot fulfil the need for a sufficient criterion of numerical identity, which alone can establish the *contrast* between numerical and merely qualitative identity for particulars.

2.1.4. Numerical Identity, Qualitative Identity, and Spatio-Temporal Continuity

Above I mentioned the view that, at least as 'merely' physical particulars are concerned, spatio-temporal continuity is the necessary and sufficient criterion of numerical identity over time. As it stands, however, this thesis is vacuous, or at least in significant respects unsatisfactory, since it does not adequately reflect the mutual logical interdependence between the notions of qualitative identity and spatio-temporal continuity in our identity-discourse on particulars. This can be brought out by the following

⁴⁷This seems to be implied in Wittgenstein's discussion of personal identity, cf. *The Blue and Brown Books*, pp. 61f.. More explicitly, such a view is to be found in Wilkerson, *Minds, Brains, and People*, pp. 27-30.

considerations:

Firstly, and most importantly, the notion of spatio-temporal continuity is by itself, as Wiggins put it, 'purely formal'⁴⁸. That is to say, it does not by itself provide a 'principle of tracing'⁴⁹ a given particular as numerically identical through space and time. Spatio-temporal continuity means something different for different *kinds* of particulars. And hence it cannot provide a principle of tracing *in abstracto*, valid for particulars of all kinds. It cannot provide a principle of tracing unless it is first specified *what* kind of thing the particular in question is, i.e. unless it is first specified '*under which sortal concept*'⁵⁰ the particular is to be traced. That is to say, we first have to have identified a particular x_1 as of a certain kind ϕ (we have to give an answer to the identification-question) before we can even try to answer the (reidentification) question whether x_1 at t_1 is identical with x_2 at t_2 . We cannot even try to give an answer to the question 'Is x_1 the same as x_2 ?' unless we first have an answer to the question 'Same *what*?'. The reason for this *essential* dependence of the reidentification question on the identification question is that, as we have seen above, only sortal concepts can provide us with criteria of identity for the particulars that fall under them. Only sortal concepts -- to attenuate the claim in the light of Strawson's criticism -- specify certain 'readinesses for variation' (criteria of identification) and 'expectations of typical continuities and modifications' (criteria of distinctness and reidentification) for all the members of its extension. In doing so, sortal concepts determine the relevant *qualitative* respects in which a particular x_1 of a given kind ϕ has to be spatio-temporally continuous with a particular x_2 in order for x_1 to count as numerically identical with x_2 . And the qualitative respects which play a role in determining what is to count as the spatio-temporal continuity of a given particular x_1 of the kind ϕ will at the same time be the qualitative respects which play a role in determining whether x_1 is to count as a ϕ at all. To illustrate this very abstract claim by a concrete example: in determining the identity over time of an artefact (e.g. a table) we will not take into account changes in its colour, its temperature, or its surface texture. But we will take into account changes in its shape, size, and, most significantly, *material constitution*. Arguably (though perhaps too strictly), a table x_1 at t_1

⁴⁸Cf. Wiggins, *Identity and Spatio-Temporal Continuity*, p.35.

⁴⁹Wiggins, op.cit. p.36.

⁵⁰Ibid..

cannot be identical with a table x_2 at t_2 unless it is constituted of at least roughly the same particles of wood. As concerns the identity over time of an animal (e.g. a dog), on the other hand, the *matter* of which a given organism is constituted at different times is quite irrelevant to determining its identity through time. A given particular animal x_1 at t_1 may be constituted of completely different matter from an animal x_2 at t_2 , and still x_1 and x_2 may be one and the same animal. In this sense, it may be maintained, a sortal term specifies the qualitative continuities relevant in determining identity-questions.

Secondly, it is worth pointing out that spatio-temporal continuity is hardly ever *actually*, or in practice, used as evidence in determining the truth or falsehood of an identity-statement of the form ' x_1 at t_1 is identical with x_2 at t_2 '. We hardly ever find ourselves in a position where we can *continuously* observe a given particular x_1 at t_1 as numerically identical with a particular x_2 at t_2 . Normally, reidentification of a given particular takes place across an 'observational gap'⁵¹, i.e. across a period of time during which we have not continuously observed the particular. The most commonly applied way of telling whether a given particular x_1 at t_1 is identical with x_2 at t_2 , or not, is by its relevant *qualities*. To give an example: When I wake up in the morning and see the chair next to my bed, I assume that it is numerically identical with the chair I saw when I went to bed, not on the grounds that it is spatio-temporally continuous with the chair I saw when I went to bed, but on the grounds that it is in the relevant respects *qualitatively identical* with it. That is to say, *qualitative*, rather than spatio-temporal, *continuity* (where qualitative continuity may be understood as qualitative similarity from any given moment to any other) would be my reason for saying that the chair I saw when I went to bed is the chair I see now (assuming, of course, there are not two or more qualitatively identical chairs in my room). This claim, however, although true, has to be further qualified.

It is certainly true that qualitative identity (in the relevant respects specified by a given sortal concepts) is most commonly used as evidence in determining numerical identity over time. And this is so because, firstly, in many cases a given particular's spatio-temporal history is in practice inaccessible to us (we do not know what spatio-temporal path the particular traced in our 'perceptual absence'), and because, secondly, most particulars are distinguishable by their perceptible qualities. Nevertheless, complete qualitative identity of two numerically distinct particulars always remains a theoretical

⁵¹Cf. Wiggins, 'The Individuation of Things and Places', p.184.

(if perhaps not practical) possibility (which is why Leibniz' principle of the identity of indiscernibles can only provide a sufficient criterion of difference, but not of identity, for particulars)⁵². And from this it follows that the qualitative identity (however complete) of a given particular x_1 at t_1 with a given particular x_2 at t_2 can only ever constitute in principle *defeasible* evidence for the obtaining of numerical identity between x_1 and x_2 . If I was told that whilst I was asleep the workmen came into my room and replaced the chair I saw when I went to bed with a qualitatively identical chair, I would revise my judgement in the light of this newly acquired evidence. And the defeating criterion in this case is, of course, spatio-temporal continuity (or rather the lack thereof). The chair I see in the morning is not spatio-temporally continuous with -- has a different spatio-temporal history from --- the chair I saw when I went to bed. Although qualitatively identical, the two chairs are distinct simply because they have traced two different paths through space and time. And distinct spatial position at any given time is itself a sufficient reason for denying numerical identity. Thus, although some kind of qualitative *continuity* may be a necessary condition of numerical identity, qualitative identity by itself can *by definition* never be a logically sufficient criterion of numerical identity, whereas spatio-temporal continuity, if it obtains at all, is both necessary and sufficient.

One may be tempted to conclude from this analysis that (1), whereas various forms of *qualitative identity* can be evidence for the obtaining of numerical identity, they cannot be criterial evidence -- since, being *defeasible*, they cannot be meaning-constitutive of the respective identity-statements; whilst (2), on the other hand, *spatio-temporal continuity* is a necessary and sufficient criterion of numerical identity of particulars, -- not, however, because it can actually be applied as evidence for the obtaining of numerical identity, but simply because spatio-temporal continuity is what numerical identity 'consists in'. However, I think that both halves of this conclusion should be resisted, since they misrepresent the logical interplay of these two notions in our identity-discourse on particulars.

⁵²Unless, of course, one adopts Leibniz' reductionism about spatial properties, and some other, highly implausible, 'revisionary' assumptions. For a criticism of Leibniz' views on numerical identity cf. Wiggins, *Identity and Spatio-Temporal Continuity*, pp.34f., and his 'The Individuation of Things and Places', where he shows that the distinguishability of particulars in terms of their spatial and temporal relations cannot be conceived as an individuation in pure predicates, but already presupposes the individuation of other particulars; cf. also Strawson, *Individuals*, pp.122/3. The points made in both discussion go back to Kant's claim that space and time are intuitions, not concepts.

Let us first scrutinize the *first half* of the suggested contrast. We have seen above that the applicability of the criterion of spatio-temporal continuity, being completely formal, is logically dependent on the specification of a sortal concept under which a given particular is to be traced. A sortal concept, we said, by specifying certain 'expectations of typical continuity' of the particulars falling under it, specifies at the same time the qualitative continuities which are relevant to determining the numerical identity over time of a given particular of the given sortal kind. We also saw that these qualitative continuities are the very same as the ones which are relevant to determining *what kind* the particular is of. For in identifying a given particular as falling under a sortal concept ϕ , we do not identify it, as it were, momentarily. Rather, the identification of a given particular as a ϕ includes within it certain expectations concerning a ϕ 's typical history and continuity. If these expectations are not fulfilled, we may come to doubt if our identification of a given particular x as a ϕ was correct in the first place. (To give an example, we may come to doubt if a 'chair' which turns into a giant frog the minute we try to sit in it, was correctly identified as a chair in the first place.) Thus, although there indubitably are various forms of qualitative identity which have to count as inductive evidence for the obtaining of numerical identity, the mere fact that a certain kind of qualitative identity is defeasible evidence for numerical identity does not by itself make it inductive. There may be various forms of qualitative continuity which are such that, if -- in a particular case -- they could not count as evidence for the numerical identity of a ϕ any more, the correctness of the identification of a given particular as a ϕ would itself become questionable. That is to say, the fact that certain qualitative continuities go together may be a condition of our identifying a particular x as a ϕ , and thus a condition of our reidentifying a particular x_2 at t_2 as the same ϕ as x_1 at t_1 . To anticipate a claim which will be spelled out only much later: identity of fingerprint may be a purely inductive (although nevertheless very reliable), and hence not meaning-constitutive, evidence for the obtaining of personal identity. Psychological continuity (i.e. similarity between P_1 at t_1 and P_2 at t_2 in terms of character-traits, memories, etc.), on the other hand, may not be purely inductive evidence, since it is partly constitutive of how we determine whether an entity is a person at all, and thus partly meaning-constitutive of statements concerning personal identity through time.

Thus, in the sense just specified, we should admit of the possibility that there are criteria of numerical identity (in the sense of *meaning-constitutive* evidence for the truth of an identity-statement) which are not logically sufficient for the obtaining of

identity, and hence are defeasible. It is crucial, however, to note that the sense in which there can be defeasible criteria of identity is significantly different from the sense in which this notion was introduced by Wittgenstein. Wittgenstein's discussions of defeasibility typically center on concepts whose meaning is determined by a *multiplicity* of criteria for the application of the given concept. The various criteria for the application of the concept may diverge in certain abnormal cases, although their coincidence in general is a background-condition of our having the given concept at all⁵³. Most importantly, however, the various criteria are *on the same logical level*. That is to say, every criterion is as good a reason for the application of the given concept as any other. And this is just what distinguishes the defeasibility of the various criteria of (e.g.) understanding from the defeasibility of various forms of qualitative identity, and what explains why the logical interdependence of qualitative identity and spatio-temporal continuity *cannot* be characterized along these lines. For qualitative identity and spatio-temporal continuity are *not* on the same logical level. The criteria invoking various kinds of qualitative identity or similarity between objects are *by definition* not sufficient for the obtaining of numerical identity. Their defeasibility in favour of the necessary and sufficient (but in practice rarely applicable) criterion of spatio-temporal continuity is just what opens up the *possibility of a contrast* between merely qualitative and strictly numerical identity for particulars. It is thus partly *constitutive* of this very contrast.

This becomes even more obvious if we have a look at the *second half* of the conclusion suggested above, namely the claim that spatio-temporal continuity is the (necessary and sufficient) criterion of numerical identity *independently* of its actual *applicability* as a criterion of (i.e. meaning-constitutive evidence for) numerical identity. (As we have seen above, 'constitutive' criteria are often distinguished from 'purely evidential' criteria by pointing out that the latter, but not the former, are generally used as evidence in *finding out* about, or getting to *know* about, the truth or falsehood of a given identity-statement.) However, although spatio-temporal continuity cannot usually be applied in practice as a criterion of identity (i.e. as evidence for determining the truth or falsehood of an identity-judgement), it is just the fact that, given the opportunity, we *would* apply it, if we could, that makes it a *criterion* of identity in the first place. If a so-called constitutive criterion were *in principle* (not just in practice) inapplicable as a

⁵³Cf. Wittgenstein's discussion of understanding, *Philosophical Investigations*, §§ 139ff..

way of telling whether identity obtains, it could not be said to be a criterion of identity at all. It is essential to its being a criterion that it may (at least in principle) be used as a *way of telling whether identity obtains*, i.e. that it may be used as an *evidential* criterion of identity. And spatio-temporal continuity fulfils this requirement. In order to find out whether a given ('merely') material particular x_1 at t_1 is identical with x_2 at t_2 , we will try to trace a spatio-temporal path between x_1 and x_2 . And although we are not generally in a position to do that (that is, although we cannot *generally*, or *in practice*, use spatio-temporal continuity as an evidential criterion), this is *what we would do*, if *we could*, and the result of our investigation would then settle the matter⁵⁴. The *theoretical* (if not practical) possibility of applying the criterion of spatio-temporal continuity in order to determine the truth or falsehood of a given identity-judgement is just what gives *sense* to our identity-statements in cases in which we are *in practice* excluded from finding out about its *truth*. Its applicability in theory (if not in practice) is what opens up the possibility of a contrast between numerical and merely qualitative identity in our conceptual scheme, even in cases in which we are in practice prevented from applying it as the (necessary and sufficient) criterion of identity (i.e. as evidence which would settle the truth or falsehood of a given identity-judgement).

2.1.5. What is a Criterion of Personal Identity?

So far we have only considered *spatio-temporal continuity* as a candidate for a necessary and sufficient criterion of numerical identity for a certain kind of particulars. However, our discussion of spatio-temporal continuity allows us to draw up a list of interrelated requirements which *any* such proposed criterion has to fulfil in order to count as a criterion of numerical identity for particulars.

1. A criterion of identity must be, although not in practice applied⁵⁵, at least in principle applicable. This *applicability-requirement* does not exclude the possibility that under certain circumstances we may be unable to find out whether a given identity-

⁵⁴Cf. Wiggins, *Identity and Spatio-Temporal Continuity*, p.34f., and 'The Individuation of Things and Places', pp.184-187.

⁵⁵*Pace* Wittgenstein, *The Blue and Brown Books*, p. 61f, and Strawson, 'Entity and Identity', p.219, who both seem to imply that a criterion of identity has to be *actually* applied. Wiggins, although he makes it plain that he does not take a criterion of identity to be usually applied (with which I agree), remains equivocal on whether he requires it to be at least in principle applicable, cf. *Sameness and Substance*, p.54.

claim is true or false. Identity-statements are true or false independently of our ability to recognize their truth or falsehood. But they do not have sense, and thus the ability to determine reality save for a Yes or No, independently of our ability to *explain what we would count* as a Yes or No -- where this explanation consists in more than just a repetition of the sentence used to make the statement in question. This is the point made by the applicability-requirement. It can also be made by stating that our criteria for *numerical identity* for entities of a given kind ϕ cannot diverge from our criteria of *reidentification* of ϕ s. Put differently, the answer to the question 'What does x 's numerical identity over time consist in?', or 'What makes it the case that x_1 at t_1 is numerically identical with x_2 at t_2 ?', must be the same as the answer to the question 'Under what circumstances would we count x_1 at t_1 as numerically identical with x_2 at t_2 ?'.

2. A criterion of identity must provide for a means of 'tracking' a particular as numerically identical over time (the *tracking-requirement*). This is just a consequence of the previous requirement.

3. A necessary and sufficient criterion of numerical identity must be able to provide for the possibility of a contrast between strict numerical and merely qualitative identity (the *contrast-requirement*). The possibility of such a contrast is just what distinguishes particulars from universals or types. Hence, if a proposed criterion of identity for a certain kind of particulars is supposed to be sufficient, it has to provide for the possibility of such a contrast, and at the same provide a *means of distinguishing* in any given case between strict numerical and merely qualitative identity.

Above I restricted the range of the claim that spatio-temporal continuity is our necessary and sufficient criterion for particulars to what I (vaguely) called 'merely' physical objects. However, I think it is easily agreed that -- in the wide sense in which the notion has been employed in the present context -- spatio-temporal continuity is the necessary and sufficient criterion not only for stones and billiard-balls, but also for artefacts, plants, and animals. Of course, how it is applied in a given case will depend on the kind of particular in question. Spatio-temporal continuity 'means' something different for stones than it means for animals. Philosophical treatments of identity, such as Locke's and Wiggins', may be understood as investigating the significance of spatio-temporal continuity for various kinds, or 'categories', of particulars (such as artefacts, plants, animals). At the most general level, however, we can say that *any* particular of *any* of these kinds is individuated via its position in our spatio-temporal framework, i.e. that its identity is determined by its spatio-temporal history.

Persons, beings like you and me, are particulars. Hence we might expect the criterion of spatio-temporal continuity to apply to their identities as well. However, in trying to do so we seem to run up against considerable intuitive obstacles. And this difficulty we have in accepting the criterion of spatio-temporal continuity as governing *our own* identity is, I suggest, what the *problem of personal identity* consists in. It consists in the fact that we are prone to believe that persons are different from *all* other kinds of particulars in that they are not individuated in space. We are prone to believe that the numerical identity of a given person over time is *not* governed by the criterion of spatio-temporal continuity of a person's *living body*, but by some radically different criterion. If we can locate the source of these beliefs and associated intuitions, we have already come half-way to a solution of the problem. I suggest that the source of these intuitions is the possibility of adopting a *first-person stance* to matters of personal identity, and that the intuitions arising from adopting this stance are (mostly) illusions. If we can expose these intuitions as the *illusions* they are, the problem of personal identity is solved.

2.2. *Personal Identity and the First-Person Point of View*

It has been claimed that, as opposed to all other kinds of identity, personal identity allows of two different ways of access, one 'from the inside', and one 'from the outside', or, in other words, one from the first-person, and one from the third-person point of view; and that it is this peculiar epistemological feature that makes personal identity special, and more worthy a topic of philosophical inquiry than the identities of dustbins and hairpins. A closely related claim often found in the literature on personal identity is the claim that a person has a 'special access' to his own numerical identity over time⁵⁶, and this claim probably goes back as far as Locke's definition of a person as a

'thinking intelligent being that has reason and reflection and can *consider itself as itself*, the same thinking thing in different times and places [my italics];'⁵⁷.

It will be part of the aim of this treatise to clarify what this 'considering oneself as oneself' really amounts to, and whether the putative special access that a person is supposed to have to his own identity really exists. On a superficial level, however, I

⁵⁶Cf. e.g. Shoemaker, in Shoemaker / Swinburne, *Personal Identity*, p.105.

⁵⁷Locke, *Essay*, Book II, xxvii 9.

think we all know perfectly well what is meant by those terms. And the imaginary cases construed in the more recent contributions to the topic of personal identity certainly manifest no dearth of appeals to the first-person perspective on matters of personal identity.

Let us recall Shoemaker's original Brownson-case⁵⁸: Imagine Brownson (i.e. the person who has Robinson's body but Brown's brain) waking up in his hospital-bed. He seems horrified to find himself in Robinson's body and, when asked, immediately answers that he is Brown, not Robinson. He remembers all, or at least most of, the things that Brown has experienced throughout his life, and he displays habits and character-traits that are very similar to Brown's. After some initial hesitation, we would, it has been suggested, finally come round to the conclusion that -- against first appearances -- Brownson is identical with (is the same *person* as) Brown. In doing so we may let ourselves be guided by various kinds of criteria of identity for persons. *We*, I say. But what about *Brownson*? In which way, i.e. on the basis of which criterion, does he go about determining his identity? Guided by which criterion does he 'self-ascribe' his identity with Brown? Does he, as it were, first *compare* his memories, beliefs, habits, character-traits, etc., with those of Brown, before -- on the basis of this comparison -- he ventures the judgement that he is identical with Brown? It would seem, No. Brownson does not have to apply any criteria of identity in self-ascribing any of his memories and beliefs. And he does not have to apply *any* criteria of identity in 'self-ascribing his identity' with Brown. Brownson, it seems, *simply knows* that he is identical with Brown. He does not have any doubts at all about, and he need not apply any criteria in order to determine, what is to count as *his* past. Brownson's self-ascription of identity is -- as I will say -- *criterialess*.

Equally, imagine the temporally reversed situation: Imagine Brown having inadvertently stumbled into the 'mad scientist's' torture-chamber as described in Williams' essay 'The Self and the Future'. There he finds himself in the hands of a mad scientist who presents him with the following choice. Tonight he, Brown, is going to be connected to a machine that will record all the features of his brain that are relevant for the dispositional and occurrent features of his personality and then imprint them on Robinson's brain who is already strapped to the bed next door. The Robinson-body person will thus emerge with what appear to be Brown's memories, character-traits,

⁵⁸Cf. Shoemaker, *Self-Knowledge and Self-Identity*, pp.22 ff.

intentions, etc.. And, *mutatis mutandis*, the same thing is going to happen to Robinson. That is to say, the information encoded in Robinson's brain will be zapped by the machine and then imprinted onto the Brown-body person's brain. And the Brown-body person will thus emerge psychologically continuous with Brown. Or so the mad scientist says. Now, he, Brown, is granted the dubious privilege of choosing which of the two, the Brown-body person or the Robinson-body person, is going to be tortured. Which one is he going to choose? Which one *ought* he to choose assuming that he wants to avoid the torture?

Now it may well be that Brown is thrown into a state of confusion over this decision, and that he does not know which option to choose. However, it would appear that there is at least one fact Brown can be certain of: either he is going to be tortured tomorrow, or he is not. Either he is going to suffer all the torturing, or no torture at all; 'no half-tortures will be laid on', as Swinburne has put it sarcastically⁵⁹. From a third-person point of view we may be tempted to treat the Williams-case just described as a borderline-case and Brown's question 'Will it be *me* who...?' as an empty question to be decided by stipulation, i.e. by the introduction of a new convention rather than by an enquiry into the facts. But, as Williams pointed out, from the first-person perspective (i.e. from Brown's perspective, or from the perspective of someone who imagines being in Brown's shoes) this reaction to the dilemma posed by our thought-experiment seems simply not acceptable. For someone who imagines actually finding *himself* in the situation described, it would seem, it just will not do to be told that one's question 'Will it be me who...?' is -- despite all appearances -- without a determinate answer. As Williams puts it:

'To be told that a future situation is a borderline case for its being myself that is hurt, that it is conceptually undecidable whether it will be me or not, is something which, it seems, I can do nothing with'⁶⁰

This seemingly natural belief of ours in the absolute determinacy of questions concerning our own identities is also brought out by other imaginary cases such as the ones envisaged by Parfit⁶¹. Imagine Brown on the night before a space-travel that in-

⁵⁹Swinburne, 'Personal Identity', p.237.

⁶⁰ Williams 'The Self and the Future' p.58; cf. also Parfit, *Reasons and Persons*, p.200.

⁶¹ Cf. Parfit, *Reasons and Persons*, pp. 200-202.

volves some sort of Startrek-teletransportation. Brown wants to know whether it will be *him* who is going to awake on Mars. He wants to have a clear-cut answer to the question: 'Will *I die* or will *I survive* the voyage? Will the future of the person waking up on Mars be *my* future?' Or imagine Brown on the night before an operation which involves some form of brain-splitting, such that the two halves of Brown's brain will be implanted into two artificially created bodies, which -- assuming the brain-transfers are successful -- will both have Brown's memories, character-traits, etc.. In this case Brown will want to know which of his two successors will be *him*. 'Will *I* be Brownson₁ or Brownson₂, or neither of them?' Brown may ask. Again, *we* may decide that Brown's question is in this particular case an empty question in the sense sketched above. Society, that is, will have to come to certain arrangements. The lawyers will have to decide on one of the two as the rightful owner of Brown's property, and the rightful husband of Brown's wife. Whatever decision the lawyers would finally come up with, their answer would appear to be highly stipulative. However, although *we* may come to treat the matter in this nonchalant manner, there would seem to be at least one person for whom these questions cannot possibly appear to be empty, and this is Brown. Brown thinks that he is perfectly entitled to demand an unambiguous answer to his questions. He thinks that there is a perfectly clear and irreducible difference between *his* and someone else's awaking on Mars, and between *his* being Brownson₁ or Brownson₂. And so would probably anyone of us in Brown's position.

Parfit has argued that what comes to the fore in the thought-experiments just described are certain deeply-rooted intuitions we cherish about ourselves and our own identities, which he calls our 'most natural assumptions about ourselves'⁶². These intuitions represent our identities as being radically and essentially different from the identities of all other kinds of particulars. Insofar as they have found their way into philosophical theorizing Parfit summarizes these assumptions under the heading of the '*Simple View*', and I will follow him in this terminology. I shall here mention just two of these 'natural assumptions'.

(1) As opposed to the identities of physical objects personal identity is always *determinate*. Whereas our concept of identity is vague and has fuzzy edges when applied to physical objects, when applied to persons it is as sharp and accurate as it could possibly be. This view has found vociferous supporters amongst philosophers. Thus

⁶² Parfit, *Reasons and Persons*, p.217.

Thomas Reid claimed that as opposed to the identity of bodies, which 'is not perfect identity' but rather 'something which for convenience of speech we call identity', personal identity 'has no ambiguity, and admits of no degrees, or of more or less'. When identity is applied to persons, Reid claims, 'the notion of it is fixed and precise'⁶³. Bishop Butler sings the same tune when he distinguishes between two different senses of the word 'the same', the 'loose and popular' one being applied to objects, the 'strict and philosophical' one to persons⁶⁴. Personal identity, it seems to us, has a much firmer ontological grounding in reality than the identity of physical objects, which, allegedly, is largely a matter of our (conventional) linguistic practices. It is, as it were, not only identity, but *self*-identity. It is the only case of genuine identity we find in the natural world; all other kinds of identity-ascriptions are only 'playing loose'⁶⁵ with the concept of identity.

(2) In matters of personal identity there can therefore be *no borderline-cases*. Questions like 'Will person P_1 be person P_2 ?' or 'Has P_1 died or survived as P_2 ?' must always have a definite and clear-cut answer (though we may perhaps not be able to find it). Questions of personal identity can *never be empty questions*, to be answered by convention or stipulation rather than by reference to the facts.

I think what lies at the bottom of these intuitions (of 'our most natural assumptions about ourselves') is a person's ability to refer to himself *without application of any criteria of his identity*. Wittgenstein named the special use of 'I' in which a person refers to himself without application of any criteria of his identity the *use of 'I' as subject*⁶⁶, and I will follow him in this usage. In the following two chapters I shall try to show that it is this ability of ours that gives rise to the idea that everyone of us has a 'special access' to his identity over time, and that this special access reveals that personal identity is -- unlike object-identity -- sharp and determinate, such that the question 'Was this person, or will this person be, *me*?' must *always*, i.e. under *any* imaginable circumstances, have a determinate answer. I shall also try to demonstrate that, and how, this

⁶³ All quotations from Reid, *Essays on the Intellectual Powers of Man*, p.206.

⁶⁴ See Bishop Butler, 'Of Personal Identity', *The Works of Bishop Butler*, vol.II, p.281.

⁶⁵ Cf. Chisholm, *Person and Object*, p.92.

⁶⁶ Cf. Wittgenstein, *The Blue and Brown Books*, pp. 66-69.

ability, and the misinterpretations that its peculiar epistemic features give rise to, influences our reactions to various thought-experiments concerning personal identity, especially those involving putative 'body-changes'.

Before that, however, I should here like to invite the reader to explore the 'phenomenology' of this intuition a bit further. At least under normal circumstances, a person's ability to refer to himself without application of any criteria of his identity extends not only to the immediate presence, but also over to his past and future. Let me express it, as is most appropriate, in the first person: I do not normally apply any criteria of my identity in judging whether what I now think to be experienced by me is in fact experienced by *me*. Nor do I do so in judging whether what I now remember as having been experienced by *me* was in fact experienced by the person I am now. And nor do I do so in judging whether what I now anticipate as being experienced by *me* will in fact be experienced by the person I am now. This idea of a person's 'special access' to his identity is, I think, what underlies the intuition that a person must -- in all imaginable circumstances -- be able to fix his identity for all the past he can remember, and all the future he can anticipate, and that he can do this simply by entertaining the thought of *himself*, i.e. simply by thinking 'I'. This conclusion seems to be warranted by the very phenomenology of this particular kind of self-reference. The awareness of oneself seems indeed to have all the characteristics of an immediate intuition. It seems to be a pure, conceptually unmediated act of consciousness in which the person who performs it can grasp his very core or essence, which, although it cannot be further spelt out or explicated, appears to determine for each person at any given time which entity was or will be him. The reader is here invited to attend to his own subjective experience: Does it not certainly *seem* to be true that my now thinking 'I' provides me with all I ever need in judging whether any given entity (about which I know all the relevant facts) was or will be me? Does my 'I'-thought not at least appear to have grasped something simple, something that determines my identity for all past and all future? The intuition which I am trying to bring out has been particularly well expressed by Nagel, when he says:

'When someone poses inwardly the question whether a past or future experience will be *his*, he has the sensation of picking out something whose identity over time is well defined, just by concentrating on his present experience and specifying the temporal extension of *its subject*'.⁶⁷

Part of the aim of the next two chapters is to show that this 'sensation' is an

⁶⁷Nagel, 'Subjective and Objective', p. 200.

illusion, and how it arises from certain epistemic idiosyncracies of a person's ability to refer to himself in the 'as-subject manner'. In the course of this discussion it will also appear that a correct account of a person's ability to refer to himself without the application of any criteria of identity whatsoever is essential to a correct understanding of certain kinds of psychological capacities that are manifestations of what Peacocke has called the 'first-person mode of presentation'⁶⁸ -- capacities which are significantly invoked in thought-experiments concerning personal identity. I shall conclude that, although a person's ability to refer to himself without application of any criteria of his identity may be a necessary ingredient of personhood, it cannot provide a person with a 'special access' to what his identity consists in. Nor can it be invoked in support of the view that questions of personal identity must always have a determinate answer.

⁶⁸ Peacocke *Sense and Content*, passim.

3. THE EPISTEMOLOGY OF SELF-REFERENCE

3.1. Some Characteristic Features of Self-Reference

Wittgenstein introduces the notion of a use of 'I' as-subject (the paradigm being 'I have a toothache') by contrasting it against the corresponding use of 'I' as object (the paradigm being 'I have grown six inches')¹. Whereas, according to Wittgenstein, the use of 'I' as-object 'involves the recognition of a particular person'² and, as a consequence of this, allows for the possibility of an error through misrecognition, this is not so in the use of 'I' as-subject. In the use of 'I' as-subject an error as to the recognition of the intended referent of 'I' is logically excluded, since, as Wittgenstein puts it, 'there is no question of recognizing a person when I say I have toothache.'³ In the *Blue Book* Wittgenstein concluded from this that 'I' in its as-subject use does not serve to distinguish one particular person from all others⁴. Later, in the *Philosophical Investigations*, he clarified and substantiated this point by claiming that, since 'I' does not draw on any criteria of identity of a person, it cannot be said to identify a particular person⁵. We shall return to this thesis below.

Adopting -- but slightly modifying -- Wittgenstein's usage of the terms 'as subject' and 'as object', I shall speak of a person *referring to himself-as-subject* (or as-object), thereby meaning that he uses 'I' in its as-subject use (or, respectively, in its as-object use) in order to refer to himself. I should want to make it plain right at the beginning -- although this will become an important topic later on (see 3.3.) -- that this distinction is drawn in terms of two different *ways of referring*, not in terms of two different *objects of reference*. Furthermore, in adopting Wittgenstein's distinction I do not intend to commit myself to any thesis as to what range of predicates the as-subject use of 'I' is supposed to cover, whether, e.g., the impossibility of misidentification in terms of which Wittgenstein draws the distinction covers only self-ascriptions of (a certain subset of) psychological properties, as Wittgenstein seems to intimate, or whether it also extends to self-ascriptions of physical properties, as Evans and other interpreters

¹ Cf. Wittgenstein, *The Blue and Brown Books*; pp. 66-69.

² Op.cit. p.67.

³ Op.cit. p.67.

⁴ Op.cit., p.68/9. Cf. Wittgenstein, *Philosophical Investigations*, §405/6.

⁵ Cf. Wittgenstein, *Philosophical Investigations* §§ 404-7.

of Wittgenstein have argued⁶. (A sceptical reader may view the use of 'I' as-subject as confined to statements of the form 'I think that p'). Nor will it be within my brief to elucidate the -- rather complex -- relations between the two ways of referring to oneself. My investigation of the epistemology of a person's reference to himself-as-subject (or a person's *self-reference*, as I shall also call it more briefly) will rather be confined to those of its aspects which may -- explicitly or surreptitiously -- give rise to the our 'natural belief' in the absolute simplicity and determinacy of personal identity over time. In his later writings Wittgenstein seems to have abandoned the distinction between a use of 'I' as-subject and as-object, and that, I suspect, for good reasons. In the light of his later philosophy, the distinction between a use of 'I' as subject and as object must have appeared as a residue of Cartesian dualism. It is indeed not evident that, as Wittgenstein seems to have thought up to the *Blue Book*, 'I' in its as-object use can always be replaced by 'this body' or 'the body which is mine' (where 'mine' is an example of the as-subject use). Nor is it obvious that the use of 'I' as object, as in 'I have grown six inches', always requires a recognition or identification of a particular person⁷. Nevertheless, I think that for the purpose of elucidating the, as it were, 'Cartesian' characteristics of self-reference which give rise to various paralogistic illusions, the distinction between the two uses of 'I', whatever its general viability, seems to me a very useful one. There are basically three such, closely interrelated, characteristic features of a person's reference to himself-as-subject, which any account of self-reference has at least to comply with and, if possible, to explain:

The first is the one invoked by Wittgenstein in drawing the distinction between the two different uses of 'I'. Wittgenstein's remarks to the effect that 'I' (in its as-subject use) does not 'provide for' the possibility of misrecognition of the person using 'I' have been restated, and further elaborated, by Shoemaker as the thesis of the '*immunity to error through misidentification relative to the first-person pronouns*' whenever 'I' is used as subject (and that is, according to Shoemaker, in first-person present-tense

⁶ Cf. Evans, *The Varieties of Reference*, pp.218-220, and Malcolm, 'Whether "I" is a Referring Expression', *passim*.

⁷Cf. Hacker, *Wittgenstein, Meaning and Mind*, pp. 481-492.

psychological statements)⁸. According to Shoemaker's definition of an error through misidentification this thesis implies that in self-reference the following mistake is excluded: A person cannot know some particular thing to be ϕ , but make the mistake of asserting 'I am ϕ ' because he erroneously thinks that the thing he knows to be ϕ is he himself⁹. A statement of the form 'I am ϕ ' (where ϕ is a predicate in connection with which 'I' is used as-subject) cannot be split up into a predication-component 'x is ϕ ' and an identification-component 'x=I'; it does not provide room for such an identification.

The second striking peculiarity of self-reference is what has been called its *essential indexicality*¹⁰. This feature was first discovered by H.N. Castaneda in his seminal essay 'He: A Study in the Logic of Self-Consciousness'¹¹, and has since then become a standard point in any account of self-reference. Castaneda's original discovery was that in ascriptions of self-consciousness to a person in sentences of the form 'A believes that *he* is ϕ ' we cannot replace the pronoun 'He' by any other singular term S (be it a definite description or any other indexical expression) *salva veritate*. For we would thereby open up the possibility that the person A to whom the self-referential belief 'I am ϕ ' is ascribed in our original sentence, may in fact not have the self-referential belief that we intend to ascribe to him by claiming 'A believes that S is ϕ ', -- which would be the case if A does not know that *he* is S. Castaneda made this point in terms of statements ascribing self-referential beliefs to others, and thus, as it were, from the third-person perspective. But this was, of course, only an indirect way of stating the (by now generally acknowledged) *irreducibility* of indexical to any kind of descriptive thought, i.e. the fact that in an expression of a self-referential belief in a sentence of the form 'I am ϕ ' the pronoun 'I' cannot be substituted *salvo sensu* by any other singular

⁸ Shoemaker, 'Self-Reference and Self-Awareness', esp. p.556. The immunity-thesis is also often stated as the 'immunity of self-reference against reference-failure' ; cf. Anscombe's well-known essay 'The First Person', p. 55.

⁹ Cf. Shoemaker, 'Self-Reference and Self-Awareness', p. 556/7.

¹⁰ See, e.g., Perry 'The Problem of the Essential Indexical'; or, alternatively, Chisholm, *The First Person*, chs. 2-4.

¹¹ See also Castaneda's 'Indicators and Quasi-Indicators', and 'On the Logic of Attributions of Self-Knowledge to Others'. Castaneda's discovery was in a way anticipated by P.T. Geach's 'On Beliefs about Oneself'.

term. In referring to himself-as-subject a person does not refer to himself under any particular description. He refers to himself simply *as himself*.

Another striking feature of self-referential beliefs is their particular *sensitivity to action and emotion*¹², i.e. their distinctive psychological role in motivating action as well as in influencing practical reasoning. Perry's belief that the shopper with the torn sack was making a mess did not cause him to rearrange the sack in his cart until he finally found out that *he himself* was the shopper with the torn sack¹³. Similarly, what the victim of Williams' mad scientist¹⁴ is primarily afraid of is that *he* will be tortured, not that some particular person A (which is in fact himself) will be tortured. (A might have forgotten that he is A.) A would have no reason to be distraught by the news that A will be tortured unless he believed that *he is A*¹⁵.

3.2. *The Distinction between Semantic and Intentional Self-Reference*

I think that much of the discussion on the nature of self-reference has suffered from a failure to determine the sense of 'reference' invoked in our speaking of 'self-reference'. This failure is in its turn due to a failure to distinguish between two senses, or perhaps rather two connotative components, of the term 'reference' in general. These are the *semantic* sense of 'reference', on the one hand, and the *intentional* sense, on the other.

The *semantic* sense of 'reference' is invoked whenever we speak of a *term*, typically a singular term, referring to a given entity. The semantic reference of a given term is typically determined by the term's *linguistic meaning* (i.e. the rule governing its use), and the *context* (i.e. the circumstances) of its utterance. It is these two factors, linguistic meaning and context of use, that a theory of semantic reference typically

¹²Cf. Perry, 'The Problem of the Essential Indexical', *passim*. Also, Rovane, 'The Epistemology of First Person Reference', (esp part 4, pp. 161-165).

¹³Cf. Perry, 'The Problem of the Essential Indexical', p.3.

¹⁴Cf. Williams, 'The Self and the Future', *passim*; cf. section 5.3. below.

¹⁵ The particular sensitivity to action and emotion applies to indexical beliefs in general: My belief that Oxford will be destroyed by a meteorite at the 9th of February 1989, at 10 pm, will not frighten me, and will not cause me to leave Oxford, unless I believe that this is *today's* date and that Oxford is *here*, i.e. where I am.

draws on. A theory of semantic reference views language as a logical corpus to be mapped onto the world. It is interested in questions such as how certain sentence-parts refer to certain parts of the world and which contribution they thereby make to the truth-condition of the whole sentence of which they form a part. (Singular terms, such as definite descriptions and indexicals, typically do so by denoting -- in a given context -- a certain *object*). It is from this perspective, and from this perspective only, that a theory of semantic reference (such as a truth-conditional semantics for formalized languages) deals with the phenomenon of human sign-use.

The *intentional* sense of 'reference' is invoked whenever we speak of a *person* - - not a term -- 'referring to' (in the sense of 'meaning' or 'thinking of') a certain object. Reference in this sense is not equivalent, nor -- at least not trivially -- reducible, to semantic reference. The relation in which a person stands to the object of his reference (if it is a relation) is of a different theoretical status from the relation in which the term he uses stands to its referent. The first relation is a cognitive or epistemological one, the second a semantic relation. Intentional reference -- although it is typically achieved by a person's using certain signs with a determinate linguistic meaning, and although it can often *only* be achieved by using signs -- is not (trivially) reducible to semantic reference.

In distinguishing between the semantic and the intentional aspect of a person's reference to a certain object, I do not want to deny that both of these aspects are systematically related to each other. To put it in a nutshell: what a person means by using a certain sequence of words is in general determined by what these words mean (plus, in some cases, the context of their use). And what a word means is in its turn dependent on what people normally mean, i.e. intend to convey, by using it. Thus semantic and intentional reference normally go hand in hand. Sometimes, however, they drift apart. This happens whenever a person draws the wrong linguistic register, i.e. when what he 'objectively' says (according to the rules determining the linguistic meaning of the words he uses) does not match with what he means, i.e. what he intends to convey by saying what he says. Then we will have to draw a distinction between what the person says and what he means (although, for sure, *what* he means, i.e. what nowadays is called the '*content*' of his thought, will again have to be expressible in one language or another).

In variation of an aphorism of Wittgenstein's¹⁶, we might represent the relation between intentionality and language in the following manner: neither is intentionality the vehicle of language nor is language (always) the vehicle of intentionality, but each of them draws the other along. Just like the blind man and the lame man, they can make progress only together. This mutual dependence of semantic and intentional reference, however, cannot, and should not, dispense us from the task of discriminating between the problems to which each of them respectively gives rise. The question of what (which rule in which context) makes a given word refer to a given object *O* is different from, and has a different theoretical status from, the question of what makes a person's intentional reference to *O* a reference to *O* -- and to no other object. It is this latter question that Wittgenstein puts forward in its full generality when in the *Philosophical Investigations* he asks epigrammatically: 'What makes my idea of him an idea of *him*?'¹⁷.

Perhaps the distinction between semantic and intentional reference can gain some further clarity by being put into an historical perspective. It was, I think, Strawson who first introduced the term 'refer' in its explicitly intentional sense into today's philosophical Esperanto, and clearly distinguished between matters of intentional and semantic reference¹⁸. Strawson, however, approached the problems of intentional reference within the framework of a communication-theoretical account of our reference to physical objects in space and time. He was therefore mainly concerned with the question of how a hearer can come to 'identify', or 'single out', or 'pick out'¹⁹ (these being Strawson's favourite terms) a certain object *as the one meant by the speaker*, who refers to the object by using a singular term. And he did not address the even more fundamental, question of how speaker and hearer can identify, or 'single out', the respective objects *for themselves*. (Obviously this question cannot simply be regarded as a borderline-case of the speaker-hearer model of communication, viz. as the case in which the hearer is identical with the speaker. In thinking of an object (by using a referring expression) I (as a kind of speaker-thinker) do not bring myself (as a kind of hearer-thinker)

¹⁶ Cf. Wittgenstein *Philosophical Investigations*, § 329.

¹⁷ Wittgenstein, *Philosophical Investigations*, p. 177. *Pace* Anscombe, I have changed her translation of 'Vorstellung' as 'image' to 'idea'.

¹⁸ Cf. his critique of Russell in 'On Referring', *passim*. Strawson then elaborated his views in ch.1 of *Individuals*.

¹⁹ Cf. e.g., 'Singular Terms and Predication', p. 63.

to think of the same object of which I intend to think as speaker-thinker. In this way I could never think of any object at all. In thinking of an object I do not use a referring expression with the intention of drawing *my* attention to a particular object to which the referring expression -- in my 'thinking-situation' -- uniquely applies. I think of the object as the thing which is given to me in a certain way, a way manifest in the referring expression I use.)

The question what conditions a thinker has to fulfil in order to identify an object *for himself* has recently been tackled by Gareth Evans. In essays and in his book *The Varieties of Reference* Evans has argued that in order to give a satisfactory theory of our reference to particular objects we do not only have to take into account the semantic relation in which a sign-user stands to an object denoted by the signs he uses, but also the 'intentional', and thus epistemic, relation in which he stands to this object²⁰. Evans maintains that in order to be able to refer to a certain object a thinker has to meet certain epistemological requirements. A thinker cannot simply refer to an object *sans phrase*, i.e. as whatever is denoted by the singular term he uses. In order fully to understand a singular term referring to a certain object, a thinker does not only have to think of the right object, but he also has to think of it in a certain way, -- a way which mediates between the thinker and the object of his thought and thus establishes the intentional relation between them. Put differently: the object of a thinker's thought has to be 'given', or 'presented', to him in one way or another. In thinking of an object a thinker has to think of it *as this or that*; he has to think of it '*under a certain mode of presentation*'. And under which 'mode of presentation', i.e. in which way, he thinks of it manifests itself in the singular term the thinker uses in referring to the object of his thought²¹.

As a paradigm of a theoretical term framed to account for different ways of thinking of an object, Evans revives the Fregean notion of *sense* ('Sinn'). In giving the sense of an expression used by a thinker in referring to an object we give one possible answer to the question of, in Evans' words, 'what makes it the case that a thought is about the object which it is about'²². It is of crucial importance here not to confuse the Fregean notion of sense with the only superficially similar notion of linguistic meaning,

²⁰ See Evans, *The Varieties of Reference*, esp. ch.1.

²¹ Cf. Evans, *Varieties of Reference*, pp. 25/6.

²² G.Evans 'Understanding Demonstratives' on p. 294.

especially since Frege's talk of 'grasping the sense' of an expression lends itself too easily to the common misinterpretation according to which this grasping consists simply in knowing the rules which govern the use of an expression. The profound differences between the Fregean notion of sense and the various modern notions of linguistic meaning have only recently been brought to light by the combined interpretive efforts of Evans, Dummett, Perry, and Burge²³. In the present context I want to point out only the two most significant differences: Firstly, Frege's notion of sense is -- as opposed to the notion of linguistic meaning -- an epistemological notion, introduced mainly with the purpose of accounting for the differences in cognitive value ('Erkenntniswert') of statements (especially identity-statements) with the same truth-value. And, second, a Fregean sense is supposed to determine its reference ('Der Sinn bestimmt die Bedeutung.'²⁴) by showing the thinker who grasps it the way to the referent.

One may have misgivings about the general viability of the Fregean (and Evansian) enterprise. One may, that is to say, have doubts that there are any *general* conditions to be found which a person has to fulfil in order to be able to refer to an object. Wittgenstein's answer to the question what makes my idea of A an idea of A would have been that typically, if I were asked, I should *say* so (or *manifest* my meaning A rather than B in some other behavioural way). And he would have maintained that over and above these behavioural criteria for meaning A rather than B, no general requirements for a thinker's meaning, or referring to, a given object can be extracted into a theory. Equally, one may doubt that the way in which we use words in order to single out objects (be it in thought or utterance) can be adequately explained in terms of the Fregean notion of sense. However, I think that Frege's distinction between sense and reference, whatever its merits in other fields of philosophical logic, can be put to good use in illuminating the epistemological peculiarities of self-reference and the 'paralogistic' illusions which I think these peculiarities give rise to. And even if we finally (as I shall suggest we should) arrive at the conclusion that the Fregean approach was misguided in the first place, we shall, I hope, have learnt something on the way to this conclusion.

²³ See Evans, 'Understanding Demonstratives', esp. pp. 293-203; Burge, 'Sinning against Frege'; and Perry, 'Frege on Demonstratives'. Cf. also Künne, 'Sinn, Indexikalität, und propositionaler Gehalt',.

²⁴Cf. Frege, 'On Sense and Reference'.

This brief characterization of a Fregean account of intentional reference in general may suffice to give us a rough idea of what such an account of intentional *self*-reference should consist in and which theoretical tools it might make use of. The questions it will have to ask are the following: What makes it the case that a person's reference to himself-as-subject is a reference to *himself*? What, in other words, is the particular mode of presentation (the '[self]-mode', as it has been called²⁵) under which a person thinks of himself when he thinks of himself-as-subject? What is the sense of 'I' when used as-subject? In which way does a person think of himself when he uses 'I' in order to refer to himself-as-subject? What constitutes the 'intentional content' of such a self-reference?

These are the questions we will have to answer (in sect.3.3.). For it seems to me quite obvious that the question of self-reference can only be regarded as a question of *intentional* self-reference. It is the intentional connotation component of 'reference' that is invoked whenever we speak of a person referring to himself-as-subject and understand this self-reference as a manifestation of what has variously been called self-consciousness, self-knowledge, or self-awareness. It is significant that viewed from a purely semantic perspective, the phenomenon of self-reference does not pose any problems at all. The linguistic meaning of 'I', i.e. the rule according to which the pronoun 'I' typically achieves its semantic reference, can in fact be quite easily explicated by stating that a given occurrence of 'I' (be it in thought or utterance) refers to whoever (or whatever) produces it. As such the linguistic meaning of the pronoun 'I' stays constant for all particular occurrences of 'I', although its referent varies from person to person. The context-dependence which manifests itself here is the most significant feature of the pronoun 'I' from a purely semantic viewpoint, and it is also the one that it shares with all other indexical expressions.

Understood as an analysis of the semantic aspect of self-reference Reichenbach's well-known token-reflexive analysis of the linguistic meaning of 'I'²⁶ was thus perfectly adequate. That a token of 'I' refers to whoever utters it is indeed all there is to say about self-reference from the semantic point of view. It represents very clearly which contribution an utterance of 'I' makes to the truth-condition of sentences in which it occurs. For, that (a given token of) 'I' refers to whoever utters 'I' (or thinks of himself

²⁵ Cf. Peacocke *Sense and Content*, p.108.

²⁶ Reichenbach, *Elements of Symbolic Logic*, pp.284-287.

by using 'I' in soliloquy) is the simple reason for the truth that 'I am ϕ ' as uttered or thought by me is true if and only if 'RS is ϕ ' is true as uttered or thought by you. For, in the given situation, 'I' and 'RS' refer to, or denote, one and the same entity²⁷. That is all that can be said about this matter, and it is not anything exciting.

We have to take care in this context not to confuse questions of semantic with questions of intentional reference. This is just what seems to me slightly misleading about the claim -- made in the train of some remarks of Wittgenstein's by Malcolm and Anscombe -- to the effect that in its as-subject-use 'I' is not a referring expression²⁸. Thus Anscombe has claimed that 'I' is 'not an expression whose logical role is to make a reference *at all*'²⁹. If it were, she argues, Descartes would have been right in his conclusion that 'I' when used as-subject stands for a Cartesian Ego. Unfortunately, Anscombe never makes quite clear what she takes to be the criteria for a term's being, or functioning as, a referring expression. Sometimes she seems to require that a referring expression has to 'latch onto'³⁰ something, and she argues that this is not the case of 'I' when used as-subject. This claim, I shall argue below, is true: in its as-subject use 'I' is not used to single out one object from a manifold of possible objects of reference (cf. sect.3.4). However, it seems potentially misleading to try to make *this* point by denying that 'I' is a referring expression. It is just the referring role of 'I' that enables you to understand whom I mean whenever I use 'I' in conversation. I mean *myself*, and if you know who I am, you know to whom I am thus referring, namely the particular entity I am, RS. From a purely semantic point of view, 'I' is clearly a referring expression. It 'stands for' an object just like a proper name (although, of course, it is not *used* just like a proper name, or the indexical 'this').

²⁷ Shoemaker refers to this trivial principle under the title of the 'logical correspondence between first-person and third-person statements' in *Self-Knowledge and Self-Identity*, (p.170). Tugendhat in *Selbstbewußtsein und Selbstbestimmung*, calls it 'veritative Symmetrie' (p. 89).

²⁸ Anscombe 'The First Person', p. 60. Cf. Malcolm 'Whether "I" is a Referring Expression' (passim). (Malcolm in fact claims that 'I' is *never* a referring expression). Cf. Wittgenstein *The Blue and Brown Books*, p. 67, where Wittgenstein counters the interlocutor's claim that 'surely the word "I" in the mouth of a man who says it; it points to himself; and very often a man who says it actually points to himself with his finger' by claiming that 'it was quite superfluous to point to himself'.

²⁹ Anscombe, 'The First Person', p. 60.

³⁰Op.cit., p.54.

The superficiality of the purely semantic perspective on self-reference is also evident from the fact that the token-reflexive analysis of the linguistic meaning of 'I' is completely insensitive to the distinction between the as-subject and as-object use of 'I'. And this is of course due to its very nature. For the distinction between two different uses of 'I' is a distinction in terms of *how* a person uses a sign in order to refer to himself -- not a distinction in terms of how the referent of a sign at a given occasion is determined by its linguistic meaning. As such this distinction must remain unnoticed by a purely semantic approach to self-reference.

More significantly, the adequacy of the token-reflexive analysis of the meaning of 'I' is also completely independent of the ontological question what the semantic referents of 'I' are. We may grant Strawson that he has conclusively shown that (at least in our conceptual scheme) persons, i.e. living embodied beings to whom both physical as well as mental attributes can be ascribed, are the ultimate referents of uses of 'I'³¹. Alternatively, we may view it as an empirical question what the referents of our 'I'-uses are. But the correctness of the token-reflexive rule is independent of these divergences of opinion. Even if one day we started to conceive of ourselves as monads, or brains in a vat, it would have lost nothing of its truth.

3.3. Does 'I' Have a Sense?

The serious, and perhaps insuperable problems that indexical expressions in general, and the indexical 'I' in particular, pose for a Fregean theory of thought and reference are by now well-known³². That such pessimism is not quite ungrounded can be gathered from a brief look at the requirements a potential sense of 'I' would have to fulfil in order to perform its task of making it the case that a thought about myself-as-subject is a thought about *myself*, i.e. the particular entity that I am: It would have to be such that (a) it uniquely determines me as its reference ('Bedeutung') whenever I grasp it, and that (b) by grasping it I am infallibly aware that it determines me as its reference (since otherwise the characteristic immunity to misidentification, which in its turn is intimately related to the other epistemological features cited above, could not be accounted for). It is not easy to see how a Fregean sense could possibly comply with both of these requirements. And it seems that Frege himself had some initial awareness

³¹ See Strawson, *Individuals*, ch. 3; and also *The Bounds of Sense*, pp. 168/9.

³²Cf. e.g. Perry, 'Frege on Demonstratives', pp. 478-485.

that the pronoun 'I', just as the other indexicals, may prove particularly recalcitrant to the standard Fregean analysis in terms of sense determining reference. In a famous passage dealing with the mode of presentation of 'I', a passage which recently has formed the topic of a lively discussion³³, Frege appears to take an essentialist way-out of the predicament created for him by indexicals. He claims that

'... everyone is presented to himself in a particular and primitive ('ursprünglichen') way, in which he is presented to no-one else. So, when Dr. Lauben thinks that he has been wounded, he will probably take as a basis this primitive way in which he is presented to himself. And only Dr. Lauben himself can grasp thoughts determined in this way.'³⁴.

The conclusion Frege seems to draw in this passage is that the sense each of us associates with his use of 'I' is primitive, and thence unanalysable and incommunicable to anyone else. It is interesting to note that this analysis of self-reference (at least in its presently suggested interpretation) is very similar to the one once propounded by Husserl in his *Logical Investigations*. There Husserl claimed that

'...the word "I" names a different person from case to case, and does so by way of an ever-altering meaning'.

Husserl concluded from this that:

'Each man has his own "I"-presentation (and with it his individual notion of "I"), and this is why the word's meaning differs from person to person'.³⁵

Such an 'individualist' conception of the nature of self-reference, which is only vaguely suggested in Frege's and Husserl's brief remarks, has in recent years been elaborated and clarified by Roderick Chisholm in his book *Person and Object*³⁶. According to Chisholm, every single person has an individual property that necessarily and uniquely individuates him, and that he furthermore necessarily knows to individuate only himself (cf. pp.31-33). Chisholm calls it a person's 'individual essence'(p.33). It is this individual essence, Chisholm claims, that forms the sense or intentional content of a person's self-reference and thus determines him as its referent, or 'Bedeutung' (cf. p.36).

³³ Perry, 'Frege on Demonstratives'; McGinn, *The Subjective View*, ch.5; Chisholm, *Person and Object*, ch. 1, esp. note 20, p. 199; Evans, 'Understanding Demonstratives', pp. 312-315.

³⁴ Frege, 'The Thought. A Logical Inquiry', on pp. 25-26.

³⁵ Both quotations from Husserl, *Logical Investigations*, pp.315/6.

³⁶ Chisholm *Person and Object*, pp. 27-41. References in text are to this work.

The only way of expressing this individual essence is by 'being I' or 'being identical with me' (p.33). Thus, to illustrate Chisholm's claim, for each person the sense of 'I have headaches' can be rendered as 'The person who is (i.e. has the essence of being) identical with *me* has headaches'. Chisholm acknowledges it as a logical consequence of his account that there are just as many individual essences as there are persons. And in one of the subsequent chapters he concludes from this conception of the nature of self-reference that a person must have an intuitive knowledge of what constitutes his identity over time (it is determined simply by his 'being identical with himself', i.e. his individual essence which cannot be further spelt out), and that therefore questions of the kind 'Will this person be me?' must always have a determinate answer -- although we may not always be able to know it (cf. pp.104-113).

It seems clear that Chisholm did not intend his account of self-reference to be understood as an attempt to *analyse* the -- according to Frege -- primitive, and thence unanalysable, mode of presentation of the referent of 'I' into more primitive elements, since it is quite obvious that the ability of grasping the sense of 'me' in Chisholm's *analysans* presupposes the ability of grasping the sense of 'I' in the *analysandum*. The sense of 'being identical with me' does not spell out or analyse the sense of 'I'; it only renders its logical characteristics more perspicuous. Chisholm's account of the sense of 'I' can thus serve to bring to light the advantages and disadvantages of the Fregean approach more clearly.

And at first thought the theoretical advantages of the Chisholm-Frege approach seem indeed to be considerable. Firstly, 'The person who is identical with me' is a singular term whose sense necessarily individuates the one who grasps it. Secondly, whoever refers to himself as 'the one who is identical with me' must do so infallibly. For, if he did not know infallibly that he refers to *himself* via this sense, he could not even grasp it. Thirdly, the Chisholm-Frege account thus harmonizes very well with the most conspicuous features of self-reference. By building the irreducibility of the 'I'-mode of presentation to any other mode of presentation (which may be held to be the reason for the above-mentioned irreducibility of 'I'-thoughts to any other thoughts) into the very sense of 'I', it may even give the appearance of providing some sort of an explanation of the 'essential indexicality' of the first person pronoun.

However, Chisholm's proposal has to pay a heavy cost for the advantages it offers. (And it seems to me that the very fact that he has to pay this price should be taken as evidence that even the advantages are obtained by some sort of logical trick-

ery). Its acceptance would force us to subscribe to a sort of solipsism regarding all 'I'-judgements. For, according to the Chisholm-Frege account, you and I could never share the same 'I'-thoughts. I could never grasp the sense of your 'I'-judgements just as you could never grasp the sense of mine. Considering that part of the very purpose of Frege's introduction of senses populating a Platonic Third Realm was to account for the possibility of different people sharing, and thus communicating, thoughts, this seems an unacceptable consequence. For, if we followed Chisholm, we would have to commit our ontology to the existence of 'private' incommunicable senses, i.e. senses which can only be grasped by the person whose 'individual essence', or 'primitive mode of presentation', they constitute. And this is to say, we would have to accept the existence of private propositions, i.e. Fregean thoughts which could be thought, and thus had, by one person only. This seems to me a highly implausible assumption. It is a simple and undeniable fact that we do understand each other's 'I'-utterances and 'I'-thoughts quite perfectly. And it seems equally evident that, in the sense in which the two of us can have the same thought at all, the thought that I would express by 'I am ϕ ' can be just the same as a thought that you would express by 'I am ϕ '. If a theory comes to contradict the simple 'phenomenological facts' it set out to explain, it is the theory, not the facts, which have to be blamed.

An alternative reading of the passage of Frege's quoted above has recently been presented by McGinn in his *The Subjective View*³⁷. He argues that we are not forced to conceive Frege's very brief and inexplicit remark as claiming that each one of us is presented to himself in a special mode of presentation which is different from the mode of presentation in which any other person is presented to himself. Frege's remark, McGinn points out, could also be read as claiming that there is a single mode of presentation in which everyone is presented to himself. McGinn thus distinguishes two very different readings of Frege's brief remarks. Let x be a person and m a mode of presentation, then Frege's remark could either mean:

$$\forall x \exists m (x \text{ is presented with } x \text{ in } m \ \& \ \sim \Diamond \exists y (y \text{ is presented with } x \text{ in } m \ \& \ y \neq x))$$

Or it could mean:

$$\exists m \forall x (x \text{ is presented with } x \text{ in } m \ \& \ \sim \Diamond \exists y (y \text{ is presented with } x \text{ in } m \ \& \ y \neq x)) \quad 38,$$

³⁷Cf. McGinn, *The Subjective View*, ch. 5.

³⁸Op.cit., p. 67.

The second version, i.e. the one favoured by McGinn, has the advantage over the first that it avoids committing us to the thesis that 'I' has a private sense for each one of us. The sense of 'I' (and thence the 'conceptual content', as McGinn says³⁹, of each person's reference to himself) is the same for all of us. In other words, it is not the cognitive aspect of a person's self-reference that makes it the case that in using 'I' a thinker thinks about himself. It is not the mode of presentation that individuates the referent of 'I', but, as McGinn puts it (with reference to indexicals in general):

'The context of thought or utterance is what ties the indexical mode of utterance down to particular things, not the concepts in the mind of the thinker; so these latter can stay constant while what is thought about varies.'⁴⁰

Applied to self-reference this thesis means: What makes a person's reference to himself a reference to *himself* is not an internal, intentional feature of this reference. The individuation of the referent of 'I' does not take place 'from within', i.e. for the thinker himself. For, the identity of the self-referring person does not in any way enter into the conceptual content of his self-referential belief. The individuation of the object of self-reference rather takes place from a third-person perspective. That is to say: which entity is being referred to in an act of self-reference is determined, not by a conceptual feature of the act of reference, but simply by the external circumstance that it is the given entity which is in fact performing the given act of self-reference. Using the terminology introduced above, we might say that the determination of the referent of 'I' is here purely a matter of semantic reference, namely of the linguistic meaning of 'I' according to which 'I' refers to whoever uses it. (Note that when McGinn claims that 'what is thought about varies' he then draws on the semantic connotation component of 'to think about'). The sense of 'I' does not contribute to the individuation of its referent.

This analysis has of course the consequence -- and McGinn is well aware of this⁴¹ -- that in the special case of 'I' sense does not determine reference. McGinn seems to think that this violation of one of the principles of Frege's thinking on sense and reference seems to be a small price to pay for an account of self-reference that does not commit us to the existence of private 'I'-senses and 'I'-propositions, and thus to the absurd conclusion that we cannot understand each other's 'I'-thoughts. And in this we

³⁹Op.cit., p.66.

⁴⁰Op.cit., p. 67.

⁴¹ Cf. op.cit., p.68.

may agree with him. At the same time, however, we may want to ask what theoretical function a Fregean sense which does not show us the way to its reference could possibly be said to fulfil. We may want to ask, that is, what role a sense which does not in any way single out its referent could possibly be said to play in an account of 'what makes' a thinker's self-referential belief a belief about himself. According to McGinn's interpretation of Frege's remark, the sense of 'I' remains a loose-running wheel. Nothing turns on it.

Perhaps this objection to McGinn's account could even be strengthened by pointing out that McGinn does not really succeed in explaining a thinker's reference to, or thought of, *himself*. For, in order to explain conscious self-reference it is not enough for a Fregean account simply to accept that the intentional object of this reference (namely the entity that I am) be determined only 'from outside', namely by dint of the *linguistic* meaning of 'I' according to which a token-occurrence of 'I' refers to whichever entity utters or thinks it. It also has to explain that it is *myself* that I thus refer to, i.e. that I *mean* myself. In other words, 'what makes' my reference to myself a reference to *myself* must be an *internal*, cognitive feature of this reference. It must be due to the intentional content of this reference -- not merely to 'the context of thought or utterance'. The object of my self-reference has to be determined by its very *content*; for otherwise it would not be *self*-reference.

Perry's version of a basically Fregean account of self-reference tries to comply with this requirement. However, it is interesting to see that it can do so only by violating certain other requirements that such an account would have to fulfil. Similarly to McGinn, Perry contends that it is not a cognitive feature of a person's self-reference by which the object of this reference is singled out⁴². Thus, just as McGinn, he accepts that in order to salvage a basically Fregean theory of thought and reference one has to sacrifice one or more of its principles when it comes to the theory of indexicals. Perry suggests that we conceive of the thought expressed by an indexical sentence of the form 'I am ϕ ' as consisting of the ordered sequence constituted by the object denoted by the indexical 'I' and the 'incomplete sense' of 'is ϕ '⁴³. This suggestion has the consequence that the thinker of an 'I'-thought becomes one of the very constituents of his

⁴²Perry, 'Frege on Demonstratives', *passim*.

⁴³Cf. *op.cit.*, pp.491-496.

thought. A sentence of the form 'I am ϕ ' can then be said to be true iff. the object denoted by 'I' -- and due to what Perry calls the 'role' (p. 479) of 'I' this is the person using it -- falls under the concept which is the Fregean reference of the incomplete sense expressed by 'is ϕ '. Thus Perry can hold that anyone of us can 'apprehend' the thought expressed by 'I am Hume' (namely by apprehending a thought containing Hume and the sense of 'is Hume'), but that only one of us can apprehend this thought by 'entertaining the sense' of 'I am Hume', namely Hume⁴⁴. Of course, anyone else can entertain the sense of 'I am Hume', but by doing so he will express a different, and, that is, a false thought. The privileged access which Hume enjoys with regard to the thought expressed by 'I am Hume' is thus not due to a feature of this thought itself (such as would be its essential privacy), but only to the sense which Hume entertains when he apprehends the thought.

Now Perry can thus certainly avoid a commitment to the existence of essentially private senses and thoughts. Nevertheless I do not think that sacrificing the principle that a thought must be transparent to its thinker is the right reaction to the problems thrown up by self-reference, since it rather obscures than clarifies its epistemological idiosyncrasies. In particular, Perry's proposal to make the concrete person who thinks a self-referential thought 'part' of this thought flies right in the face of the alleged immunity to error through misidentification characteristic of self-reference. It is just characteristic of the use of 'I' as-subject that I do not have to know who (i.e. which worldly entity) I am in order to be able thus to refer to myself. A related difficulty in Perry's account is brought out by his (I think mistaken) assumption that a thinker's grasping the Fregean sense of 'I' consists in his knowledge of the rule which constitutes its linguistic meaning (its 'role', as Perry calls it), namely that any token of 'I' refers to the one who (mentally or publicly) utters it⁴⁵. For this assumption seems to render self-reference impossible: in order to apply this rule at any given moment a thinker would already have to know who it is who applies it. Thus, instead of explaining it, Perry's analysis already seems to presuppose self-consciousness.

Thus, the dilemma that an account of self-reference along basically Fregean lines faces can be summed up in the following manner: *Either* it maintains that the identity

⁴⁴Cf.op.cit. p.492.

⁴⁵Cf. Evans' critique of this assumption in 'Understanding Demonstratives', pp. 320/1.

of the thinker *does not enter* into the 'conceptual content' of his self-referential thought, and then it does not really explain *self-reference*. That is to say, it cannot explain in which sense a person can be said to mean *anything* when using 'I' as-subject. *Or* it maintains that the identity of the thinker *does enter* into his thought, and then it cannot account for the immunity to misidentification which is characteristic of self-referential thought. That is to say, it cannot explain why the thinker is immune to misidentification in referring to himself-as-subject. Either way, it cannot comply with all the requirements for an account of self-reference.

I think what we are witnessing here is a model shaking itself to pieces. The only adequate conclusion to draw is, it seems to me, to simply accept that 'I' used as-subject has *no sense (mode of presentation)* at all, and, that is to say, that there is *no object being presented* in this particular kind of self-reference. And I think this is also how we should understand Anscombe's claim that statements of the form 'I am NN' are not *identity-statements*⁴⁶. On the superficial semantic level statements of the form 'I am NN' may be said to be identity-statements simply because, when RS utters or thinks 'I am RS', then 'I' and 'RS' refer to the same entity and because when you do so, 'I' and 'RS' refer to different entities. And this is all a purely semantic account of self-reference, such as a truth-conditional semantics for a given language, is interested in. However, the statement 'I am RS' is in significant respects very different from a paradigmatic identity-statement such as 'The morning star is the evening star', since it cannot be analysed in terms of two different Fregean senses showing us the way to the same reference. For 'I', when used as-subject, has no sense at all. *Prima facie*, this claim may appear outrageous. But perhaps it can be made to appear somewhat less so if we put the search for the Fregean sense of 'I' into an historical perspective.

I think it is partly the very talk of self-reference (or, alternatively, self-consciousness or self-awareness) that misleads us, and has misled many of our predecessors, in this context. For talking of self-reference suggests that -- since in using 'I' as-subject in order to refer to himself a person does not have to draw on any of his features as the concrete, worldly entity he is -- he gets hold of a very special object: namely *his self*, the very core and essence of his being which is left over after all his inessential features have been subtracted, an entity even the complete amnesiac (who is on top of that deprived of all his sensory input) could still hold onto. (The move from 'myself' to 'my

⁴⁶ Cf. op.cit., p.60.

self' is a swift and hardly noticeable one.) The awareness of this special object is then conceived as being immediate, and the reference to it, in Anscombe's caricaturing words, as 'sure-fire'⁴⁷. Such a view of the nature of self-reference is, although intuitively appealing, deeply flawed. We are here 'held captive by a picture', -- a picture that has indeed permeated and informed most of traditional thinking on the nature of intentional self-reference. It is the picture of a self continuously emitting a kind of radar-beam of intentionality which it sometimes -- in a mysterious act of *reflection* -- re-directs towards itself. This picture of self-reference as a kind of self-reflexivity has informed theories of self-consciousness from Fichte to Husserl, and has only recently been revived by Robert Nozick⁴⁸. Its 'paralogistic' consequences will be further spelt out below, sect.4.1.. What is important in the present context is that the understanding of self-consciousness in terms of the optical model of self-reflexivity and the postulation of an individuating sense of 'I' are just two sides of the same coin. The presumption that the sense of 'I' has to individuate its respective thinker, i.e. that it must be the intentional, cognitive content of his self-reference that makes his reference a reference to *himself*, is intricately bound up with the illusory idea that by thinking 'I' a person gets hold of something that other people cannot get hold of, something essentially private to which only the person himself has access⁴⁹. The two ideas support each other like two bricks in an arch, which, once they are fitted in, cannot be easily pulled out again. In realizing that the way in which a person thinks of himself-as-subject does not draw on any of the features which constitute him as a particular person amongst others, we are tempted to conclude that this *way of thinking* must determine a special referent different from, or at least 'more inward' than, the person performing the act of self-reference -- a special entity reference to which is rendered sure-fire by the 'primitive' sense of 'I'. But the

⁴⁷ Anscombe 'The First Person', p.55.

⁴⁸ Particularly nice examples of this sort of imagery are to be found in Husserl, *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Erstes Buch*, (esp. §§ 36-37 and §§ 45-46), e.g. 'der immanente 'Blick-auf' das Objekt...der aus dem 'Ich' hervorquillt' (p. 65.), meaning something like 'the immanent gaze-at the object...which emanates from the I'. Fichte's attempts at coming to grips with self-consciousness have been summed up by Henrich, 'Self-Consciousness'. (Fichte's works are unfortunately hardly available in English). Cf. also Nozick, *Philosophical Explanations*, pp.71-86.

⁴⁹ Chisholm, for one, is quite explicit about this, cf. his *Person and Object*, chapter 1: 'The Direct Awareness of the Self', pp. 23-46.

temptation to account for the epistemological idiosyncrasies of a person's self-reference in terms of a special referent of 'I' has to be resisted if we want to avoid the fatal epistemological as well as ontological consequences which the postulation of individuated 'I'-senses brings in its train. However, once we have abandoned the idea of a special referent of 'I' as-subject, the postulation of a special sense becomes equally superfluous. And instead of half-heartedly clinging on to the idea that 'I' has a sense which however plays no role whatever, we might as well cut the string and simply say that 'I' has no sense at all.

3.4 *The Matter of Identification in Self-Reference*

In the previous section we have seen that, within a broadly conceived Fregean framework, a person's self-reference by 'I' cannot be understood as achieving its reference by any of its mere cognitive features. It is not the sense of a given use of 'I' that determines which person is the referent of a person's self-reference. In this section I want to address the related, but more general issue of whether self-reference can be said necessarily to involve an *identification*, or *identifying knowledge*, of the self-referring person. The term 'identifying knowledge' can here be understood in two ways: it can be understood in a strong sense as knowledge of *what kind* an entity is of, and it can be understood in a weak sense as '*discriminating knowledge*'⁵⁰, i.e. knowledge of *which entity* (which of all entities) is referred to by a certain singular term.

The question whether a person's self-reference requires, or involves, identifying knowledge in a strong sense, is evidently of utmost relevance to any account of the theoretical relations between self-reference and personal identity. For, if it could be shown that a person's self-reference requires identification in the strong sense, then a correct theory of self-reference could indeed predetermine questions of personal identity. For if, in other words, it could be demonstrated that for a person to be able to refer to himself-as-subject he must have knowledge of whatever constitutes his essence, and thence (given the arguments in sect. 2.1.1.) knowledge of whatever constitutes his identity over time, then the question what personal identity consists in could be answered, as it were, purely from the first-person perspective. If a person, in order to think of himself-as-subject, really had to have knowledge of *what kind of entity* he is (and

⁵⁰ Evans introduces the notion of discriminating knowledge in ch. 4 of *Varieties of Reference* under the name of 'Russell's Principle'.

hence of his criterion of identity), then this would indeed be a decisive point in favour of a Simple View of personal identity. Then, and only then, genuine indeterminacy as to whether a certain future person will be *me* could be apriori excluded in terms of purely epistemological considerations on the nature of self-reference. Attempts to substantiate the thesis that self-reference requires identifying knowledge in the strong sense have in recent years been made, with varying degrees of explicitness, by Chisholm and Maddell⁵¹. Ultimately, of course, this line of thought goes back to Descartes, who famously concluded that, since he could abstract in thought from all his bodily features, he could determine his essence (and hence his criterion of identity) as that of a *res cogitans*. However, the identification of the self as a *res cogitans* is only one possible version of such a Simple View. Others may be less ontologically committed whilst still sticking to its basic premise. I think that all attempts directly to gain any criteria of identity of the self-referring thinker from the very act of self-reference ultimately rest on the idea that the notion of a 'self' (understood as whatever is referred to whenever 'I' is used as-subject) is itself a sortal term which specifies criteria of numerical identity for whatever falls under it. That this idea embodies a paralogistic misconception will be shown in more detail in section 4.1..

For the moment I should want to focus on the weaker, and hence more promising, claim that self-reference necessarily involves *self-identification* in the weak sense, i.e. the claim that in order to refer to himself-as-subject a thinker has to have *discriminating knowledge* of himself, i.e. that he has to know *which of all* entities he is. This weaker claim, if it could be shown to be true, could still be turned into an argument in favour of a Simple View of personal identity. For, it may be argued, for a thinker to be able to identify, or single out, a given object from all others, he has to have at least a vague idea what *kind* of thing the entity in question is, since, arguably, in order to distinguish an object from all others a thinker has to 'conceptualize' the object in some way. He cannot distinguish it as a pure 'this'. I shall try to show that there is no sense in which self-reference can be said to involve identifying knowledge even in this weaker

⁵¹ For Chisholm cf. section 3.3. above.. For Madell cf. his argument to the effect that 'I' is a 'purely referential term' in *The Identity of the Self*, pp.38-46, which is in turn a defense of McTaggart's argument to the same effect in *The Nature of Existence*, vol.2, ch.36. However, although only few theorists explicitly argue for the view that self-reference is necessarily *self-identification*, it seems to be implicitly shared by many other theorists of personal identity; cf. the discussion of Williams' argument against the possibility of conceptually undecidable cases of personal identity below, sect. 5.6..

sense.

What does it mean to say of a thinker that he has identifying knowledge of a particular object? According to Strawson⁵², it means to ascribe a capacity to him, a capacity which can be manifested in various ways: (a) By picking the object out of one's current field of perception (e.g. by pointing to it), or (b) by recognizing it by one of its features when presented with it, or (c) by giving a description of it which identifies it uniquely. What then about self-reference? In which sense, if any, can self-reference be said to presuppose, or involve, identifying knowledge of the self-referring person?

Let us start with the last criterion given by Strawson: A person's self-reference by 'I' certainly does not require knowledge of a *description* uniquely identifying him. This is, as we have seen above, in fact just definitive of the use of 'I' as-subject. For with respect to any description true of me there is at least the logical possibility that I do not *know* that it is true of me. And in that case a reference to myself via this description would not be a reference to myself-as-subject. The only possible candidate for a description which a thinker has necessarily to know to be true of himself is perhaps something like 'the thinker of this thought' or 'the subject of these experiences'. This possibility was in fact once entertained by Russell who claimed that

'... "I", as commonly employed, must stand for a description; it cannot be a proper name in the logical sense, since true proper names can only be conferred on objects with which we are acquainted... We may define "I" as the subject of present experience.'⁵³

However, the demonstrative identification allegedly achieved by saying 'this thought' or 'these experiences' is obviously a pseudo-identification. In 'inwardly pointing' at one of my thoughts or experiences in the solipsistic vein I do not single out these thoughts from a possible manifold. The only thoughts or experiences to which I possibly could thus refer are by definition mine. (Of course, if we lifted the solipsistic veil, i.e. if we opened up the possibility that the description 'the thinker of this thought' could possibly single out someone else's thought, then this criticism would not apply. But at the same time there would be no guarantee that the definite description refers to *me*.) Thoughts and experiences are identificationally dependent on the particulars (i.e. the persons)

⁵²Cf. 'Identifying Reference and Truth-Values', p. 77.

⁵³Russell, *Theory of Knowledge, The 1913 Manuscript*, pp. 36f..

whose thoughts or experiences they are⁵⁴. Hence, far from indirectly singling out their 'subject' or 'owner', the definite descriptions 'the thinker of this thought' or the 'subject of these experiences' already presuppose the previous and independent identifiability of the person *whose* thoughts and experiences they are. Anscombe's claim that the sense of the proposition 'I am this thing here' can be rendered by saying that

'this thing here is the thing, ...of whose action *this* idea of action is an idea, of whose movements *these* ideas of movements are ideas, of whose posture *this* idea of posture is an idea...' ⁵⁵,

seems to me not entirely immune to a criticism along these lines. I shall not attempt such a criticism in the present context. However, I shall take it for granted that self-reference cannot be said to involve a descriptive, or descriptive-cum-demonstrative, identification.

Nor, however, to move on to another of Strawson's criteria cited above, can a person's self-reference be understood as involving a special kind of *perceptual knowledge*. The absurd consequences to which the notion of an 'inner sense' necessarily leads are too well-known to be repeated here⁵⁶. Self-reference is not some sort of mental pointing to an 'intrinsically subjective' entity, called the self, accessible only to the person whose self it is. As Hume famously remarked, there is no entity a thinker could 'stumble on' when he 'enters most intimately into what he calls himself'⁵⁷. And even if he did stumble upon such an entity, his constant perceptual knowledge of this entity could not be invoked to explain self-reference. For how could the thinker identify this entity as *himself*? By what property could he tell his self apart from all other selves? The absurdity of the question is just a manifestation of the underlying absurdity of the perceptual model of self-reference.

This is also the reason why, finally, self-reference cannot be said to involve the recognition of a particular entity by one of its features. What could such a feature be?

⁵⁴Cf. Strawson *Individuals*-argument to the effect that "states, or experiences...owe their identity as particulars to the identity of the person whose states or experiences they are." (p.97.), and Wittgenstein's reflections in private ostensive definition, *Philosophical Investigations*, §§246-253, esp. §253; cf. also Evans, *The Varieties of Reference*, pp.252f., who explicitly considers and rejects this possibility.

⁵⁵Anscombe, 'The First Person', p.61.

⁵⁶ Cf. Shoemaker, *Self-Knowledge and Self-Identity*, ch.2/3.

⁵⁷ Hume, *A Treatise of Human Nature* Book I, part IV, section VI.

What property of mine could be such that, whenever I encounter it, I infallibly know that it could only belong to me, i.e. that it singles out me uniquely? The only property that could possibly fulfil such a stringent requirement is the pseudo-property of 'mine-ness' or 'being identical with me', and we have already seen above which fatal consequences this suggestion necessarily runs into⁵⁸. What we have to explain is the fact that it is impossible that I should know that some person or other is in pain, and erroneously take that person to be myself. Immunity to error through misidentification is here trivially guaranteed for the simple reason that an identification -- understood as recognition of an entity by one of its features -- does not take place. If *per impossibile* a person would have to apply criteria of recognition in order to refer to himself he could -- as both Shoemaker and Anscombe have demonstrated in their own ways⁵⁹ -- never be sure that he has singled out the right entity, nor that there is only one such entity he calls his self.

Thus the view that self-reference necessarily involves a self-*identification* of the referring subject seems to find no support in an examination of self-reference with regard to the conditions for identification of particulars in general. To be sure, when -- under normal circumstances -- I think 'I am in pain' I *do* have identifying knowledge of myself. That is, I do know *who* it is who is in pain in the sense that I could cite a whole cluster of definite descriptions which uniquely identify RS, the person I am. (And of course I also have knowledge of myself via visual perception and proprioception.) But that I have this knowledge is not a necessary requirement for the possibility of self-reference. As may be brought out by considering Anscombe's case of the completely anaesthetized, radical amnesiac, I would not need to know who (which of all entities) it is who thinks the thought 'I am in pain' in order to think truthfully and with justification that *I* am in pain. This seems to me the point of Wittgenstein's therapeutic, but otherwise misleading, statement that 'when I say "I am in pain", I do not point to a person who is in pain, since in a certain sense I have no idea *who* is'⁶⁰ Clearly, I nor-

⁵⁸Cf. for such a suggestion, Maddell, *The Identity of the Self*, pp.135f., and Chisholm, *Person and Object*, cf. the discussion above. sect.3.3..

⁵⁹ Cf. Shoemaker in Shoemaker/Swinburne, *Personal Identity*, pp.103-105; and Anscombe, 'The First Person', p. 52/3. Cf. also Strawson's argument from a thousand souls, directed against the Cartesian, in 'Self, Mind, and Body', pp. 174/5.

⁶⁰Wittgenstein, *Philosophical Investigations*, §404.

mally have a very good idea who is in pain when *I* am in pain. It is RS, the worldly particular that I am, about whom I know a good many other things. The 'certain sense' in which I may be said to have no idea who is in pain can only amount to the truth that, in order truthfully and with justification to say that I am in pain, I do not have to know who, which of all entities, I am.

A point which is sometimes presented as an objection against the non-identification view of self-reference is the putative *de re-ness of self-referential beliefs*⁶¹. This objection is, I think, based on an illegitimate inference from the truth that all 'I'-beliefs can in principle always be *de re-ascribed* to the falsehood that 'I'-beliefs have in themselves *de re-character*. To put it very briefly: it is indeed true that from a belief-ascription of the form 'A believes that *he* (himself) is ϕ ' we can always proceed *salva veritate* to the corresponding *de re*-ascription 'There is an *x* such that $x=A$, and *x* believes of *x* that *x* is ϕ '⁶², whereas we cannot generally do this from a belief-ascription of the form 'A believes that *S* is ϕ ' (where *S* is supposed to be a singular term, which may, or may not, single out a given entity). Someone who believes that he is ϕ can -- in one sense -- also truly be said to believe *of himself* that he is ϕ . Self-referential beliefs, that is to say, can always be *de re*-ascribed. But, I shall argue, this does not imply that the self-referential belief expressed by 'I am ϕ ' in itself has *de re-character*?

Let us have a look at some of the criteria for a *de re*-belief cited in the literature. One of the criteria Chisholm gives is that in order for me to have a *de re*-belief about an object, I have to have 'encircled' it as the focus of a whole bundle of definite descriptions⁶³. Only then can I be said to stand in a relation of 'epistemic immediacy' to it. It is evident, however, that this is not the case for self-reference. In order to be able to think of myself, I do not have to know any description true of myself, and not stand in any kind of 'epistemic immediacy relation' to myself. On the contrary, I may adduce as many of those descriptions as you will, they would not enable me to develop

⁶¹ Cf. Chisholm *The First Person*, esp. pp. 18-20, and Lewis, 'Attitudes de Dicto and de Se', *passim*.

⁶² Cf. Chisholm, *The First Person*, p.18.

⁶³ Cf. Chisholm, *The First Person*, p.115. According to Chisholm's definition of 'epistemic immediacy', which is in its turn supposed to define the strong *de re-ness* of a belief, the degree of a person's intimacy with an object is the higher the more identifying facts he knows about it (cf. *op.cit.*, pp.109-111).

self-consciousness. Lewis gives two more criteria for the *de re*-ness of belief: in order for a subject to have a *de re*-belief about some object Y, the subject has to refer to Y

‘under some description Z such that either (1) Z captures the essence of Y, or (2) Z is a relation of acquaintance that the subject bears to Y’⁶⁴.

I think we have already seen above that none of these requirements is fulfilled by self-reference. When I refer to myself ‘under the description “I” when used as-subject’ (if one wants to use this misleading form of expression) I do not grasp my essence (cf. also sect.4.1.). But nor can I be said to be ‘acquainted’ with myself in any normal sense of ‘acquaintance’. Lewis concludes:

‘Self-ascription of properties is ascription of properties to oneself under the relation of identity. Certainly identity is a relation of acquaintance par excellence. So belief *de se* falls under belief *de re*’⁶⁵.

This claim, I confess, I simply do not understand. One may have doubts as to whether identity is a relation at all. But it certainly is not a relation of acquaintance.

It seems to me that it is the expression ‘of himself’ that is tricking us here, and misleads us into ascribing *de re*-ness to self-referential beliefs. In a way, the *de re*-ascribability of self-referential beliefs is simply due to the fact that someone who ascribes such a belief to a person in a sentence of the form ‘A believes that he is ϕ ’ has to make a presupposition of existence with respect to the person to whom he ascribes self-consciousness. His statement cannot be true unless A exists, since A is *per definitionem* the entity of which A is conscious in *self*-consciousness. And thus there is no danger of the referential component of the belief ascribed failing to have a referent. However, it is not this triviality that the *de re*-theorist, such as Chisholm, wants to make his thesis. He wants to interpret the ‘of himself’ in the *de re*-ascription of a self-referential belief in a stronger sense, namely as implying that the *de re*-ascription really ascribes a *de re*-belief to A (viz. a *de re*-belief with respect to *himself*). But for that he would have to adduce reasons which are independent of the fact that a *de re*-reading of an ascription of a self-referential belief is in principle always possible. I do not think that he could adduce such reasons, and I suspect that his claim rather rests on essentialist intuitions which make the step from ‘of himself’ to ‘of *his self*’ seem a very short, and indeed inevitable one. Could the complete anaesthetised amnesiac (who is deprived

⁶⁴Lewis, ‘Attitudes de Dicto and de se’, p. 542.

⁶⁵Op.cit., p.543.

of all sensory input and hence does not know who (i.e. which entity 'in the objective order' he is) console himself with the fact that at least he believes *of himself* that he does not know who he is? Could he console himself by the thought that in one sense at least he knows who he is, namely *himself*? Of course the convinced essentialist could always insist that our amnesiac knows that *he* does not know who he is, and that this 'knowledge that *he*...' is all he ever needs to ascertain his identity, i.e. that this is in fact just what knowledge of one's own identity consists in. But at this point he seems to have come full circle round in his argumentation. Perhaps the 'individualist's' view cannot be directly refuted. But, at least, it can be shown that much speaks against it, and nothing in its favour.

To sum up: A person's self-reference cannot be said necessarily to involve a self-*identification*. In order to refer to *himself-as-subject* a person does not have to pick or single himself out of all objects in the world. In referring to himself-as-subject he does not, as it were, have to choose from a manifoldness of objects before he can get hold of the right one (as would be the case if self-reference were a kind of perception). From the thinker's point of view, there is no such getting hold of, or grasping, of an object in self-reference. There is no identification, and therefore no possibility of mis-identification. It is, I think, mainly the similarity in linguistic surface structure between sentences of the form 'I am ϕ ' and sentences of the form 'A is ϕ ' that misleads us into assuming that self-reference involves the identification of a particular object. For, whereas in order to be able to grasp the content of a sentence of the form 'A is ϕ ' I have to identify the object A in one way or another, an analogous requirement does not apply to a sentence of the form 'I am ϕ '. I do not have to identify the entity I in fact am in order to be able to refer to, or think of, myself-as-subject (although, for sure, my utterance of 'I am ϕ ' typically enables you to identify me as the utterer of the sentence 'I am ϕ '). The grammatical multiplicity of 'I'-sentences does not correspond to their epistemological multiplicity, but rather disguises it. Put into Evansian terminology, we might make this point by saying that in order to think of himself-as-subject, a person need not have '*discriminating knowledge*' of himself. He need not have to know -- although he normally does know -- *which of all* objects in the world he is. There is here no need for discriminating knowledge because there is just no possible gap to be bridged between a person and his thought of himself-as-subject. In self-reference a person does not turn its feelers, which are normally stretched out onto the world, back upon itself. He simply *draws them in*.

4. THE ONTOLOGY OF SELF-REFERENCE

In the previous chapter I have tried to show that -- from a first-person perspective -- self-reference cannot be said to presuppose, or involve, an identification of the referent of 'I'. In this chapter I shall focus on the consequences which this epistemological result has with regard to ontological questions of self-reference. In particular I shall try to refute certain misconceptions, or systematic illusions, concerning the nature of the referent of 'I' to which the previously discussed epistemological characteristics of self-reference are prone to give rise. I shall argue that these misconceptions rest on a 'paralogistic' misinterpretation of the epistemological features of self-reference as ontological features of its referent, a misinterpretation well exposed by Kant in the 'Paralogisms'-chapter of the *Critique of Pure Reason*. As will become more obvious in Chapter 5, these 'paralogistic' misconceptions of the self have influenced (explicitly, but more often implicitly) much of the contemporary discussion on personal identity. By showing that the epistemological features of an entity's self-reference cannot be accounted for in terms of the ontological nature of this entity, and that therefore considerations concerning self-reference cannot predetermine questions of personal identity, I shall try to lay the groundwork for the discussion of personal identity in the following chapter.

4.1. The Simplicity and Particularity of the Subject

Cartesianism is a doctrine that has officially been declared dead long ago. But its spirit is still alive haunting much of contemporary thinking on self-reference. The presuppositions underlying the Cartesian conception of self-reference form the basis and starting-point of other, less disreputable, theses concerning personal identity which, in combination, may be characterized as the Simple View of personal identity. As I understand it, the Simple View of personal identity¹, is a syndrome of various, mutually supportive, theses concerning self-reference and personal identity. Below (in section 5.6.) I shall scrutinize the Simple View's account of *personal identity over time*. In the present context, I should like to scrutinize certain assumptions concerning *self-reference* on which the Simple View's account of personal identity -- sometimes explicitly, but more often implicitly -- is based.

The most characteristic feature of the Cartesian account is, of course, the thesis

¹The expression was first coined by Parfit in his original essay 'Personal Identity' in order to denote a cluster of theses which he took his Reductionist View to refute.

that the entity I refer to when I refer to myself-as-subject is an (essentially and exclusively) *thinking* substance (a *res cogitans*). According to this account, the criterion of *my* identity over time is the criterion of identity of entities of the sortal kind '*res cogitans*'; and what constitutes *my* survival is the further existence of a certain example of this kind (namely the particular *res cogitans* which is mine, or which I am). The hypostazation of a special referent of 'I' (when used as-subject) as an essentially and exclusively *thinking* substance is the most significant trait of the Cartesian account of self-reference. But, arguably, it is not this particular mistake which gets the Cartesian into his predicament in the first place. This is rather the more fundamental, and more widespread, assumption that the as-subject use of 'I' determines a *special referent* of 'I' (be it an exclusively thinking substance or not) which is conceived to be different from (or at least 'more inward than') the publicly observable person thus referring to himself, namely the very core of personhood. In referring to myself (as-subject), it is assumed, I refer to a special entity, namely *my self* -- the entity I am essentially, whose criterion of identity therefore constitutes the criterion of *my* identity, and whose further existence guarantees *my* survival. It is this assumption that, I think, constitutes the *nucleus* of the Cartesian as well as all other Simple Views of self-reference and personal identity.

Let me clarify these allegations: What characterizes the Cartesian in particular, and thus sets him apart from other adherents of a Simple View of persons, is his conceiving of the referent of his as-subject reference as an exclusively thinking thing, i.e. an entity whose essence is exhausted by its being the subject of *his*, the thinker's, experiences, and which therefore cannot have the bodily properties which he, as a concrete person, at least seems to have. But the Cartesian account is only one possible manifestation of the more general misconception that a thinker in referring to himself-as-subject identifies a given substance or particular *at all* -- be it as an essentially thinking thing or as whatever. The train of reasoning by which the Simple View's account of self-reference -- explicitly or, more often, implicitly -- arrives at its suggestive conclusions can perhaps be best spelt out into the following logical steps²:

² I do not want to imply that this way of reasoning is explicitly invoked by all, or even any, 'simple' accounts of personal identity. I only claim that it is surreptitiously at work in many of those accounts. Examples are McGinn, *The Character of Mind* pp. 111-116, Madell, *The Identity of the Self* pp. 122-140, and Chisholm, *Person and Object* pp. 104-113.

(a) A person's reference to himself-as-subject presupposes, or involves, identifying knowledge of the entity thus referred to. It presupposes, or involves, identifying knowledge in the double sense of knowing *which* particular entity is thus referred to (namely the particular self that I call mine), and *what kind* of entity (namely a self).

(b) In the case of self-reference this knowledge cannot consist in anything else but knowledge of one's 'individual essence'. For, as can be gathered from considerations of extreme cases such as the sensory-deprived amnesiac, a thinker may be deprived of all knowledge of an objective world (including his own body), but still be able to refer to himself-as-subject. (And, on the other hand, no descriptive or perceptual knowledge of what is objectively the case could ever take the place of the particular kind of knowledge required for self-reference.)

(c) In using 'I' as-subject, a person grasps his individual essence. This essence, although it cannot be further spelt out, but only be stated in indexical terms such as the Chisholmian 'being I', determines the very core of a person's being, his *self*. My self is what I am essentially. It is what remains after all my contingent features have been abstracted. It is an entity whose essence is wholly exhausted by its being the object of *my* reference to myself-as-subject. My self is the ontologically irreducible persisting bearer of all my mental and physical states, which by having these states gives unity to my life. It is -- as McGinn has most clearly expressed it -- 'a simple substance whose essential nature can only be captured in non-reductive terms.'³

(d) Since my self is what I am essentially, the further existence of this entity is what constitutes my survival. Since this entity is a simple substance, my (as every other person's) identity cannot but be a simple and unanalysable fact, a 'further fact over and above facts about mental [or any other kind of] connectedness'⁴. Furthermore, I (as every other person) have infallible knowledge of what the fact of my identity consists in. It consists simply in its being *me* who is the same at different times.

The two most important (putative) consequences of this hypostazation of the referent of 'I' are, of course, firstly, that cases of indeterminacy with regard to personal identity seem to be logically excluded, and secondly, that self-reference seems to provide a person with a privileged access to himself (or rather 'his self') and his (or rather its)

³ McGinn, *The Character of Mind* p.115; cf. McTaggart, *The Nature of Existence*, vol.2, ch.36, and Maddell's defence of his position in *The Identity of the Self*, pp.23-49.

⁴ McGinn, *The Character of Mind*, p.113.

identity over time. However, these consequences fall together with their premises. It is, in particular, the first premise, namely that a person has to have identifying knowledge of himself in order to refer to himself-as-subject, where, as I tried to show above, the Simple View of self-reference goes wrong. The Simple View realizes correctly that in referring to himself-as-subject a person does not refer to himself *as* the particular person, i.e. the particular limited entity in the world, that he is. It realizes, to put it differently, that a person's identity as a particular entity with all its concrete abundance does not enter into the 'intentional content' of his self-reference. But he mistakenly concludes from this that by 'I' used as-subject he does not (semantically) refer to this particular entity, but to something more inward, namely the core or the essence of his person, the Self or Ego. In other words, the Simple View concludes from the truth that there is a way in which (trivially) only he can refer to himself (namely by using 'I', in utterance or soliloquy) to the falsehood that there is an entity (his self) to which only he can refer. He thus illegitimately moves from a logical to an ontological privacy. As Wittgenstein puts it in the *Blue Book*:

'We feel then that in the cases in which "I" is used as subject, we don't use it because we recognize a particular person by his bodily characteristics; and this creates the illusion that we use the word to refer to something bodiless, which, however, has its seat in our body. In fact, *this* seems to be the real ego, the one of which it was said: "Cogito, ergo sum"'⁵.

Evidently, this criticism of the view that the entity referred to by 'I' is a necessarily simple substance is closely related to Kant's and Strawson's analyses of the Cartesian fallacy. Following Kant, Strawson has represented this fallacy as an illegitimate inference from the fact that a person can ascribe experiences to himself without thereby drawing on any criteria of his identity to the illusion that the 'I' by which experiences are thus self-ascribed (obviously an instance of the as-subject use) can therefore be used to achieve a 'purely inner reference'⁶; -- a reference to 'an absolutely simple, identical, immaterial individual'⁷. Strawson does not want to deny to the Cartesian that there is indeed a sense in which the use of 'I' as-subject can be said to refer to, or single out, a certain individual. But this, he explains, is due only to the fact that 'I' is normally used as a communicative device (that, we might say, it has the person who uses 'I' as

⁵ Wittgenstein, *The Brown and Blue Books*, p. 69.

⁶ Strawson, *The Bounds of Sense*, p.165.

⁷ Op.cit., p.166.

its semantic referent) and is in this way linked to the empirical criteria we use in identifying a certain person. The links between the criterionless self-ascription of experiences and the empirical criteria of the identity of the person who self-ascribes certain experiences are not, as Strawson puts it, '*in practice severed*'⁸. But they are severed in the Cartesian attempt to interpret the use of 'I' as referring to a purely inner object, i.e. an entity whose essence is conceived as being fixed merely by its being the object of *my*, the thinker's, self-conscious thought. The Cartesian thus cuts the as-subject use of 'I' off from its background-conditions which first endow it with its 'referential force'⁹ -- or, as I would prefer to put it, its identifying' character -- while at the same time closing his eyes to the fact that this procedure has in fact deprived the use of 'I' of this force.

Kant, however, was well aware of this. When he speaks of the 'transcendental subject' of the thought 'I think' he does not mean a sharply individuated entity which makes itself the object of one of its thoughts when it accompanies any of its representations with the 'I think'. As Kant makes sufficiently clear in his 'Paralogisms of Pure Reason', the pure transcendental self-consciousness is not to be understood as the consciousness of a substance. It is not the consciousness of a particular entity, but the consciousness of a unity, -- a unity constituted by the connectedness of experience which, according to Kant, is ensured by the applicability of the categories. The 'I think', he says in one passage, is 'not a representation distinguishing a particular object, but a form of representation in general.'(B404)¹⁰. Transcendental self-consciousness is the consciousness of a unity simply because its 'referent' -- if a 'pure form' may be called a referent -- has been deprived of any inherent determinations. As Kant puts it:

"'I' means a something in general (transcendental subject) the representation of which must, no doubt, be simple if only for the reason that there is nothing determinate in it.'(A355)¹¹.

Any empirical content is, as it were, cleared out of the notion of a transcendental subject. And it is just this clearing out of all empirical content from the representation 'I think' which, at the same stroke, makes it possible for a subject to refer to himself in

⁸Op.cit., p.165.

⁹Strawson,*The Bounds of Sense*, p. 165.

¹⁰References in text are to the two editions of Kant's *Critique of Pure Reason* in Kemp-Smith's translation.

¹¹Cf. also B404.

the peculiar as-subject manner, and makes it impossible for pure self-consciousness to be, as Kant puts it, 'knowledge' of an object. By depriving 'I' of its individuating force, we deprive its putative referent of his ordinary criteria of identity. What remains is an abstract, general unity. The 'rational psychologist', however, neglects this consequence. In assuming that the phenomenological features (i.e. unity, singularity, simplicity, etc.) of the contentless unity of consciousness are perceived properties of an entity, he -- as Kant puts it -- misconceives 'the unity in the synthesis of thoughts as a perceived unity in the subject of these thoughts'(A402), or -- as Strawson puts it more elegantly -- he 'confuses the unity of experience with the experience of a unity'¹².

The contentless, non-individuating nature of transcendental self-consciousness also explains why Kant could keep his 'I think'-postulate (i.e. the postulate that it must be possible for the 'I think' to accompany all representations which a subject is to count as his own¹³) entirely independent of any considerations concerning the nature of the thinker, i.e. considerations concerning what *kind* of entity the thinker is. A thinker's reference to himself-as-subject neither draws on, nor hence determines, any features of his essence. It does *not draw on* any features of his essence, since in order to refer to himself-as-subject a thinker does not have to draw on any criteria of his identity. He does not need to know who (which of all entities) or what (what kind of entity) he is. His nature does not enter into the intentional content of his self-reference. Nor does such a reference *determine* any of these features; what a thinker is is a matter completely independent of the nature of his self-reference. I, the thinker, might be a *res cogitans*, I might be a human being, or simply a material body, or a brain in the vat, or what you will. But which of all these, which sort of entity I am, is completely irrelevant to the questions concerning the nature of self-reference. For in referring to myself-as-subject I do not refer to myself *as* the particular person, the determinate entity in the world, that I am.

We may agree with Strawson that Kant was successful in his attempt to discard as an illusion the idea that the referent of 'I' can be shown by introspection to be a necessarily simple, unitary substance. As a criticism of Cartesianism and other varieties of a Simple View, Kant's reflections on self-consciousness are of a lucidity seldom to

¹² Strawson, *The Bounds of Sense* p.162.

¹³Cf. Kant, *Critique of Pure Reason*, B 131.

be found in his work. However, when it comes to Kant's own positive account of self-consciousness, matters become more complicated and obscure. There is of course, first of all, Kant's 'transcendental idealist' view that the 'I' of transcendental apperception also (obliquely) refers to a *thing in itself*, namely me as I am in myself, the noumenal or metaphysical subject¹⁴. This obscure view seems at times to have tempted Kant to assert that the subject is in itself unknowable, and that we thence cannot know what constitutes its criterion of identity. One must not underestimate the influence of his transcendental idealist convictions on Kant's conception of self-consciousness. However, I do not think it is generally as significant (and hence as malignant) as particularly Strawson has made it out to be. Let me briefly express my qualms:

In *The Bounds of Sense* Strawson claimed that for Kant the 'I think' of apperception 'represents the tangential point of contact between the field of noumena and the world of appearance'¹⁵. He has since elaborated this suggestion by attributing a dual view concerning the reference of 'I' to Kant¹⁶. According to the first part of this dual view, the 'I' of apperception has a 'purely non-denotative significance'¹⁷, referring, if only 'formally', to the unity of consciousness. According to the second part, the 'I' of apperception refers to the subject as it is in itself, which, as Kant claims in B422, 'cannot be known'. Now, it is indubitable that various passages in both editions of the *Critique* may suggest that Kant held such a dual view (e.g. the ones cited in footnote 14). However, it seems significant to me that most of these passages were dropped in the B-edition, where in one passage Kant explicitly rejects the view imputed to him by Strawson. namely in B423 (note), where he says that in self-consciousness

'something real ... is given, given indeed to thought in general, and so not as appearance or as thing in itself (noumenon)'.

Similarly, in B429, Kant states that in self-consciousness

'I...represent myself to myself neither as I am nor as I appear to myself. I think

¹⁴Cf. Kant, *Critique of Pure Reason*, A350 where he speaks of 'the subject in itself which as substratum underlies this 'I' as it does all thought', or A355, where he claims that 'in attaching 'I' to our thoughts we designate the subject of inherence only transcendently, without noting in it any quality whatsoever.'

¹⁵Strawson, *The Bounds of Sense*, p.173,

¹⁶Cf. his 'Kant's Paralogism: Self-Consciousness and the "Outside Observer"', esp. pp. 205-212.

¹⁷Op.cit., p.207.

myself only as I do any object in general from whose mode of intuition I abstract.'

I am inclined to think that in all those passages in which Strawson takes Kant to refer the 'I' of apperception to the 'subject in itself', Kant really only refers to what he calls the 'transcendental subject of the thoughts = x' (B404), which is conceived by Kant as a counterpart to the 'transcendental object = x' (A250) of the Transcendental Deduction. Both are conceived as mere unities of the manifold, void of all empirical content, and, that is, they are conceived in the sense of the first part of Strawson's dual view. I am inclined to think that it is Strawson's failure to distinguish between the transcendental subject/object and things in themselves that tempted him to attribute to Kant a view he arguably did not hold. However, I shall not try to substantiate this claim here, since this aspect of Kant's theory seems to me of interest only to Kant-scholarship.

The problem I should like to address instead is a different, although closely related, one. It can be brought out by raising the question in what sense the referent of the 'I' of apperception, allegedly a pure 'form without content' or 'mere unity of consciousness', can, as Kant claims, be said to *exist* and to be *numerically identical*. It may make sense to say of a form that it exists, and perhaps even that it is numerically identical. However, this is neither the sense in which each of us thinks of himself that *he* exists and is numerically identical over time, nor -- as we shall see -- is it the sense in which Kant invoked these notions with respect to apperception. The sense in which each of us thinks of himself that he exists and is numerically identical over time is, it seems evident, that of a *particular's*, not a form's, existence and persistence. And this is to say, as we saw in sect.2.1.4., that he conceives of himself as an entity for which there is the possibility of a contrast between numerical and merely qualitative identity (which does not exist for forms or universals). The issues raised by discussing Kant's attempts to reconcile this basic phenomenological fact with his conception of apperception as a mere form or unity of consciousness are of a wider significance, which reaches beyond an interpretation of Kant's thought, and will occupy us again below (cf. sections 5.4. and 5.5.).

As I see it, the difficulty that self-consciousness poses for Kant is that it is supposed to combine elements of generality and particularity at the same time. One way in which this difficulty manifests itself in Kant's work is in his wavering attempts to subsume the phenomenon of self-consciousness under his dichotomy of *intuition (particularity)* and *concept (generality)*. Thus, in the A-edition, Kant claims that the 'I' of apperception is 'as little an intuition as it is a concept of any object; it is the mere form

of consciousness' (A382). Similarly in the B-edition, after first having referred to 'the concept, or, if the term be preferred, the judgement "I think"' (B399), Kant repeats his denial of the conceptual character of apperception when he claims that 'we cannot even say that this [i.e. apperception] is a concept, but only that it is a bare consciousness which accompanies all concepts' (B404). Nor, of course, is transcendental apperception to be characterized as an intuition, since in pure apperception (as opposed to inner sense) nothing is given for thought to which the category of substance could then be applied. (This is just the mistake of the rational psychologist.) Kant variously tries to capture the unique epistemological status of the 'I think' by explaining that it is 'merely the consciousness of my thoughts' (B413) or 'a mere form of consciousness' (A382); or that it is 'not itself a representation distinguishing a particular object, but a form of representation in general' (B404); that it 'serves only to introduce all our thought as belonging to consciousness' (B400). And in a passage in the 'Refutation of Idealism' he makes another attempt. He says:

'The consciousness of myself in the representation 'I' is not an intuition, but a merely intellectual representation of the spontaneity of a thinking subject' (B278).

It seems evident that the entirely opaque characterization of self-consciousness as an 'intellectual representation' is more a desperate subterfuge than an attempt at clarification. But it is also evident why Kant must have felt himself driven to deny both the intuitional as well as the conceptual character of apperception. For if apperception were an intuition, then it would have to be said to yield knowledge of a *particular*, i.e. of something to which the category of *substance* could be applied. If it were merely conceptual, then it could not be said to include the idea of particularity (and thence numerical identity) of the entity referred to by 'I', and thence it could not be said to include the certainty of the *existence* of a certain particular.

However, as is often overlooked, this is exactly what Kant wants to hold. For Kant *did* agree with Descartes in that 'the "I think"... contains within itself the proposition "I exist"' (B422, note), and thus that the *Cogito*-argument proves the existence of a given *particular*, not just the existence of a general form or unity. What he found fault with in Descartes' *Cogito*-argument was -- apart of course from the 'rational psychologist' conclusions Descartes drew from it -- only Descartes' (occasional) tendency to represent the relation between his 'cogito' and his 'sum' as one of *inference*. Kant's claim is that the 'cogito' and the 'sum' are actually *identical*, one and the same proposition, as is evident from the following passage: 'My existence cannot be regarded as an

inference from the proposition ‘‘I think’’...but is identical with it.’(B423)¹⁸. This equation of ‘cogito’ and ‘sum’ is also manifested in a reprise of the famous formula of §16 of the Transcendental Deduction which goes as follows:

‘Certainly the representation ‘I am’, which expresses the consciousness that can accompany all thought, immediately includes in itself the existence of a subject...’-(B278).

Obviously, Kant must have thought that in the case of the subject there has to be a much closer connection between its concept of itself and its existence (its intuition) than for any other entity. But he felt unable to explain this intimate connection in terms of his concept-intuition dichotomy. His dilemma was that he could not account for the certainty of the existence of a particular involved in apperception by saying that apperception is an intuition; and that at the same time he could not discard apperception as purely conceptual, since this would have made it impossible for him to claim -- as he wanted to -- that apperception does contain the certainty of the existence of a *particular* (a substance). In a surprising passage dealing with the Cogito-argument, Kant states that ‘the ‘I think’ expresses an indeterminate empirical intuition, i.e. perception...’ (B423), and he goes on to explain:

‘An indeterminate empirical perception here signifies only something real that is given, given indeed to thought in general, and so not as appearance, nor as thing in itself (noumenon), but as something which actually exists, and which in the proposition ‘I think’ is denoted as such.’(B423).

We can see why the (paradoxical) characterization of self-consciousness as an indeterminate empirical perception must have sounded appealing to Kant. For the perceptual or intuitional aspect could account for the certainty of the existence of a particular involved in it, its indeterminacy for the fact that this certainty cannot be said to involve any knowledge of the ‘perceived’ particular. But the paradox involved in calling a particular indeterminate is perhaps more a sign of the inapplicability of Kant’s logical dichotomy between intuitions and concepts to self-consciousness than the expression of a deep insight.

Kant’s difficulties in characterizing the logical status of self-consciousness may tempt one to claim that self-consciousness, and thence the notion of a subject of thought,

¹⁸Cf. A345: ‘...what is referred to as the Cartesian inference *cogito, ergo sum* is really just a tautology, since the *cogito* (*sum cogitans*) asserts my existence immediately.’

can be understood only *via negativa*¹⁹. And one may even come to doubt -- as Lichtenberg seems to have done²⁰ -- that so much as the idea of the *particularity* of the subject enters into a person's reference to himself-as-subject, or whether this idea is not rather imported from the normal background-conditions of our application of the first-person pronoun²¹. However, I think this would be going too far. Kant certainly thought that there is a sense in which what Williams once called 'the pure point of view of consciousness'²² provides us with the idea of a particular, i.e. the idea of something for which there is a contrast between strict numerical and merely qualitative identity. This is brought out by Kant's continued application of the concept of numerical identity to the subject of apperception, which would not make much sense if he conceived of that 'subject' as a mere form or unity. Thus in one passage he claims that

'as self-consciousness is a transcendental representation, numerical identity is inseparable from it, and is a priori certain.'(A113)²³.

I think we can agree with Kant that self-reference is not general, but *singular* (though not identifying) reference. We may, that is, agree with him that the first-person perspective *does* provide for the idea of the particularity of the subject of thought, and thence for the contrast between numerical and merely qualitative identity which is characteristic of particulars, as opposed to other kinds of individuals. What the first-person perspective *cannot* do by itself is give substance to, or fill in, this contrast. That is to say, it cannot by itself provide any criteria of identity, any counting and reidentification rules, for the entity thus referred to. (And the Cartesian mistake consists in just that, i.e. in falsely assuming that the first-person perspective *can* provide such criteria, and thus constitute a special kind of entities, *res cogitantes*.) One may, that is to say,

¹⁹Cf. Vendler, *The Matter of Minds*, p.110.

²⁰Cf. Williams' discussion of the matter in his *Descartes: The Project of Pure Enquiry*, pp. 95-101.

²¹ As Strawson sometimes seems to suggest in *The Bounds of Sense*, p.166; and in 'Self, Mind, and Body', p.176.

²²Cf. Williams, *Descartes: The Project of Pure Enquiry*, p. 100. Of course, it is dubious, to say the least, that the very notion of a 'pure perspective of consciousness' is intelligible. However, I think it put to good use in elucidating the illusory character of the 'paralogisms'.

²³Cf. also A107, A108.

maintain (with Kant) that the thought of oneself-as-subject is -- although 'not itself a representation distinguishing a particular subject' (B404) -- nevertheless the thought of 'a something in general' (A355)²⁴, as long as one holds at the same time that it must be possible to *identify objectively* the *very same* particular that is *non-identifyingly referred to* by the 'I' of apperception. One may, in other words, admit to Kant that the subjective, or first-person perspective provides not only the idea of a mere *unity or form* of consciousness, but that of a *particular* (for which there is a contrast between numerical and merely qualitative identity), as long as one remains aware that it is the objective, or third-person, perspective which determines what *kind* of thing this particular (the referent of the 'I' of apperception) is. For *only* the objective, or third-person perspective can determine what kind of thing this particular is and, in doing so, provide criteria of its numerical identity and at the same time fill in, and give substance to, the idea of a contrast between strict numerical and merely qualitative identity of the entity referred to by 'I'. This thesis will be further explored in the next section.

4.2. The 'Logical Identity' of the Subject

Arguably, the knowledge of what it is for a particular to be of a certain kind (i.e. what constitutes the essence of a certain particular) and the knowledge of what it is for *one* item of a given kind to be the same at different times (i.e. what constitutes the identity over time of a particular of the given kind) are systematically related. A thinker could not be said to know what kind of thing a given particular is unless he had at least a vague idea of what is to count as the numerically identical persistence over time of a particular of that kind (cf. section 2.1.above). However, personal identity may be thought to be an exception to this rule. It may be held that, although, admittedly, self-consciousness cannot by itself provide a person with an insight into whatever constitutes his nature or essence (and therefore cannot reveal a person to be a necessarily simple substance), it may nevertheless provide a person with a special *first-person access to whatever constitutes his identity over time*. Although, that is to say, self-consciousness remains silent about what *kind* of entity an 'I'-thinker is, it at least provides a person with all the knowledge he ever requires to determine his identity for all past

²⁴ Cf. Kant's talk of the 'transcendental subject = x' in B 404, or of 'this I or he or it (the thing) which thinks'(ibd.).

and future time -- *whatever* he is. A person's self-reference, one may be tempted to think, can -- at any given moment -- fix that person's identity once and for all. A person can define his present identity and -- in the Nagelian vein cited above (sect.2.2) -- extrapolate it into the past and future simply by referring to himself as the subject of a present experience of his. The mere thought 'I-NOW' may thus seem suited to fix a person's identity in a way which excludes the possibility that there ever could be a genuine indeterminacy of whether a past or future person was or will be *him*. Just as a laser-ray, shooting from a given point of reference into infinity, determines, for every given object of the universe, whether it is touched by it or not, the thought 'I-NOW' seems to determine, for every given future situation, whether it will happen to *me* or not. And the fact that we are not able to spell out what this (putative) reference-fixing knowledge of onself consists in, may just be said to be due to its very character, namely that of an immediate and intuitive acquaintance.

This is a weaker claim than the one dealt with in the previous section, since it does not require the subject to have identifying knowledge of himself (understood in the strong sense of knowledge of what *kind* of thing the subject is). Nevertheless, I think, it is based on a related paralogistic misconception. Being more modest in its pretensions, and at the same time much 'airier' in its argumentation, it is much more difficult to expose it as the illusion I think it to be. It is, what is more, one of those illusions whose attraction recurs even after insight into their illusory character has been gained. But it *is* an illusion, as Kant tried to demonstrate in an acute analysis by the title 'Third Paralogism of Pure Reason'. In the course of this discussion, in which he deals with the issue of a person's first-person access to his identity over time, Kant states that changes could occur in the referent of 'I' such that

'despite the logical identity of the "I" such a change may have occurred in it as does not allow of the retention of its identity, and yet we may ascribe to it the same-sounding "I" which in every different state, even in one involving change of the thinking subject, might still retain the thought of the preceding subject and so hand it over to the subsequent subject'²⁵.

In a fascinating footnote to this passage Kant illustrates his claim by an analogy between conscious monads and elastic balls: just as a series of elastic balls may collide with each other, each one transferring its 'motion' to the next, so a series of monads

²⁵ Kant, *Critique of Pure Reason* A 363. This thought-experiment is reminiscent of Locke, *Essay on Human Understanding*, Book II, ch. XXVII, sect- 13.

may each transfer its 'consciousness' to the next, each successor retaining the memory and the consciousness of its respective predecessor²⁶. In such a case, Kant seems to suggest, each token of utterance (or 'thought') of 'I' would have a different referent, viz. the monad that produces just this token of 'I'. What is more, the referent may even change during the thinking of 'I', if we imagine the change of monads happening fast enough. In envisaging this possibility, Kant certainly did not mean to make a positive suggestion to the effect that the real foundation of our consciousness is such a kind of monad-billiards. What then, however, is the meaning and purpose of this Kantian 'thought-experiment'?

One way of understanding the passage just cited would be to interpret it as an allusion to, and elaboration of, Locke's distinction between the 'consciousness that makes a man be himself to himself' (and is thus what personal identity consists in) and the one or more 'individual substances' which, according to Locke, underlie this consciousness²⁷. Locke famously argued that personal identity over time is radically different from the identity of material things insofar as it does not require the persistence of any particular substance or matter, but simply the identity of a given 'consciousness', which could in principle be transferred from one body to another. And Kant -- in the passage just cited -- might be understood as trying to substantiate this Lockean claim by illustrating it from the point of view of a subject whose consciousness is transferred from one 'substance' to another.

We shall return to the Lockean suggestion that personal identity consists in 'identity of consciousness' (or, in its updated version, 'psychological continuity') below. However, although one should not underestimate Locke's influence on Kant's discussion of the 'Third Paralogism', it seems *prima facie* unlikely that Kant should here have intended to support Locke's theory of personal identity from a 'transcendental perspective'. For, as Strawson has recently pointed out²⁸, we have to keep in mind that Kant's main aim in the 'Third Paralogism' (as opposed to Locke's aim in the *Essay*) was not

²⁶ Cf. Kant, *Critique of Pure Reason*, A 363/364 .

²⁷Cf. Locke, *Essay Concerning Human Understanding*, Book II, ch. xxvii, sect. 9, cf. also sect. 13.

²⁸Cf. Strawson, 'Kant's Paralogisms: Self-Consciousness and the "Outside Observer"', pp.212-219, which is presented as a criticism of Bennett's interpretation of the 'Third Paralogism', in Bennett, *Kant's Dialectic*, pp. 93-113.

so much to defend or attack a certain theory of what personal identity 'consists in', but to rebut certain unjustified pretensions to knowledge of the 'rational psychologist'. According to Strawson, Kant's aim was not so much to dispel certain philosophical illusions we might have about ourselves as *empirical* selves, i.e. about ourselves as 'objects of experience', but rather to expose as illusions the 'transcendent' knowledge-claims of the rational psychologist. The rational psychologist, working from his 'sole text', the judgement 'I think' (A343), presumes that 'the pure perspective of consciousness' can reveal a person to himself *as he is in himself* (as he is as a *thing in itself*). This leads him to the conclusion that what is thus revealed is a single, identical, simple, immaterial substance. And, as we have seen, this conclusion is shown by Kant to be a delusion based on a misconception of 'the unity in the synthesis of thoughts as a perceived unity in the subject of these thoughts' (A402).

Strawson certainly has a point here. Nevertheless, I think that Kant's discussion of the Third Paralogism contains more than just a rejection of some, now slightly antiquated, 18th century conceptions of the self -- conceptions of a Cartesian or Leibnizian ancestry. A thoroughgoing discussion of Kant's Third Paralogism would have to examine its relation to Hume's claim that 'when he enters most intimately into what he calls himself' he does not 'stumble over' a certain particular that he would call his self, and the conclusion Hume draws from his failure to find the self in consciousness as to its fictional character²⁹. A detailed interpretation would also have to take into account the role of the 'outside observer' in Kant's treatment of the topic. However, I shall not make an attempt at such an interpretation here³⁰. I shall instead focus on one particular aspect of Kant's reflections, namely his assertion of a certain shortcoming of the 'pure point of view of consciousness' as a means of determining matters of personal identity, which in turn is closely related to the distinction Kant draws between the numerical identity over time of the subject as conceived in the *ontological* (or 'rational psychologist') sense, and the purely '*logical identity*' which Kant himself attributes (as *a priori* certain) to the subject of thought. *Pace* Strawson, I think that Kant's reflections contain insights which can be critically applied not only to the transcendent claims of the rational psychologist, but also to the topic of 'empirical' personal identity. I think they can show

²⁹Cf. Hume, *A Treatise of Human Nature*, p.252.

³⁰Cf. Powell, *Kant's Theory of Self-Consciousness*, pp. 130-174.

something with respect to the theoretical value of at least some first-person considerations in thought-experiments concerning personal identity.

One recurring *motif* of Kant's otherwise obscure discussion of the 'Third Paralogism' is his claim that the purely first-person perspective cannot provide us with a 'permanent element' (A362) in experience onto which we could latch in making a substantial claim as to the numerical identity of the subject of experience. Kant explains what he takes to be the rule of the application of the concept of numerical identity over time to a given substance in the following way:

'If I want to know through experience, the numerical identity of an external object, I shall pay heed to that permanent element in the appearance to which as subject everything else is related as determination, and note its identity through the time in which the determinations change.' (A361f.)³¹

And he goes on to point out that, since self-consciousness does not provide us with such a 'permanent element' in appearance (which, I take it, would have to be provided by a sortal concept³²), it cannot provide us with a *means of tracing* a given particular as numerically identical through time. Kant sometimes makes this point by saying that the first-person perspective cannot provide us with any firm '*correlatum*' (A366) against which we could, as it were, hold our identity-claims such as to confirm or disconfirm their truth. It cannot, as it were, provide us with a firm stance against the 'fleeting representations', which would allow us to distinguish 'the permanent element in appearance' from its ever-changing modifications. Thus the passage in A366 reads in full:

'...if I want to observe the mere 'I' in the change of all representations, I have no other *correlatum* to use in my comparison except again myself, with the universal condition of my consciousness'.

This, Kant seems to claim, has the consequence that -- from the first-person perspective -- the *appearance* of numerical identity over time is *inevitable* and thence *trivially* guaranteed as a necessary by-product of all consciousness. Since '...in the whole time in which I am conscious of myself, I am conscious of this time as belonging to the

³¹Cf. also A362, B413.

³²As I understand Kant's talk of the 'permanent element', it is what is specified by a given sortal concept as the nominal essence of whatever falls under it. Thus, if we want to find out if a horse H_1 at t_1 is numerically identical with a horse H_2 at t_2 , we 'pay heed' to the required 'permanent element' specified by the concept 'horse' (i.e. the necessary and sufficient criterion of a horse's identity), to which the horse's *accidentia* are 'related as determinations'. If we find the 'permanent element' to have continuously existed between t_1 and t_2 , H_1 is identical with H_2 , if not, then not.

unity of myself', Kant claims, '...in my own consciousness ...identity of person is unfailingly met with' (A362). And in another passage Kant tries to explain this 'phenomenological' necessity by pointing out that

'since we reckon as belonging to our identical self only that of which we are conscious, we must necessarily judge that we are one and the same throughout the whole time of which we are conscious'(A364).

He also underlines the trivial or, as he puts it, 'tautological'(cf. A366), and hence empty, character of this sense of identity by claiming that 'the proposition, that in all the manifold of which I am conscious I am identical with myself, ...is an analytic proposition.'-B408)³³.

Kant concludes from these remarks that 'we are unable from our own consciousness to determine whether, as souls, we are permanent or not.' (A364). This, of course, sounds very much like a 'transcendental idealist' denial that we cannot really know ourselves as we are in ourselves (as things in themselves). However, we might be able to give Kant's reflections a more 'austere' interpretation, which can show us what is wrong with Nagel's idea that a thinker can extrapolate this identity for all past and all future 'just by concentrating on his present experience and specifying the temporal extension of its subject'.

We said above (in our discussion of the putative simplicity of the self) that the 'pure perspective of consciousness', although it may provide us with the *idea* of the particularity of the subject of thought, cannot *by itself* give substance to, or fill in, this idea. It cannot by itself *establish* the possibility of a contrast between strict numerical and merely qualitative identity (which is characteristic of particulars as opposed to types or universals), since it cannot by itself provide any criteria of identity of the particular thus (non-identifyingly) referred to. And it cannot provide any such criteria since in referring to himself-as-subject a person does not draw on, or invoke, any criteria of his identity. I suggest that we understand Kant's talk of the purely 'logical identity' of the subject as applying this basic insight to the idea of a person's putative first-person

³³The passage continues: 'But this identity of the subject of which I can be conscious in all my representations, does not concern any intuition of the subject, and cannot therefore signify the identity of the person, if by that is understood the consciousness of the identity of one's own substance, as a thinking being, in all change of states'. Cf. also A 364. Kant argues similarly in A 116, B 138. Cf. also his representation of the paralogism of simplicity in the B-version, B 407/8.

access to his identity *over time*.

We saw above that the unity of apperception (into which all my representations are united insofar as I can 'accompany' them with an 'I think'-judgement) can be characterized as simple -- if only for the reason that it is deprived of any inherent determinations (or manifoldness). As Kant puts it in B429, in self-consciousness 'I think myself only as I do any object in general from whose mode of intuition I abstract'. However, just as it is inevitable that something that is referred to in this general or formal way *as* a bare bearer of qualities (determinations) must appear as simple (since, lacking 'determinations', it could not possibly appear non-simple), something that is *repeatedly* (i.e. at different times) referred to as such a bare bearer of qualities (determinations) must appear as numerically identical *throughout these times* (since it could not possibly appear non-identical). There could not possibly be an 'identity-gap' between different instances of a person's reference to himself-as-subject, since none of these instances of self-reference invokes any criteria of identity of the entity thus referred to. And since none of these instances invokes any criteria of identity, several of them together cannot do so either. Just as, from the 'pure point of view of consciousness', one single act of self-reference cannot serve to establish a substantial identity of the self, several of those acts taken together cannot do so either. And this is why a claim as to the numerically identical persistence of a given 'self' based purely on the first-person perspective must turn out to be either ungrounded or empty. It is ungrounded if the claim to numerical identity is understood in an ontological sense. It is empty if it is understood in the purely 'logical' sense suggested by Kant.

The fact that a person's thought of himself-as-subject rather consists in an abstraction from all his inherent determinations than in an identifying ('reflexive') reference to a given particular is also the reason why the first-person perspective cannot provide us with a 'permanent element in appearance', and hence with no means of tracking a particular. It cannot provide us with a permanent element since a person's thought of himself-as-subject just consists in an abstraction from all the determinations in relation to which it could appear as permanent. And it cannot provide us with a means of tracking a particular falling under it, since a person's thought of himself-as-subject is not an identifying reference, i.e. since it does not make use of a sortal concept which would specify certain requirements of permanence for anything falling under it. Just as from the 'pure point of view of consciousness' no determinate particular is *identified*, no particular is *traced*.

What Kant calls the ‘logical identity’ of the subject is thus a trivial concomitant of the ‘coherence’³⁴ which a person’s consciousness has to display in order to represent an objective world, i.e. in order to be consciousness of (or at least *as of*) an objective world. Being a necessary by-product of all consciousness, it cannot *by itself* warrant any conclusions as to the identity of the person *whose* consciousness it is. The idea of an identical subject that persists as the same throughout all consciousness is nothing more than the subjective upshot of this coherence. (Or perhaps better: the talk of the ‘coherence’ or ‘connectedness’ of experience is just a way of objectifying the formal rule-governed coherence that consciousness has to display in order for me to be able to accompany any and all of my representations with the ‘I think’). The logical identity of the subject is in this sense a phenomenological necessity. It is subjectively necessary that I experience myself as the same in all experiences which are connected together into *one* consciousness. (This is just how *one* consciousness is defined.) The ‘constant logical subject of thoughts’ (A350) stretches into all past I can remember and all future I can anticipate. Even if I had entered into existence with all my apparent memories only a moment ago, my ‘logical identity’ over all my imaginary past would still be maintained, although there could not even be so much as a candidate for my ‘ontological identity’ at the time of my non-existence.

This, I take it, shows two things: it shows, *firstly*, that the logical identity of the subject cannot by itself guarantee personal identity in the substantial sense required by Nagel’s idea of a reference-fixing mental act or by an adherent of the Simple View, who maintains that matters of personal identity can be settled purely from the first-person perspective. Persons are particulars, and personal identity consists in the numerical identity of a given particular over time. The ‘logical identity of the subject’ *cannot* guarantee the identity over time of a certain determinate particular and hence cannot by itself guarantee the numerical identity of the given particular that I am³⁵. Kant’s reflections on the ‘logical identity of the subject’ show, in other words, that personal identity cannot be conceived as whatever constitutes *my* identity, where the referent of ‘my’ is

³⁴Cf. A363: ‘The identity of the consciousness of myself at different times is therefore only a formal condition of my thoughts and their coherence, and in no way proves the numerical identity of my subject.’

³⁵A similar point has been made recently by Carol Rovane in her ‘Branching Self-Consciousness’, esp. pp. 376-371. However, I should not want to go along with some of the ‘reductionist’ conclusions she draws from it.

supposed to be picked out purely from the first-person perspective. For, as we have seen above (sect.3.4.), no particular entity is thus singled out.

It shows, *secondly*, -- and this is just the other side of the same coin -- that there are no criteria of identity for 'subjects' or 'selves' or 'consciousnesses' as such. We saw above that a person's thought of himself-as-subject does not draw on any criteria of that person's identity. Arguably, the ability to refer to oneself-as-subject is a necessary feature of personhood. We would not count an entity as a person unless it displayed this ability. However, the ability to refer to oneself without applying any criteria of one's identity does not *define* a person's identity. In other words, the pure perspective of consciousness cannot provide us with any criteria of identity of a 'self' or 'consciousness' which could be grasped *independently* of the criteria of identity governing the persons (i.e. the particular entities, the genuinely persisting substances) *whose* 'selves' or 'consciousnesses' they are. This is, of course, a point well and repeatedly made by Strawson³⁶. However, I think that its relevance to questions of personal identity has not always been fully appreciated in contemporary discussions of the topic. I shall discuss the matter in more detail below (chapter 5), where I shall argue that the basic Kantian insight that there are no criteria of identity for 'selves' or 'consciousnesses' as such can help us to see what is wrong with various version of the Simple View, as well as Lockean (and neo-Lockean) theories, of personal identity.

At present, however, I should like to address a different, though closely related, question. We have already noted above that -- in spite of his own warnings as to its misleading character -- Kant had no scruples referring to the 'purely logical identity' of the subject as a case of *numerical* identity. This is manifest from various passages, e.g. A113, where -- in the context of the Transcendental Deduction -- Kant claims that

'since self-consciousness is a transcendental representation, numerical identity is inseparable from it and is a priori certain' (A113)³⁷

or A365, where -- in the context of the Paralogisms -- Kant speaks of 'that numerical identity of our self which we infer from identical apperception' (A365). As we know, Kant views the unity or numerical identity of apperception as a 'transcendental ground',

³⁶Cf. Strawson, 'Self, Mind, and Body', p. 174/5; and *The Bounds of Sense*, pp.162-169.

³⁷Cf. also A362, where he says that 'the personality of the soul has to be regarded ... as a completely identical proposition of self-consciousness in time...and is valid *a priori*'.

i.e. a necessary condition, of our having experience of (or at least *as of*) an objective world, which he takes himself to have established in the Transcendental Deduction. Some Kant-interpreters have interpreted Kant's talk of the numerical *identity* (as opposed to the mere *unity*) of apperception as implying a genuine contrast³⁸. They claim that - in the context of the Transcendental Deduction Kant invokes the *numerical identity*, as opposed to the mere *unity*, of apperception because he thought he could deduce certain conditions of experience (which the manifold of appearance has to fulfil) from the identity of apperception which he thought he could *not* deduce from its mere unity. I shall not try to examine the textual evidence for such a contrast (which to me seems fairly slim). However, in the light of what was said above it seems to me extremely doubtful that in the context of the Transcendental Deduction the phrase 'numerical identity of apperception' means anything different, or is supposed to establish anything more, than the mere 'unity of apperception. Such a contrast between the mere unity and the numerical identity of apperception would make Kant's Transcendental Deduction liable to the criticism of having committed a related, though more sophisticated, 'paralogistic' mistake which he criticises so well later in the *Critique*. I shall not go into the matter³⁹. Instead, I shall consider a different justification of Kant's use of the notion of numerical identity with respect to the 'subject of thought', a justification that we

³⁸Cf. notably Henrich, *Identitaet und Objektivitaet*, esp. pp. 55-58, and his 'The Identity of the Subject in the Transcendental Deduction', where he attempts a detailed critical reconstruction of the transcendental deduction, articulating first the conditions which have to be met, and then developing a step-by-step argument which fulfills them in a way such that the 'I think' of the deduction can appear as a subject with self-conscious identity over time and personhood.

³⁹Put in a nutshell, I should state my qualms as follows: since, in the Paralogism-chapter, Kant shows that the numerical identity of apperception has to be understood in a purely logical sense, the drawing of a contrast between the unity and the identity of apperception in the context of the Transcendental Deduction would either have to take recourse to an '*empirical*' notion of identity of the subject, or (alternatively) take recourse to an '*apriori*' notion of numerical identity of the subject. The problem with interpreting Kant's talk of the 'identity' (as opposed to the mere 'unity') of apperception in an '*empirical*' way (cf. e.g. Bennett, *Kant's Analytic*, pp. 117-125) is that it cannot do justice to the theoretical 'dignity' (A91) of the Deduction. The problem with invoking a sense of numerical identity of the subject which is neither empirical nor equivalent to the mere unity of apperception (cf. Henrich 'The Identity of the Subject in the Transcendental Deduction', esp. pp. 261-279) is how to establish a sense of 'identity of the subject' which evades the paralogistic trap. For reasons of space, I cannot go into the matter here.

already envisaged above.

I said above that although the pure perspective of consciousness provides for the idea of particularity of the subject of thought (and thus for the contrast between numerical and merely qualitative identity), it cannot by itself fill it in. And we concluded that - - because of this necessary shortcoming -- the pure perspective of consciousness is *essentially* dependent on the objective, or third-person perspective, to give substance to this contrast, and that is, provide criteria of identity of the entity referred to by 'I'. These criteria will then determine what kind of thing the 'I'-thinker is, and thus what constitutes his numerical identity *over time*. Kant himself may have thought that these criteria are in principle inaccessible to us, since they concern the identity of ourselves as we are 'noumenally' (i.e. as things in themselves). However, as Strawson has repeatedly pointed out, we need not follow Kant in this in order to make sense of the more intelligible aspects of his reflections on self-consciousness. The contrast between numerical and merely qualitative identity may equally well be filled in by the empirically applicable criteria of identity of ourselves as we know ourselves (i.e. as a certain kind of perceptible and persistent particulars, namely persons)⁴⁰.

Talking of the identity of the subject is thus in a way misleading. For it suggests the identity of a certain particular, or substance, which is not what is meant by it. The talk of the 'subject of thought' is rather a hypostazation of a particular ability that may be said to be essential to personhood, namely the ability to refer to oneself-as-subject (i.e. without application of any criteria of one's identity). We may, however, accept this talk as a borderline case of our identity-discourse. The identity of the 'subject of thought' -- understood in this sense -- can be compared to the identity of an 'extensionless point' (to which, according to Wittgenstein's *Tractatus*⁴¹, everyone of us 'shrinks' when he abstracts from all the determinations which constitute his identity as a particular person). Just as extensionless points cannot be distinguished from each other by their internal features (they have none), subjects of thought are subject to neither counting nor identifying rules. The notions of numerical identity over time and difference at one and the same time do not in a strict sense apply to them. And this is just to say that 'sub-

⁴⁰Cf. Strawson, *The Bounds of Sense*, p.164. This may have been the point of Kant's introduction of the 'outside observer' into the discussion of the Third Paralogism. Cf. B415: 'Its [the soul's] permanence during life is of course evident, since the thinking being (as man) is itself likewise an object of the outer senses'

⁴¹ Cf. Wittgenstein, *Tractatus Logico-Philosophicus*, 5.64.

jects of thought' are not concrete particular, but just a *façon de parler*, a way of talking about a certain aspect of such concrete particulars as persons indeed are.

The 'rational psychologist' misunderstands this *façon de parler*. He mistakes the 'logical identity of the subject' for a non-trivial, ontological identity. In this respect his mistake is similar to the one committed by the solipsist Wittgenstein attacks in the *Blue Book*, who tries to express his putative insight by claiming that 'when anything is seen..., it is always I who see it'⁴². Wittgenstein replies to his assertion: 'What should strike us about that expression is the phrase 'always I'. Always who? For... I don't mean: "always LW"'⁴³. In other words, by 'I' the solipsist does not mean, as Wittgenstein also puts it, 'a particular entity 'I''⁴⁴; he does not mean the particular entity that he *in fact* is. Wittgenstein makes the same point by saying that the solipsist uses the word 'I' without a 'grammatical neighbour'. By saying 'I', the solipsist does not intend to 'point to certain objects...as opposed to others'⁴⁵. And Wittgenstein admonishes us: 'if ... I believe that by pointing to that which in my grammar has no neighbours I can convey something to myself (if not to others) I make a mistake...'⁴⁶

The 'I' of transcendental apperception is similarly without a grammatical neighbour. When Kant says in the Transcendental Deduction that 'it must be possible for the "I think"' to accompany all my representations' (B131) -- since, if I could not combine them into one self-consciousness, 'they would not be one and all *my* representations' (B132) -- he does not mean that they would not be one and all the representations of a given particular, namely the particular I *in fact* am, RS. The 'I' and the 'my' are not supposed to individuate in this context. They are supposed to be understood from the pure perspective of consciousness, and that is, in abstraction from their genuinely individuating objective background. And this is perhaps what gives Kant's use of the first-person pronoun its 'transcendental twist' and makes it, at times, so difficult to grasp. Of course, this is not to say that it cannot 'convey' anything any more. Kant's 'epistemological ascent' from the individual person to the transcendental subject, the 'I' of the 'I

⁴²Wittgenstein, *The Blue and Brown Books*, p. 61.

⁴³Ibid..

⁴⁴Wittgenstein, *The Blue and Brown Books*, p.63.

⁴⁵Op.cit, p.71.

⁴⁶Op.cit., p. 72.

think', which 'in all consciousness [no matter whose] is one and the same' (B132), does however have one significant consequence: namely that -- strictly speaking -- it makes no sense to apply the notion of numerical identity to the transcendental subject (neither in its synchronic, nor in its diachronic application). Wittgenstein compares the solipsist's fruitless attempts to express his putative insight with the procedure of someone who 'constructed a clock with all its wheels, etc., and in the end fastened the dial to the pointer and made it go round with it'⁴⁷. In a similar fashion the theorist who mistakes the logical identity of the subject for an ontological, and thence informative and non-trivial, identity has overlooked that just as a clock whose dials are fastened to the pointer cannot tell the time, the logical identity whose identity is a trivial upshot of the coherence of consciousness, and thence cannot by itself determine the identity of the person *whose* consciousness it is. The logical identity of the subject, as it were, falls through itself. Or, to use another metaphor, there are no possible gaps between different instances of a subject's 'I'-thoughts; there is here no possibility of non-identity. Any attempt to gain a non-trivial, substantial sense of identity of the subject by artificially stretching its identity over a temporal extension must fail. For as soon as -- *per impossible* -- we try to apply any actual criteria of identity to this temporally extended pseudo-entity, its identity 'snaps together' again into an extensionless point.

4.3 *Experiential Memory and Personal Identity*

I think that Kant's notion of a purely logical identity of the subject can find a useful and elucidating application in discussions of the relation between personal identity and what has been called 'experiential memory'⁴⁸ or 'memory from the inside'⁴⁹. In contrast with *factual* memory, which may be defined as the retention of previously acquired knowledge (e.g. remembering that the battle of Hastings was in 1066), *experiential* memory is conceived as the memory of events *witnessed* or *experienced* by the remembering person. It is characteristically expressed in the form 'I remember x's ϕ -ing.' (as opposed to 'I remember *that* x ϕ -d'). The experiential memory of *one's own*

⁴⁷Op.cit., p.71.

⁴⁸ Cf. Wollheim, 'Memory, Experiential Memory, and Personal Identity'.

⁴⁹ Cf. Shoemaker in Shoemaker/Swinburne, *Personal Identity*, p. 105.

experiences and actions (typically expressed by statements of the form ‘I remember (myself) ϕ -ing’) is obviously a subgroup of this particular kind of memory. It is *first-person* experiential memory that we will be exclusively concerned with in what follows, since it is this kind of memory that -- by making, as Locke put it, ‘a man be himself to himself’⁵⁰ -- has been thought to provide a criterion for personal identity. The question whether first-person experiential memory can function as a criterion of personal identity will be discussed in more detail below (sect.5.4.1.). In the present context I would like to discuss a related claim which, if true, would render viciously circular from the very start any attempt to give an account of personal identity in terms of experiential memory. This is Gareth Evans’ claim that

‘if a subject has, in virtue of the operations of his memory, knowledge of the past states of a subject, then that subject is himself’⁵¹.

Evans’ thesis is directed against the view (endorsed by, amongst others, Parfit and Shoemaker) that first-person experiential memory is a way of possessing knowledge about an object which leaves open the identity of that object. He accuses such a view of inferring wrongly from what holds true in the outlandish cases of fissions and fusions described by Shoemaker and Parfit in connection with their discussions of *quasi-memory*⁵², to what holds true in the normal case of memory which does not have to take into account such abnormal circumstances. Evans’ own view, as I understand it, is that first-person experiential memory necessarily presupposes, and provides a subject with knowledge of, the persistence of a *particular object*⁵³.

Now, I think, we can grant Evans that judgements about oneself based on experiential memory do not -- under normal circumstances -- contain an ‘identification-component’ in the sense that the content of such a memory-claim could be split up into two different cognitive components, one telling me that ‘someone experienced x’ and the other one that ‘this someone was me’⁵⁴. That is to say, I do not normally have to

⁵⁰Locke, *Essay Concerning Human Understanding* II, xvii.9.

⁵¹ Evans *The Varieties of Reference*, p. 245.

⁵² Cf. Parfit, ‘Personal Identity’, esp. pp. 15-17, and Shoemaker, ‘Persons and their Past’, esp. pp. 24-49.

⁵³ Cf. Evans, *The Varieties of Reference*, pp. 237-248, esp. p. 246, note 64.

⁵⁴ Cf. Evans, *The Varieties of Reference*, p.237.

infer from my remembering something ‘from the inside’ to its having been experienced by *me*. That it was *I* who experienced what I experientially remember is indeed something that, as one might put it, is already ‘built into the very content’ of an experiential memory. (That is to say, if I was asked to say *what* I thus experientially remember, I would typically respond with a statement of the form ‘*I* ϕ -ed’ or ‘I remember (myself) ϕ -ing’). Nor will we want to deny that first-person experiential memory-claims are ‘immune to error through misidentification’, i.e. that, as Shoemaker puts it, ‘where the present tense version of a judgment is immune to error through misidentification relative to the first person pronouns this immunity is preserved in memory’⁵⁵. But from all this, I shall argue, it does not follow that experiential memory by itself can guarantee the continued numerical identity over time of a certain *particular*. First-person experiential memory, I shall suggest, is immune to misidentification for the simple reason that it does not contain an identification-component *at all*.

Shoemaker’s introduction of the notion of *quasi-memory* is best understood as a removal of what he calls the ‘previous awareness condition’⁵⁶ from our normal concept of memory. Shoemaker acknowledges that we would not normally say that someone remembers experiencing or doing something if it was not *himself* who experienced or did what he thus claims to remember. But this restriction, Shoemaker argues, can be lifted in order to frame our new concept of quasi-memory, which does not presuppose the identity of the remembering person with the referent of the ‘myself’ appearing in the expression of the content of the memory⁵⁷. In his essay ‘Personal Identity’ Parfit concluded from the conceivability of the notion of quasi-memory that

‘it cannot be part of what I seem to remember about this experience that I, the person who now seems to remember it, am the person who had this experience’⁵⁸.

⁵⁵ Shoemaker’s definition in ‘Persons and their Past’, p.21.

⁵⁶ Shoemaker, ‘Persons and Their Past’, p. 24.

⁵⁷ Cf. ‘Persons and Their Past’, pp. 34-39. Shoemaker sometimes speaks as if, by introducing this new concept of quasi-memory, he has also demonstrated the possibility of a new faculty of gaining knowledge of the world. Shoemaker claims that this can be achieved by adding a *causal* requirement to the notion of quasi-memory (cf. op.cit., p.24), which he thinks is already implicit in our present notion of memory. Evans (I think correctly) criticises this particular tendency of Shoemaker’s account; cf. Evans, *Varieties of Reference*, p.248. (Cf. also the discussion below, sect.5.5.1.)

⁵⁸ Parfit, ‘Personal Identity’, p. 15.

Similarly, Shoemaker has argued that, with respect to memory from the inside,

'one's self enters into the content of what is directly known, not as the subject of the past action..., but as the subject of the present remembering.'⁵⁹

Evans criticizes these conclusions for -- he thinks, wrongly -- assuming that 'the identification which 'I' effects (for me now) is an exclusive present-tense identification'⁶⁰. Evans' view is that, on the contrary, 'it is of the essence of an 'I'-Idea that it effects an identification which spans past and present'⁶¹. I shall not attempt to discuss the, fairly complicated, reasons Evans adduces for this claim (which would have to discuss Evans' complex theory of 'ideas'⁶². Let us instead approach the matter from the Kantian perspective outlined above.

It is, of course, true that *what* one remembers in an experiential memory is *oneself* doing or experiencing something. That it was *I* who experienced what I experientially remember is indeed something already 'built into the content' of an experiential memory -- be it a veridical or an only apparent memory. For if I was asked to recount what I thus remember my report would typically be of the form 'I remember (*myself*) ϕ -ing'. This, it is important to note, would also be true for a *quasi* experiential memory, which is not *necessarily* a memory of what was done or experienced by the *same* person as the one who now remembers having done or experienced it. (Only by gaining further information can I come to know that it was in fact only a the 'quasi memory donor' who experienced what I quasi-remember. But this would not affect the 'content' of my memory). Now if the use of 'I' in the expression of the content of a first-person experiential memory involved the identification of a certain particular object, then Evans would indeed be right in his claim that the notion of quasi-memory is fundamentally incoherent. For then the identity of the particular person who *has* the given memory with the particular person who is the 'intentional object' of this memory (i.e. the person the memory is *of*) would already be presupposed by the very content of such a memory. Only A (the person having the memory) could (logically) have had A's memory of (*him-*

⁵⁹ Shoemaker in Shoemaker/Swinburne *Personal Identity*, p.105.

⁶⁰ Evans, *The Varieties of Reference*, p. 246.

⁶¹ Ibid.

⁶²Cf. the definition in Evans, *The Varieties of Reference*, p.104. In general, it seems to me that Evans' account of 'self-identification' suffers from his attempt to assimilate self-reference to reference to objects.

self=A) swimming in the Pacific Ocean. For since the very content of the given experiential memory ('*what* is remembered') individuates A as its owner, only A could possibly grasp the content of *this* particular memory of *himself* swimming in the Pacific Ocean. And thence only A could have been the person who actually swam in the Pacific Ocean. To put it differently: If I could 'grasp the content' of any of *my* first-person experiential memories only if what I thus remember was in fact experienced by the person (the particular entity) that *I* am, then identity of the remembering with the remembered person would indeed be guaranteed. But it would be *trivially* guaranteed. And the fact that it would be *trivially* guaranteed would at the same time prevent experiential memory from functioning as a *criterion* of identity. For if only *I* could have had *this* memory of swimming in the Pacific Ocean (for only I can grasp its content), only I could possibly have done or experienced what I thus remember. This observation may be viewed to bring out the truth in Reid's and Butler's indictment of the so-called memory criterion of personal identity as being irremediably circular (cf. below, sect. 5.4.1.). For, according to the conception of experiential memory just sketched, the identity of the 'subject' of an experiential memory would not only be indexed to the person having the memory, but -- via the 'content' of the memory -- to the person who actually experienced what is thus remembered; which would make it trivially true that a person with *this* particular memory (or set of memories) would have to be *this* particular person⁶³.

I think it is evident from the previous discussion where this view of experiential memory goes wrong. It goes wrong in its crucial premise that the use of 'I' in the expression of a first-person experiential memory involves the identification of a certain *particular object*. One might say that it is, on the contrary, just the fact that the use of 'I' in the expression of the memory does *not* involve an identification of a certain particular that makes it into an experiential memory in the *first-person* (rather than the *third-person*) 'mode'. In other words, the identity of the remembering person does not

⁶³Reasoning along these lines, Chisholm has arrived at the conclusion that, although character-traits, beliefs, and even memories could be 'transferred' from one person to another, first-person experiential memories could (logically) not; cf. Chisholm, *Person and Object*, p.107: '...my memories could even be said to be transferred from me to you, if what you remember, or think you remember, the same things I do. (But if I remember *my* doing the deed, the content of that memory could *not* be transferred to you).' Cf. the discussion of Chisholm's 'Simple View' of personal identity below, section 5.7..

only *not* enter into the content of a first-person memory qua *remembering* person; it does not enter into the content of the memory *at all*. And this also shows why -- in the sense in which two persons' memories can be identical at all -- *my* first-person memory of *my* swimming in the Pacific Ocean can be just the same as *your* memory of *your* swimming in the Pacific Ocean. And, at the same time, it shows why first-person experiential memory cannot be said to guarantee the persistence of a given particular object. Using Kant's terminology we might make this point by saying that the 'logical identity' of the remembering person with the 'subject' of the experiential memory is indeed built into the content of any experiential memory of mine (quasi or non-quasi). It is trivially true that it was *I*, the 'logical subject' of this memory, who did or experienced what I thus remember. But this logical identity of the subject does not by itself guarantee numerical identity of person. Experiential memory does not guarantee the *persistence* of a certain determinate particular, since it does not involve an *identification* of such a particular. The 'I' in a description of the content of an experiential memory can be compared to an 'idle wheel'. Once again nothing turns on it.

4.4. *Subjective Imagination and Nagel's 'Objective Self'*

A number of related points can be made with respect to a distinction between two different kinds of imagination, which I shall call, respectively, *subjective and objective imagination*. The distinction between these two ways of imagining was first drawn by Bernard Williams (although not under this label) and then elaborated by Zeno Vendler and Richard Wollheim⁶⁴. One convenient way of introducing it -- although a way not essential to the distinction itself -- is in terms of two different ways of understanding an invitation of the form 'Imagine yourself ϕ -ing!'. To take Vendler's example⁶⁵: there are two clearly distinguishable ways of complying with the request 'Imagine yourself swimming in that water!' (made, e.g., by someone standing next to me on a high cliff staring down into the water beneath).

(a) I can imagine myself, i.e. *RS* (the person I am), swimming in the water below. This is typically done by visualizing me, *RS*, swimming in the water, i.e. by

⁶⁴ Cf. Williams, 'Imagination and the Self', pp. 26-46; see esp. p. 42; cf. Wollheim, *The Thread of Life*, pp.73-75; cf. Vendler, *The Matter of Minds*, ch. 3.

⁶⁵ Vendler, *The Matter of Minds*, pp. 43-4.

imagining seeing RS swimming there (preferably, but not necessarily, from above). This would be a clear case of *objective* imagination.

(b) I can imagine (*myself*) swimming in the water. (In this second case the 'myself' can be omitted without any loss of meaning.) What I then imagine is not perceiving myself as a particular person, i.e. RS, swimming in the water. Rather I imagine *having the experiences* of someone (not necessarily RS) swimming in that water. That is to say: I do not imagine a specific person (myself=RS) having these experiences, but simply the experiences without, one might say, any 'identity-index'. I imagine the situation, as it were, 'from within' (not, as in the first case, 'from outside'). I imagine it -- we can say -- *subjectively*.

Unlike in the first case, in the second case my identity as a particular person does not enter into the content of my imagination. For even though there is a perfectly legitimate sense in which -- when imagining the situation subjectively -- I can be said to imagine *myself* swimming in the water, this is not a sense in which I can be said to imagine a particular person swimming in the water. The 'myself' in a description of the content of a subjective imagination does not have an individuating role. 'I' and 'myself' are not here used to distinguish the imagining person from all other entities.

This becomes even clearer when we consider more extreme cases of subjective imagining, such as cases of imagining oneself being another person than the person one is. With respect to such 'feats of transference' the view that by 'I' (or 'myself') in the expression of the content of such an imagination a person identifies a particular entity, would lead to an indeed absurd consequence: I can imagine myself being, or rather having been, Napoleon. What is involved in this imagining? How do I bring about this feat of imagination? Certainly not by imagining RS being Napoleon, since RS being Napoleon is an *obvious* logical impossibility. It is logically impossible for any particular person -- except, of course, Napoleon -- to be Napoleon. Now if the 'myself' in the imaginary state of affairs 'myself being Napoleon' were interpreted as identifying a certain individual, then -- since, I take it, it is impossible to imagine an *obvious* logical impossibility -- my imagining being Napoleon would itself turn out to be impossible, which it obviously is not. Therefore, what I imagine when I imagine myself being Napoleon cannot be a particular person being Napoleon (nor, more generally, can it be a particular entity, such as a 'transcendental subject', being Napoleon). My identity as a particular, individual entity does not enter into the content of subjective imagination. What I imagine is better described as experiencing the world from Napoleon's *point of*

view, i.e. seeing what he saw, feeling what he felt, etc.. In order to be able to imagine being Napoleon at the battle of Waterloo (e.g., in order to be able to write a 'stream of consciousness'-novel about what went on in him at this event), I have to know quite a bit about Napoleon: his education, his character, the history of his time, and, more specifically, the mood he was in at the day of the battle of Waterloo, his local position in the battle, and many more similar details. Such an extensive knowledge, however, would not be required for an elementary 'impersonation'⁶⁶ of Napoleon. A person's imagining being Napoleon is a matter of degree, not a matter of all or nothing. It can be more or less detailed, more or less successful. For it involves a whole cluster of various acts of subjective imaginations, not -- *per impossibile* -- imagining one individual being another.

I think that the correct interpretation of the phenomenon of subjective imagination can find a useful application to the 'philosophical puzzle'⁶⁷ which Thomas Nagel (in his book *The View from Nowhere*) has alleged to find in the truth that *he* is TN. Nagel confesses that he finds it puzzling how *he* (as what he calls the '*objective self*', i.e. as the pure thinking subject of the all-encompassing 'objective conception' of the world) can be identical with *TN* (a particular, limited entity *within* the world). He finds it puzzling how he as the subject of this conception which 'might as well view the the world from the perspective of a different person' (p.60) could be tied down to anything as particular as the person TN. The thought-experiment by which Nagel wants to induce an analogous astonishment in his readers consists in an absolute *abstraction* from all the determinations which constitute him as a particular entity (a person) within the world, followed by an equally absolute *self-identification* with this entity. I shall argue that the interpretation Nagel gives to this thought-experiment rests on a paralogistic misconception of subjective imagination of the kind sketched above.

Let us have a closer look at what is involved in Nagel's stunning feat of philosophical imagination. It consists basically in two subsequent steps: In a first step Nagel clears out all empirical content from his conception of himself, thus abstracting from all

⁶⁶ Cf. Williams comparison of subjective imagination with 'impersonation', in 'Imagination and the Self', pp.40-45.; similarly Wollheim, *The Thread of Life*, pp. 76-78.

⁶⁷ Nagel *The View from Nowhere*, ch.4: 'The Objective Self', p. 57. All references in the text are to this book.

the properties which constitute him as a particular object within the world. In a second step, the reversal of the first, he re-locates this 'objective self', which he has just abstracted from the person he is, back into the world. He -- in Nagel's words -- 'identifies himself' (p.60) with TN. He, as it were, draws the contentless, pure unity of consciousness, which is conceived to be weightlessly hovering above the world, back down into the world again and places it into the particular person he is.

Represented in terms of our distinction between subjective and objective imagination, Nagel's thought-experiment can be described in the following manner: When Nagel imagines the world from the 'centreless' or objective point of view (a world which contains besides many other things the person TN), then he imagines himself objectively; -- not in a particular situation, such as swimming in the water, but in the most radically objective, and thence *detached*, way possible. When, in the second step, he 'identifies himself' with TN, then he subjectively imagines (himself) being TN. That is to say, he does not, as in the first step, *objectively* imagine *TN being there*, alive and kicking, in the world, but he *subjectively* imagines (himself) *being TN*. He imagines experiencing TN's experiences, or -- as Nagel would prefer to say -- experiencing the world from TN's point of view.

Nagel professes to find a 'metaphysical mystery'(p.56) hidden in his thought 'I am TN'. It seems to me that although it would take a philistine to deny that there are many things mysterious to individual human existence the fact that Nagel wants to express by saying 'I am TN' seems hardly one of them. In particular, it is, I think, Nagel's failure to acknowledge the non-individuative role of the first-person pronoun in the description of the content of subjective imagination that makes it appear to him that there is a mystery embedded in his being who he is. And this failure on Nagel's part seems just a particular manifestation of the deep-rooted paralogistic illusion that 'I', after having been deprived of all the usual background-conditions of its application, can still be used in a way which fixes a referent.

Nagel is not entirely unaware of the fact that the 'I' in his 'philosophical' understanding of the thought 'I am TN' cannot be taken to achieve a genuinely identifying singular reference. For, *on the one hand*, he denies that the objective self is to be conceived as an entity which is 'distinct' from the particular person he, TN, is. As he puts it:

'The objective self...is not unique: each of you has one. Or perhaps I should rather say each of you is one, for the objective self is not a distinct entity.' (p.63).

On the other hand, however, Nagel still wants to interpret the 'I' of the philosophical thought 'I am TN' as making a pseudo-reference to a pseudo-entity, an entity whose identity is supposed to be determined only by its being 'the subject of the impersonal conception of the world' (p.64). Nagel calls it TN's 'real me', which just happens to 'occupy' the publicly identifiable person TN⁶⁸.

Now it seems to me that this putative entity is an ontological hermaphrodite; it is not a particular entity within the world, but nor does it (somehow) exist outside it. In a way the 'objective self' is only the hypostazation of the objective conception of the world, since it is 'incapable of being anyone in particular' (p.61). But it is nevertheless supposed to be 'something real' (p.66)⁶⁹ -- after all it is what *I* am, it is, as Nagel puts it, 'this thinking subject' (p.60). There is not *one* objective self for *all* of us (Nagel's objective self is not a world soul), but there does not seem to be *one* for *each* of us either (for how could two different objective selves be distinguished?). Thus, *on the one hand*, Nagel wants to hold that objective selves are particular entities, and that there are just as many of them as there are persons. For, after all, he somehow has to account for the truth that 'I am TN' is an *identity*-statement (although a very peculiar one), which is true when thought by TN, and false when thought by me. And from this it would seem to follow that, since I (RS) and TN are different persons, we cannot have the same objective selves. *On the other hand*, however, Nagel wants to treat the objective self rather as a certain 'aspect' of particular persons, namely their capacity to conceive of the world objectively (which is the same for all of them), a form in which every being capable of having 'I'-thoughts necessarily partakes⁷⁰. But with regard to this reading the question arises how such a general form or aspect could be *me*.

I therefore conclude that the interpretation which Nagel forces on his 'I am TN'-thought must be flawed. The misconception from which Nagel's astonishment about his being a particular person arises is basically nothing different from the familiar Cartesian fallacy of moving from an epistemological to an ontological privacy, i.e. of inferring

⁶⁸ Cf. op.cit., p.60/1: 'The real me occupies TN, so to speak; or rather the publicly identifiable person TN contains the real me. From a purely objective point of view my connection with TN seems arbitrary.' It is useful to ask here: *whose* connection with TN?

⁶⁹ Cf. also p. 65: 'The objective self functions independently enough to have a life of its own.'

⁷⁰ Cf. op.cit., p.65.

from the truth that, for every thinker, there is a way in which only he can refer to the entity he is to the falsehood that there is a special entity (his 'real me') to which only he can refer. Nagel realizes that in referring to himself in the 'I'-mode he does not refer to himself *as* TN, i.e. the particular limited person he in fact is. But he 'paralogistically' concludes from this that the pronoun 'I' in his philosophical thought must have a separate referent different from TN. Just as there are only two persons involved in my imagining being Napoleon (namely Napoleon and RS)⁷¹, there is only one person involved in my (alternately) imagining being, and not being, RS. And this person is RS. There is no need for an obscure Nagelian 'objective self'. In saying this, I do not want to deny that what might be called the capacity of '*absolutely subjective imagination*' (i.e. a person's ability to abstract from *all* the determinations which constitute him as a particular, limited being within the world) shows something significant about human beings⁷². It may even be an essential ingredient of personhood. But it hardly warrants the sweeping ontological conclusions Nagel wants to draw from it. As it stands, Nagel's notion of the objective self is not more than a misguided reification of the (distinctively human) capacity of 'absolutely subjective imagination', a most sophisticated variety of the classical paralogism of 'confusing the unity of experience with the experience of a unity'⁷³. There is no 'real me', neither one for all of us, nor one for each of us. Although there is a way in which only I can refer to myself, there is no entity to which only I can refer.

[It seems to me that a similar analysis could be given to the putative problem that the indexicality of our speaking and thinking about the world is sometimes seen to

⁷¹Cf. Williams, 'Imagination and the Self', p. 45.

⁷² Nagel's problem about his identification with TN finds a counterpart in the inability to subjectively imagine one's own death. I can easily imagine RS's death (I can, e.g., imagine RS's corpse being buried, etc.). But there is a sense in which I cannot really imagine *my* death. That is, I cannot subjectively imagine my being dead. For this would -- *per impossibile* -- be a case of subjective imagination in which *no experiences* at all are imagined. It is significant that we do not generally take this inability of ours as proof that we can survive the deaths of the particular persons we are. I suspect that considerations of this kind may also have lead Wittgenstein to claim, in *Tractatus* 6.431, that 'in death the world does not change, but cease.' If my identity were indeed that of an 'extensionless point' (5.64), then my death would not make the world any lighter. It would not be a further fact over and above all the facts which constitute the world. Cf. also Chapter 6.

⁷³ Strawson, *The Bounds of Sense*, p.162.

pose for any purely 'objective' or 'absolute' conception of the world and our position within it. Thus Castaneda has suggested that we have to admit of the possibility of 'essentially indexical facts' (such as that *I* am RS, that it is *now* 10pm, 1/4/1992, or, less convincingly, that Oxford is *here*) which are defined by their relation to the 'I think'⁷⁴. And, similarly, Nagel maintains that there are truths that can be grasped only from a certain 'perspective', or 'point of view', but which are nevertheless 'ineliminable' from any complete conception of the world, -- the paradigm case of such an 'essentially subjective' truth being the by now familiar 'I am TN'⁷⁵. This is, Nagel claims, a truth which, although of the utmost significance for TN, cannot possibly be captured by any purely objective view of the universe, however encompassing; it is a truth for which 'there is no room in the world conceived as simply there, and centreless'⁷⁶. It rather has to be understood as expressing an 'irreducibly first-person fact'⁷⁷, i.e a fact which can only be known from the 'perspective' of a particular person, namely TN; a fact which -- *paradoxically* -- at the same time must, but cannot, form part of the all-encompassing objective view of the world.

Perhaps we could, tentatively, say that an understanding of the three 'essential indexicals' *in combination* is constitutive of what it means for a given subject or person to 'be in the world'. But it seems hardly necessary to describe this 'being in the world' in terms of a certain self or subject occupying a certain *perspective*, or *point of view*, on an 'objective' world. In fact, the optical metaphor of a given perspective can in this context be positively misleading. It is true that the point defined by 'I-Now-Here' is for every person, at every time, in every place, the starting-point from which he has to trace his epistemic path (his 'experiential route'⁷⁸) through the world. It is, in a way, the *point of origin* of the coordinate system in which all our identification of things, times and places takes place. It is, as it were, the point where the 'epistemic grid' which we lay over the world -- in order to find our way through it and act according to its demands -- is anchored. This manifests itself most distinctly in the fact that even the

⁷⁴Cf. Castaneda in 'The Phenomeno-Logic of the I', *passim*, and 'The Self and the I-Guises', esp. pp. 120-134. Cf. also McGinn's claim that 'all indexical modes of presentation go back to, and are anchored in, conscious presentation of the subject of the indexical thought in question' (McGinn, *The Subjective View*, p.17.) and that it is the 'definite relation to the self' (*op.cit.*, p.19) built into the content of all indexical thoughts which accounts for the special subjectivity, or 'egocentricity', of these thoughts.

⁷⁵ Cf. Nagel, *The View from Nowhere*, p. 57, and his earlier work on the topic.

⁷⁶ Nagel, *The View from Nowhere*, p.56.

⁷⁷*Op.cit.*, p.57.

⁷⁸ This is Strawson's expression, cf., e.g., *The Bounds of Sense*, p.128.

completely impartial way of locating a place in terms of degrees of longitude and altitude can make sense to a subject only if he knows where the place which functions as the zero-point of this system of measurement (e.g., in our case, Greenwich) is from *here*, i.e. where *he* is located in relation to this place. And equally, the location of an event in time by assigning it a certain date according to a certain calendar is for its practical viability dependent on our knowing how far in the past or future the date which we have chosen as the point of reference of our system of measurement (e.g. the year zero) is from *now*. Identification of objects in time and space is thus intrinsically, it may be held, identification in relation to *oneself*. However, this should not be taken to suggest that we should conceive of the entity picked out by 'I-Now-Here' as a kind of 'subjective centre' within the objective world, i.e. the eye-point of a given perspective on the spatio-temporal objective world. For the artificial expression 'I-Now-Here' cannot by itself be taken to identify, or single out, a determinate particular (although it may well be said to 'pin it down'). Just as I do not have to know *who* I am in order to refer to myself-as-subject, I do not have to know *what time* it is in order to refer to the present time by 'now', or *what place* I am in in order to refer to this place by 'here'. Just as by saying 'I' I do not *identify* myself (for myself), by saying 'now' I do not *identify* a certain time, and by saying 'here' I do not *identify* a certain place.

'I'-statements are -- as all other indexical statements -- indeed essentially and irreducibly subjective as to their characteristic meaning (or 'conceptual content'), but they do not therefore express essentially subjective facts. Just as there are no essentially private 'I'-propositions which can be grasped by only one thinker, there are no essentially *subjective* 'now'-propositions or 'here'-propositions', which can be grasped only from a certain point of view. And this implies that they do not pose any problem for a purely 'objective' conception of the world -- whatever this may mean precisely. What is the case if an 'I'-statement is true can be restated in 'objective' terms without further difficulties simply by replacing the 'I' in question by a non-indexical expression identifying the 'worldly' person who used 'I' in order to refer to himself. Thus, just as there is a way in which only I can think of myself, but no entity only I can think of, there is a way in which only I can express the fact that I am RS, but no fact which only I can express. 'I'-thoughts (as all other indexical thoughts) are indeed irreducibly subjective, but they are also only subjectively irreducible.]

5.SELF-REFERENCE AND PERSONAL IDENTITY

In the previous two chapters I tried to dispell some of the 'paralogistic' illusions to which the epistemological peculiarities of a person's reference to himself-as-subject give rise. I argued that a person's self-reference cannot be held to presuppose, or require, identifying knowledge of the particular thus referred to, in the double sense that it neither requires any knowledge of *what kind* of thing is thus referred to, nor any ('discriminating') knowledge of *which* particular is thus referred to. A person's self-reference, although it may -- from an objective point of view -- be said to 'pin down' a certain entity (namely, due to the semantic reference of the pronoun 'I', the self-referring person), does not -- from the point of view of the self-referring person -- involve an acquaintance with, or grasp of, a certain entity. My knowledge that *I* am ϕ is not for me an answer to the question *who* is ϕ .

The 'paralogistic' illusion is based on a misinterpretation of these epistemological features of a person's reference to himself-as-subject. It misconceives self-reference as an identifying reference to an entity *sui generis*, a simple, determinate substance, called the 'self' or 'subject'; an entity which is conceived as constituting a person's core, and whose persistence is then claimed to be what a person's identity over time consists in. Following Kant, I argued that such a view of self-reference mistakes the cited epistemological features of self-reference for ontological features of its referent.

Having thus rejected the view that a person's self-reference determines a special referent whose criteria of identity are more determinate than those of the self-referring person may appear to be, I went on to consider a weaker claim -- namely the claim that a person's self-reference, although it cannot reveal to the person what constitutes his essence, can nevertheless provide him with a 'privileged access' to what constitutes his identity over time. I argued that this claim is based on a similar paralogistic illusion, namely the idea that a person can at any given time, as it were, fix his identity for all past and all future simply by thinking 'I'. In my attempt to lay bare the hidden workings of this illusion I fell back upon Kant's notion of a purely 'logical identity of the subject', which as such cannot guarantee personal identity. It cannot, that is to say, guarantee the *persistence over time* of a certain determinate *particular* (such as a person).

I ended my discussion by applying these points to the interpretation of various psychological phenomena exemplifying 'first-person modes of presentation', and to some general problems of indexicality.

It is now time to take a more straightforward approach to the so-called 'puzzle-

cases' which have formed the topic of recent discussions of personal identity, and to scrutinize how the abovementioned misconceptions concerning self-reference have influenced and, at times, distorted their interpretation. Before we proceed to this task, however, it will prove useful to step back for a moment and reflect in a more general and methodological way on the theoretical role which thought-experiments can play in answering the question of what personal identity consists in.

5.1. *Thought-Experiments*

The nearly entire lack of any methodological reflection on the theoretical role and status of thought-experiments has long been a major *lacuna* in contemporary treatments of personal identity and has only very recently been brought into (mainly critical) focus. Thus in her book *Real People* Kathleen Wilkes has claimed that thought-experiments are 'highly misleading as a philosophical tool'¹, since firstly, 'they cannot be realized in the real world'², and secondly, because in imaginary cases that could not possibly be realized in real life, our intuitions become 'increasingly dubious, uncertain and contestable'³.

Against Wilkes' *first* criticism Paul Snowdon has, I think successfully, pointed out that although thought-experiments are not real experiments with real objects being acted upon, they do not purport to envisage situations that could not possibly be realized in the real world⁴. It is, on the contrary, just their possible factuality that endows the imagined scenarios with their (at times) disturbing quality. In other words, the touchstone for the validity of thought-experiments is not whether they are physically, but whether they are *logically* possible.

Wilkes' *second* objection to the method of thought-experiments seems more serious. Our intuitions are indeed often deceptive and likely to point us in the wrong direction. And we would have no reason to believe that the intuitions evoked in us by

¹Cf. Wilkes, *Real People*, pp. 1-48. In his article 'Human Beings' Mark Johnston pursues a similar line. However, I think that his particular line of criticism has been successfully rebutted by Paul Snowdon in 'Personal Identity and Brain Transplants', pp. 123-125, and thence I will not consider it here.

²Wilkes, *Real People*, p.2.

³Wilkes, *Real People*, p.47.

⁴Cf. Snowdon 'Personal Identity and Brain Transplants', pp. 118ff.

the consideration of puzzle-cases could teach us anything valid about ourselves and our true nature, if questions of our true nature were to be understood in an empirical, or, alternatively, a deep metaphysical sense (which would have to be elucidated first). If, that is to say, the necessities which are supposed to be revealed by certain thought-experiments were meant to be of a *nomological* (or *Kripkean*) rather than *conceptual* kind, then there would be no reason whatsoever to assume that our intuitions concerning these thought-experiments could have any validity as to the question what personal identity 'consists in'⁵. However, Wilkes' objection does not, it seems, get a hold against theories of personal identity which, as I suggested above (chapter 2), understand the question what personal identity 'consists in' as a question to be answered by conceptual analysis, i.e. a question to be answered by an elucidation of the notion of numerical identity as applied to persons. For such a theory may insist that, although thought-experiments cannot tell us what a person *actually* can and cannot survive (e.g. various forms of lobotomy), they can tell us what we would *count as* a person's survival, i.e. *under what circumstances* we would count a person as having *survived* a certain change, and, equally, under what circumstances we would count a person as having *perished* (be these circumstances actual or only imagined). According to such a conception, what we typically do in inventing thought-experiments is present ourselves with a situation in which, by stipulation, we know the truth and falsehood of all the states of affairs that are normally used as (criterial and inductive) evidence in determining matters of personal identity (although of course not the identity-fact itself). And we then try to answer the pertaining identity-question. And since, by stipulation, we know the truth and falsehood of all the relevant criterial and inductive states of affairs, any indecision in our answering the identity-question cannot be ascribed to ignorance of certain facts, but only to an indeterminacy in, or inapplicability of, our concept of personal identity. Any indecision in our answering the pertaining identity-question can only be viewed to be

⁵ Thus Parfit seems to take the question in such a metaphysical sense when he distinguishes between purely semantic intuitions (which he describes as 'beliefs about words', i.e. beliefs about how we would use a certain word on a certain occasion) and intuitions expressive of our 'most natural assumptions about ourselves', which he describes as 'beliefs about ourselves' and our own nature (cf. Parfit, *Reasons and Persons*, p.200). I have to admit that I do not quite grasp the distinction. Certainly, whether we use a certain word or not -- and the word in question is of course 'the same person as' -- depends on whether we think it applies or not. And that is just a belief about what is the case, i.e. it is a belief about ourselves. Cf. the discussion of Parfit below, sect. pp. 5.7..

a conceptual matter, since what -- in a given imaginary case -- we would accept as the truth or falsehood of the pertaining identity-claim cannot depend on our knowledge or ignorance of certain *facts*. It must depend entirely on what we take the respective identity-claim to *mean*⁶. And thus the thought-experiment may be said to have contributed to an elucidation of the *concept* of personal identity. For by evoking our intuitions with respect to the given imaginary case (intuitions which are no more than reflections of the implicit 'grammar' of the concept in question), the thought-experiment forced us to make explicit our implicit criteria for the application of the concept. And it thus enabled us to isolate the 'core' of our notion of personal identity. That is to say, it enabled us to separate what is *essential* to a person's persistence through time from what is only *contingently* the case whenever a person persists. (Of course an adherent of a Platonic or 'realist' view of concepts, according to which the meaning of our concept-terms somehow transcends, or is fixed independently of, our *ability to explain what we mean* by using these terms, would not go along with these claims, since he would hold that an elucidation of the circumstances under which *we would count* a given concept as instantiated could never exhaust its meaning. However, I have argued above that Wittgenstein's rule-following considerations provide a good argument against any such view.)

It would be a misunderstanding to object to *this* conception of thought-experiments on the grounds that the scenarios envisaged in them will never *actually* occur. For, as Snowdon has put it, 'the normal view of language is that we have conferred upon its terms an interpretation in such a way that determines them as true or false of certain merely possible (but non-actual) situations'⁷. Thus the fact that the imaginary case is merely possible (but non-actual) cannot imply that there is nothing that is true, and false, of it. Rather, in evoking in us an intuition as to whether a given identity-claim would be true of the imaginary case if it *were* real, the thought-experiment teaches us what *criteria* we would apply in general in order to establish whether an identity-claim of the given kind is true or not. And the thought-experiment can thus be said to have

⁶In real life, of course, the situation is exactly reversed. There our ignorance of whether a given identity-claim is true or not is usually due to ignorance of certain evidential facts, i.e. facts which may be cited as (criterial or inductive) evidence in finding out about its truth.

⁷Snowdon, 'Personal Identity and Brain Transplants', p. 121.

contributed to an elucidation of the meaning of such identity-claims.

Thus *prima facie* there do not seem to be any good reasons for denying that thought-experiments can play an interesting and fruitful role in an elucidation of the notion of numerical identity as applied to persons. There is, however, a more general line of criticism of the conception of the role of thought-experiments just sketched. This line of criticism is based on more general considerations concerning the nature of our concepts and their origin in human life. Quine has presented this objection paradigmatically in the following passage:

'The method of science fiction has its uses in philosophy, but...I wonder whether the limits of the method are properly heeded. To seek what is 'logically required' for sameness of person under unprecedented circumstances is to suggest that words have some logical force beyond what our past needs have invested them with'⁸.

And in his article 'Entity and Identity' Strawson has given expression to similar reservations concerning the theoretical usefulness of imaginary cases. He says:

'I have remarked that a substance-kind concept involves a whole cluster of expectations of typical continuities and discontinuities and, one might add, of typical terminations, to which individuals of the kind are subject. History sometimes, and mildly, and ingenious philosophers more frequently, and starkly, present us with cases in which these expectations receive a certain sort of shock. The recipe for administering the shock is to contrive some striking dislocation of normal continuities, perhaps in such a way that we may seem to be presented with rival claimants to a single identity. It is of the essence of such cases that they should be quite outside the range of normal experience. Indeed, obviously, they could not have the shock-effect I have mentioned if they were not. Whatever is commonplace, or merely unusual, is taken in the stride of our normal conceptual apparatus; and the same would hold of the philosophers' imaginings if they *became* commonplace. They would be absorbed by conceptual adjustments: additions, refinements, decisions. We can amuse, and instruct, ourselves by imaginative anticipation of such adjustments; but we should recognize *this* exercise for what it is.'⁹

The lesson that we are supposed to learn from these remarks -- a lesson that both Strawson and Quine only intimate by way of irony -- may perhaps be spelt out in the following way: Admittedly, we do conceive of our terms as being able to determine reality (even only imaginary reality) save for a Yes or No. However, there are limits to this ability. As long as we stay within those limits, thought-experiments may well assist us in clarifying and elucidating those concepts. That is to say, they can help us in getting clear about what criteria we actually do invoke in determining whether a concept

⁸ Cf. W.V. Quine, reviewing Milton K. Munitz, (ed.) *Identity and Individuation*, p.490. Wiggins, *Sameness and Substance*, pp. 188/9 is equally suspicious of how well we really understand descriptions of imaginary cases.

⁹Strawson, 'Entity and Identity', p. 213f.

applies or not, and they can thus contribute to an elucidation of the concept. However, once we stray beyond the limits within which a concept can determine reality save for a Yes or No, we cannot expect this procedure of elucidation to yield any results any more. For we cannot expect to discover implicit rules for the application of a concept where rules have not yet been laid down.

This, however, is exactly what we do in some of the more *outré* thought-experiments concerning personal identity. We traverse the bounds of the applicability of this notion, but still expect it to be decidable whether it applies to the envisaged situation. And thence the intuitions evoked by thought-experiments which face us with 'unprecedented circumstances' that go beyond the limits of the applicability of our notion of personal identity cannot be taken to achieve an elucidation of our *present* concept of personal identity. At best these intuitions can be seen as instinctive and subliminal attempts at *concept formation*, i.e. as attempts at a conceptual adjustment of our present concept of personal identity to cope with the unprecedented circumstances envisaged in the thought-experiment. At worst they are simply unwary reactions to the way the envisaged situation is described. (And the addition, or subtraction, of even a single tiny detail to, or from, the imaginary story can often make all the difference in this matter and cause our intuition to go the other way.)

I think that this line of criticism of the elucidatory power of (at least some) thought-experiments is basically justified. And I shall try to substantiate this claim in my discussion of various 'puzzle-cases' below. In order to be a convincing objection to the theoretical value of the incriminated kinds of thought-experiments, however, the Strawson/Quine-objection will have to be clarified further. For neither Strawson nor Quine say anything about what draws the line between (*on the one hand*) those cases which are 'merely unusual' and which are 'taken in the stride of our normal conceptual apparatus' (Strawson), and (*on the other hand*) those which are genuinely 'unprecedented' (Quine) or 'quite outside the range of normal experience' (Strawson), and the occurrence of which is not provided for, and thence cannot be mastered by, our normal conceptual apparatus. What separates the 'merely unusual', with reference to which it is still decidable whether a given concept applies, from the genuinely outrageous, with respect to which it has become inapplicable? What *delimits* the range of applicability of our concepts? As long as we do not have any, if only vague and initial, answers to these questions the cited criticism remains empty.

The idea that there are limits to the applicability of our concepts, or that the

applicability of our concepts rests on the obtaining of certain *background-conditions*, is often associated with Wittgenstein's later philosophy. It is thence to him that we should turn for an answer to our questions.

According to Wittgenstein, concept-formation is only one aspect of human behaviour and as such. Like all human behaviour, it is based on, and responsive to, contingent aspects of the general course of events in nature. Concepts are not made in heaven; they are man-made and as such framed on the background of 'certain very general facts of nature'¹⁰. Unfortunately, Wittgenstein was not very explicit on the topic, and the scarce remarks he makes about it are scattered throughout his later work. However, even though Wittgenstein's views on the topic are admittedly fragmentary and unsystematic, a brief sketch of these views may, by contrast and analogy, help us in arriving at a more perspicuous notion of a background-condition of the applicability of a given concept.

Arguably, the 'very general facts of nature' alluded to in the *Philosophical Investigations* were meant to include both (a) facts concerning our *common biological nature* (i.e. shared desires, needs, wants; shared perceptual and other capacities; shared primitive psychological responses) and (b) certain *very general, but contingent, regularities* displayed by the course of events¹¹.

Our *common biological nature* may be said to *condition and restrict* our choice of concepts in the sense that we *could* not have certain concepts if we did not have certain capacities¹², or that we *would* not have certain concepts (or, as Wittgenstein preferred to put it, 'we would not play certain language-games') if we did not display certain common desires, aims, wants, needs, etc. without which the application of a certain concept would 'lose its point'¹³.

The *very general regularities* displayed by the course of events may be said to

¹⁰Wittgenstein, *Philosophical Investigations*, p.230.

¹¹ For the following sketch of Wittgenstein's treatment of the matter, cf. Baker/-Hacker, *Rules, Grammar, and Necessity*, pp. 234-238, and pp.318-338.

¹² Thus in *Zettel* §368 Wittgenstein mentions that we would not be able to participate in the language-game of people who distinguish different shades of yellow by means of a binary notation, for the simple reason that we lack their acute perceptual capacities: these people 'can do what we cannot do' (*Zettel* §368). Similarly, the blind cannot participate in our language-game with colour simply because they cannot employ the samples we typically employ in giving ostensive definitions of colour-terms.

¹³Cf. *Philosophical Investigations* §142: if our common behavioural manifestations of mental phenomena such as pain, fear, or joy were not relatively regular and uniform our normal language-games of ascribing pain, fear, or joy to people would 'lose their point' (§142).

form the background on which the application of concepts is based. If these regularities ceased to obtain many of our present concepts, Wittgenstein claimed, would become pointless. Many of the 'language-games' we presently play would go out of use. Thus, to cite at least one of the examples given by Wittgenstein: if we inhabited a world in which the colour of things changed constantly, or alternatively, a world in which everything was the same colour, we would find 'little or no use'¹⁴ for our present colour-vocabulary. Wittgenstein's reflections on those very general (but contingent) features of the world we inhabit, which form the *background-conditions* of the application of our concepts, are few and far between, and he does not always distinguish them very clearly from the restrictions placed upon our concept-formation by our common biological nature. Wherever these reflections appear, they usually do so in the context of considerations concerning Wittgenstein's thesis of the autonomy of grammar. That is to say, reflections on the background-conditions of the applicability of certain concepts are usually invoked in order to shake the grip of the preconception 'that certain concepts are absolutely the correct ones.'¹⁵, i.e. that our concepts correspond to something that lies deep in the nature of things. And those reflections are then immediately combined with the consideration of alternative concepts, alternative forms of representation, which, Wittgenstein seems to have thought, we can make intelligible to ourselves by imagining some of the background-conditions of a community's playing these alternative language-games to be sufficiently different from ours¹⁶.

One may have Davidsonian¹⁷ or 'naturalistic'¹⁸ doubts concerning the intelligibility of such alternative concepts, especially when the alternative concepts in question

¹⁴ *Remarks on the Philosophy of Psychology* II, §198, and also §199.. Cf. also Zettel §350, and *Remarks on the Foundations of Mathematics*, p. 200, where he claims that, if physical objects kept vanishing of themselves all the time (i.e. if 'confusion supervened'), the activity of counting things would probably go out of use, and thus the mathematical sentence ' $2+2=4$ ' would 'become unusable' (op.cit., pp. 51/2)

¹⁵ Wittgenstein, *Philosophical Investigations*, p.230.

¹⁶ Cf. Wittgenstein, *Philosophical Investigations*, p. 230. Cf. also Zettel, §§387/8, where he claims that only by imagining people with a different education and different interests can we imagine 'essentially different' concepts; cf. also Wittgenstein's examples of alternative, though analogous, forms of representation, such as alternative colour-geometries (cf. Zettel §§345, 368), alternative ways of measuring (cf. *Remarks on the Foundations of Mathematics*, pp. 38, 91) or of calculating the price of wood (cf. op.cit., pp.93/4).

¹⁷ Cf. Davidson, 'On the Very Idea of a Conceptual Scheme'.

¹⁸ Cf. esp. Stroud, 'Wittgenstein on Logical Necessity', pp. 512-517.

are supposed to be alternatives to concepts that we would regard as central to our present 'conceptual scheme'. However, these doubts need not disturb us in the present context, since the reflection on various background-conditions of our *present* concepts (which is all I shall be concerned with in the following) does not by itself depend on the intelligibility of *alternatives* to these concepts.

Regardless, therefore, of whether we may want to accept all of the more questionable premises of Wittgenstein's, admittedly fragmentary, reflections on the background-conditions of our concepts in general, I think we can draw two lessons from those reflections which may help us in dealing with our more specific problem of the background-conditions of the applicability of the notion of *personal* identity.

The *first* lesson is that the background-conditions of a given concept's applicability to a given situation (or, put differently, the *background-conditions* of a given concept's ability to determine reality save for a Yes or No) have to be distinguished *both* from the *human capacities* that make it possible for us to have a given concept, *and*, more importantly, from the *human needs, wants, and desires* that give the application of this concept a 'point' (i.e. a functional, *non-logical*, significance in our various practical pursuits). This distinction seems to have been conflated by Quine when he says that it was 'past needs' that invested our words with their 'logical force' (which I take to mean the ability to pose a Yes/No question to reality). It may be 'past needs' that caused us to introduce certain words into our language ('to develop a certain language-game'), and it may be the persistence of these needs which still give the use of these word a 'point' in contemporary human life. But it is its *meaning* as specified by its criteria (if it is the kind of word whose meaning is explained in this way) that gives it 'logical force'.

Keeping these two issues separate is particularly important when dealing with questions of personal identity. For an investigation into our concept of personal identity is unique in that the entities applying the concept are at the very same time the entities the concept is applied to. And this makes it very difficult for us to distinguish the question of the *applicability* of the concept of personal identity to a given situation from the question of whether in the given situation the application of the concept would *have a point* (i.e. practical significance). Thus if the scenarios imagined in some of the more far-fetched thought-experiments concerning personal identity were to become commonplace, our interests, needs, aims, and purposes in life would probably change as well. To anticipate an example that will be discussed in detail later on, suppose some form

of cloning were to become practically feasible and, arguably, we could all prolong our 'lives' indefinitely by having our 'personalities' transferred from one (artificially created) body to another. In that case many of our interests, expectations, aims, and purposes in life would change as well. We might, for example, not feel the need any more to apply the notion of numerical identity to persons, since its application would -- in a sense -- have become pointless. However, this does not by itself imply that the concept would have become inapplicable (although, as I will argue below, this may well be true). Thus, although it would arguably be very difficult *in practice* to keep the question of applicability of the concept of personal identity to a given situation separate from the question of how our interests, expectations, aims and purposes would be affected if we found ourselves in the envisaged situation, it will prove useful, and possible, to keep the two questions separate *in theory*¹⁹.

The *second* lesson that we can learn from Wittgenstein's reflections pertains not -- as the first does -- to radical changes in our shared *human* nature (needs, desires, interests, etc.), but to radical changes in the *general course of events in the world*. This is the lesson (which, it seems, Wittgenstein himself did not always properly heed) that we have to distinguish *imagined changes in the background-conditions* of a given concept which deprive the concept of its *applicability* to a given situation from imagined changes which only deprive the concept of its *having a point* in human life. The application of a concept can lose its point because some general features of the course of events have changed in such a way that nothing hangs any more on whether the concept applies or not. But a concept can have lost its point and still be applicable. Thus, to take Wittgenstein's own example cited above: in a world in which things changed colours all the time, there would not be much of a point in predicating colour-terms of objects, and thence 'our language-game with colours' would probably go out of use. But this is not to say that in such a world colour-concepts would have become *inapplicable*, i.e. that it would be *senseless* to say that at a given time *t* an object *x* was, say, red. (We merely would not have any good reason to assume that the object is still red now). Wittgenstein's comparison of words to tools may have caused him to overlook, or at least ignore, this distinction.

In the following section I shall give a more precise definition of the notion of

¹⁹I suspect that Parfit's treatment of personal identity suffers from just such a failure to distinguish between the two cited issues, cf. sect. 5.7..

a background-condition of the applicability of our concept of (numerical) personal identity. I shall argue that the limits within which thought-experiments can contribute to an elucidation of our concept of numerical identity with respect to persons are determined by the obtaining of these background-conditions. And I shall argue that in at least some thought-experiments these limits are traversed. What happens in those thought-experiments is that one or more of the background-conditions of the applicability of the concept of personal identity to the envisaged scenario are imagined not to obtain. I shall try to show that it is only with respect to those thought-experiments (but then for good reason) that our intuitions as to the obtaining of personal identity (lacking the guidance of the underlying concept which has become inapplicable to the envisaged situation) begin to waver and to be lured here and there by the addition of inessential details to the imaginary story. With respect to those thought-experiments, and with the cited qualifications, Strawson's and Quine's (and, in a way, Wilkes') reservations about the theoretical value of thought-experiments will be vindicated. They were right in pointing out that one cannot expect thought-experiments to unearth implicit rules for the application of our concept of personal identity where no rules have been laid down. However, I shall also argue that this is not to say that these ventures into conceptual no-man's land have no use whatsoever. By pushing our concept of personal identity to its limits, thought-experiments may help us to locate the bounds within which application of the concept is possible. In other words, they may help us to delimit the bounds within which identity-claims with regard to persons make sense. If we cross these bounds, that is to say, if the *non-obtaining* of the situation the obtaining of which our thought-experiment asks us to imagine is one of the background-conditions of the applicability of the notion of personal identity, then we cannot expect our identity-claims with respect to this situation to make sense any more. That is to say, we cannot expect the question 'Is it still the same person?' to be answerable by a clear Yes or a No. Whether in imagining a given situation we have crossed the bounds of sense drawn by our present concept of personal identity cannot be determined apriori. It has to be determined for individually for each imaginary case. For, arguably, only by crossing the bounds of sense can we find out where they are.

5.2. Inductive Evidence, Criteria and Background-Conditions

The way in which problems of personal identity have customarily been discussed is by searching for the necessary and sufficient *condition* of personal identity, which at

the same time is conceived to be the necessary and sufficient *criterion* (i.e. meaning-constitutive evidence) for the obtaining of personal identity. As was pointed out in the previous section, thought-experiments are supposed to be instrumental in this search: they are supposed to enable us to distinguish the genuinely criterial, and hence meaning-constitutive, from the only inductive, or symptomatic, evidence for personal identity. And this distinction is then supposed to determine what is essential to a person's persistence over time from what is only contingently the case whenever a person persists.

In what follows I shall (where appropriate) supplement the customary tests for necessity and sufficiency of a given kind of evidence with a third test, which will check whether the general obtaining of the evidential relation between a given kind of evidence and the obtaining of personal identity is one of the background-conditions of the applicability of the notion of personal identity. If so, the given kind of evidence would have to be conceived as meaning-constitutive and thus genuinely criterial of such identity-statements, without being either necessary or sufficient of their truth. If not, it would have to be conceived as merely inductive or 'symptomatic' evidence for the obtaining of personal identity.

The introduction of such a test is supposed to make room for the possibility, mentioned in sect. 2.1.4. above, that, although there indubitably are various forms of qualitative identity which have to count as inductive evidence for the obtaining of numerical identity, the mere fact that a certain kind of qualitative identity is defeasible evidence for numerical identity does not by itself make it inductive.

Thus, to illustrate this claim by the following table citing various forms of evidence for the obtaining of personal identity -- some of them genuinely criterial, some of them purely inductive or symptomatic, some actually applied, others in practice or even in principle inapplicable. (I shall refer back to it repeatedly in the following discussion).

Kinds of evidence for: P_2 at $t_2 = P_1$ at t_1

(1) P_2 at t_2	has the same bloodgroup as	P_1 at t_1
(2)	" has the same fingerprints as	"
(3)	" has the same structure of genes as	"
(4)	" has the (numerically) same brain as	"
(5)	" has the same physical appearance as	"
(6)	" has the same (experiential) memories as	"
(7)	" has the same (experiential) memories, intentions, beliefs,	

- | | | |
|-----|---|---|
| | character-traits, intellectual abilities, etc. as | " |
| (8) | " has the (numerically) same mind (or 'self') as | " |
| (9) | " has the (numerically) same living body as | " |

With respect to each of these states of affairs we can ask:

(a) whether the obtaining of the respective state of affairs is *necessary* for the obtaining of personal identity. (The way to answer this question is to try to imagine a situation in which we would assert the obtaining of personal identity, but not of the respective state of affairs).

(b) whether the obtaining of the respective state of affairs is *sufficient* for the obtaining of personal identity. (The way to answer this question is to try to imagine a situation in which we would assert the obtaining of the respective state of affairs, but not of personal identity).

(c) whether the *obtaining of the evidential relation* between the respective state of affairs and personal identity is one of the *background-conditions* of our application of the concept of numerical identity with respect to persons.

If it is *not*, then the obtaining of the respective state of affairs will have to count as purely *inductive* evidence for the obtaining of personal identity, and in that case the following will be true: if the evidential relation does not, or is *imagined* not to, obtain, in a given situation (that is to say, if in the given situation we cannot presumptively conclude from the obtaining of the respective state of affairs to the obtaining of personal identity), our concept of personal identity remains untouched. It remains decidable whether in the given situation P1 at t1 is the same person as P2 at t2.

If, *however*, the obtaining of the evidential relation between the respective state of affairs and personal identity *is* one of the background-conditions of our application of the concept of numerical identity with respect to persons, then the obtaining of the respective state of affairs will have to count as genuinely *criterial* evidence for the obtaining of personal identity, and in that case the following will be true: if the evidential relation does not obtain (or is *imagined* not to obtain) in a given situation, then our concept of personal identity loses its applicability with respect to this situation and thus its ability to determine reality save for a Yes or No. That is to say, it becomes undecidable whether in the given situation P1 at t1 is the same person as P2 at t2. In that case, the *contingent* fact that the obtaining of the respective state of affairs and the obtaining of personal identity *usually* go together has been shown to be one of the *meaning-constitutive background-conditions* of our being able to apply the concept of numerical identity.

ty at all. And thus the general obtaining of the evidential relation has been shown to constitute part of the *framework* within which our identity-discourse concerning persons makes sense, and without which it has lost its bearings.

It is important in this context not to conflate the sense in which a *background-condition* may be said to be meaning-constitutive with the sense in which *defeasible criteria* are sometimes called meaning-constitutive. However, it is crucial in this context to keep these two senses separate. A defeasible criterion for, e.g., pain may be said to be meaning-constitutive in the sense that (a) we learn and explain the concept of pain by reference to one of its defeasible criteria, and (b) the obtaining of the criterial state of affairs, e.g. the manifestation of a certain kind of pain-behaviour, is a *logically good, albeit defeasible*, reason for the ascription of pain to the person who manifests the pain-behaviour. The sense in which a background-condition may be said to be meaning-constitutive of a given concept-term is *purely negative*, or *counterfactual*. It amounts to saying that the application of the concept-term would not make sense *unless* the background-conditions for its applicability were fulfilled. However, it would be wrong to assume that one explains a concept-term by reference to one of its background-conditions, or that one justifies one's claim that a given concept applies by pointing out that one of its background-conditions obtains. What is true is that all such explanations and justifications can only take place on the background of the very general, but contingent, regularities which constitute the conditions of applicability of the concept.

The idea of such a 'Criterial-Test', and the corresponding notion of a background-condition of personal identity, will become clearer below, when they will actually be put to use. I shall try to show that although most of the various puzzle-cases concerning personal identity *are* conceptually decidable in a straightforward way, a particular kind of them is not. What happens in those latter thought-experiments is that certain background-conditions of the applicability of the concept of personal identity are imagined *not* to obtain. In which of these thought-experiments this happens, and hence what must *not* be the case (if only in the imagination) in order for our notion of personal identity to be applicable, cannot be determined *ab initio*. It can only be found out by trial and error, i.e. by a scrutiny of the theoretical status of the various forms of evidence for personal identity listed above.

5.3. Brain-Transfers and Personal Identity

One argument that is sometimes adduced as an argument against the view that personal identity consists in the spatio-temporal continuity of a given animal body is the (so far only logical) possibility of *brain-transfers*. This possibility was first envisaged

by Shoemaker in his original *Brown-Brownson-case*²⁰. The case is described as follows: During a surgical operation which requires the temporary removal of the brain from the skull, Brown's brain is inadvertently placed back into Robinson's body. Now imagine Brownson (i.e. the person who has Robinson's body but Brown's brain) waking up in his hospital-bed the next morning. He seems horrified to find himself in Robinson's body and, when asked, immediately answers that he is Brown, not Robinson. He remembers all, or at least most of, the things that Brown has experienced throughout his life, and he displays habits and character-traits that are very similar to Brown's. After some initial hesitation, we would, it has been suggested, come to the conclusion that -- against first appearances -- Brownson is identical with Brown. In doing so, we let ourselves be guided by what Noonan has called the 'brain criterion'²¹ (our number (4) in the above list, cf. sect.5.2.). An adherent of the brain-criterion takes the possibility of brain-transfers to demonstrate that what is essential to personal identity is not the persistence of a whole animal organism, but rather the persistence of a *very specific part* of such an organism, namely the part which is in some sense 'causally responsible for a person's mental continuity'. Since, as the slogan has it, 'the person goes where the brain goes', identity of the brain is viewed to be the necessary and sufficient criterion of personal identity.

I have already argued above (pp.7 ff.) that the view that *brain-identity* is what *personal identity* 'consists in' is either based on a *misunderstanding*, or (as in Nagel's and Mackie's case) a different (although arguably misleading) understanding, of the search for the criterion of personal identity. Just as persons are not identical with their brains, personal identity is not constituted by the numerical identity of the brain, although the brain may well be called the causal vehicle, or the causal *chassis*, of a person's identity -- in the sense that a person could survive without his left thumb, and even without his (own) heart, but not (as far as we know) without his brain. If this is what the claim that brain-identity constitutes personal identity is supposed to amount to, then it is (contingently) true; but it is not what we are looking for when we are searching for a *criterion* of personal identity. A criterion of personal identity is supposed to provide us with *meaning-constitutive* evidence for the obtaining of personal identity. In

²⁰Cf. Shoemaker, *Self-Knowledge and Self-Identity*, ch.1.

²¹Noonan *Personal Identity*, pp. 3-5.

other words, it is supposed to elucidate our *concept* of personal identity; it is supposed to spell out the circumstances under which we would *count* a person P_1 at t_1 as numerically identical with a person P_2 at t_2 . That it is a person's brain, rather than his heart or his left thumb that causally 'underlies his mental life', is something scientists have found out *about* persons, but it is not *conceptually* constitutive of personal identity.

From a purely logical point of view the so-called brain-criterion is thus on the same level as other clearly *inductive* or *symptomatic* forms of evidence such as the ones cited as numbers (1)-(3) in the above table (namely identity of bloodgroup, fingerprints, and genetic structure). And this is to say that, being only inductive evidence for the obtaining of personal identity, it is *not a criterion at all*. Of course, this is not to impugn the fact that -- just as identity of fingerprint and identity of genetical structure are *in fact* highly individuating of persons, indeed individuating to such a high degree that, at least in some countries, they may be used as final ('beyond reasonable doubt') evidence in a court of law -- as a matter of *empirical fact* personal identity and brain-identity always go together. However, just as we can easily imagine that two candidates for numerical identity, P_1 at t_1 , and P_2 at t_2 , have the same bloodgroup, identical fingerprints, and identical genetical structure, without being the same person, we can easily imagine situations (although such situations may be *physically* impossible) in which we would count a person P_1 at t_1 as numerically identical with P_2 at t_2 without P_1 and P_2 having the same brain, just as (vice versa) we can imagine brain-identity without personal identity. In order to do that we just have to imagine human beings differing from the real ones only in that their hearts, as opposed to their brains, are causally responsible for their mental continuity. (This may be physically impossible, but it is certainly not conceptually impossible). If this were the case, we could not any more safely conclude from identity of the brain to identity of the person (e.g. if brains could be transplanted like hearts or livers without there being any change in psychological qualities on the part of the brain-donor and the brain-receiver). However, our concept of personal identity would remain untouched by this imagined change. Since the putative criterion of brain-identity does not play any role in our present identity-discourse concerning persons, a distortion of the evidential relation between brain-identity and personal identity would not affect it in any way.

It is also interesting to note (and this I take to be further proof for the purely inductive character of the so-called brain-criterion) that even in the imaginary brain-transplant cases it is *never* the numerical identity of the brain that is actually *used* as the

decisive evidence in answering the respective identity-question. To cite Shoemaker's original Brown-Brownson case as an example: in this brain-transfer case we do not say (though, I shall argue below, we should not) that Brownson is (numerically identical with) Brown *because* he has the same brain. We say it because Brownson in what he says and does *manifests psychological continuity* with Brown. It is psychological continuity that here serves as our criterion of identity. If we mistakenly planted Brown's brain into Robinson's body, and the awakening Robinson-body displayed all the usual Robinson-like character-traits, memories, capacities, etc., we might be astonished, since this result would contradict a large section of what we presently take to be scientific knowledge. But we would not hesitate to conclude that the person inhabiting Robinson's body is in fact no-one else but *Robinson*.

Thus if we are looking for a serious opponent to the bodily criterion of personal identity, we shall have to consider psychological continuity, not the (putative) brain-criterion. This will be done in the next section. Before that, however, I should like to briefly examine a weaker version of the brain-criterion of personal identity, which was suggested by Wiggins in his *Identity and Spatio-Temporal Continuity*.

According to this conception, what is essential to personal identity is not bodily continuity as such, but rather the continuity of a person's 'life and vital functions' which, Wiggins says, 'define an individual in the category of substance. They define a person.'²² A person's 'life and vital functions', Wiggins goes on to claim, have to be causally anchored in some way in some part of a person's body. Wiggins calls this special part of a person's body, which is supposed to be the 'seat of memory and other functionally characteristic capacities', a person's '*individuating nucleus*'²³. *De facto*, the role of being a person's 'individuating nucleus' is played by the human brain. However, Wiggins emphasizes, it is not logically necessary that it is a person's brain that should play the role of individuating nucleus. All that is logically or conceptually necessary is that some part of a person's body, some 'parcel of matter'²⁴ (be it a person's heart or liver or whatever) should play this role. Thus Wiggins can conclude that

'coincidence under the concept *person* logically requires the continuance in one organized parcel of all that was causally sufficient and causally necessary to the continuance of essential and characteristic functioning, no autonomously sufficient part achiev-

²²Wiggins, *Identity and Spatio-Temporal Continuity*, p.51.

²³Ibid..

²⁴Ibid..

ing autonomous and functionally separate existence.’²⁵

In other words, for P_1 at t_1 to be numerically identical with a person P_2 at t_2 , P_2 need not necessarily have the same brain as P_1 . But he must have as the seat of his ‘memory and other functionally characteristic capacities’ the very same *bodily part* which had earlier been P_1 ’s ‘seat of memory and other functionally characteristic capacities’. Hence, Wiggins concludes, it is logically possible that one and the same person should have different bodies at different times, namely in case the relevant bodily part (the ‘individuating nucleus’) has been transferred from one body to another.

Wiggins’ approach is ingenious, but hardly plausible. As was first pointed out by Shoemaker²⁶, there does not seem to be any a priori logical reason that a person’s psychological characteristics should be continuously ‘causally seated’ (whatever that means) in any one particular part of his body. There is, for example, no apriori good reason to exclude the possibility that the two hemispheres of the brain, or indeed the brain and the heart and the liver, should alternate at regular intervals in being the ‘seat of memory and other functional characteristics’. And therefore there is no apriori good reason to assume that the transfer of any one particular bodily part at any one time should result in a person’s ‘change of body’, since there is no apriori reason to assume that a given bodily part should at this particular time fulfil the role of being the ‘seat of memory’, and hence no reason to assume that the transfer of this particular ‘parcel of matter’ from one body to another should result in a ‘transfer’ of the relevant characteristics from one body to another. Thus, once again, we are driven back to the so-called ‘psychological continuity criterion’ of personal identity, and it is to this that we must now turn.

5.4. ‘Body-Changes’ and Psychological Continuity

5.4.1. Locke, Quasi-Memory, and Psychological Continuity

The history of the psychological continuity criterion of personal identity goes back as far as Locke’s definition of a person as a

‘thinking intelligent being that [...] can consider itself as itself, the same thinking thing in different times and places’,

and his corresponding claim that

‘consciousness [...] is that that makes everyone to be what he calls *self*, and thereby distinguishes himself from all other thinking beings: in this alone consists *personal identity*, i.e. the sameness of a rational being. And as far as this consciousness

²⁵Op.cit., p.55.

²⁶Cf. Shoemaker, ‘Wiggins on Identity’, esp. pp.114 ff..

[i.e. the consciousness of one's self as having sensations and perceptions] can be extended backwards to any past action or thought, so far reaches the identity of that person.²⁷

There have been various attempts to elucidate Locke's claim that the identity of a person consists in the identity of a given 'consciousness' which is alleged to 'extend backwards' to a person's past²⁸. One possible way of understanding Locke's theory of personal identity would be to understand Locke as holding something like a (first-person) *experiential memory criterion* of personal identity, as expressed in number (6) of the above table. According to this criterion a person P_1 at t_1 is identical with a person P_2 at t_2 iff. P_1 and P_2 have the same first-person experiential memories, i.e. the same first-person memories of their past actions and experiences (cf. sect. 4.3. above). The problems with this suggestion are well-known²⁹. People do forget many of the things they used to remember without thereby becoming different persons; senility is a case in question. More precisely, a person P_2 at t_2 may have lost many, or even all, of the first-person experiential memories that he used to have as P_1 at t_1 , and still P_1 and P_2 may be one and the same person. This is why several theorists of personal identity have tried to fill the flaws in Locke's account by sharpening the supposed memory-criterion into a *memory-continuity* criterion. The criterion then reads:

P_n at t_n is numerically identical with P_1 at t_1 if and only if P_n at t_n is linked by continuity of memory to P_1 at t_1 , where this is to mean that P_n at t_n now experientially remembers many of the things he used to remember as P_{n-1} at t_{n-1} , and P_{n-1} at t_{n-1} remembered many of the things he used to remember as P_{n-2} at t_{n-2} , etc..

However, it is evident that even this more sophisticated criterion cannot cope with the possibility of *complete* memory breakdown, as in a case of total amnesia. Amnesia, however, even if irreversible, is not considered equivalent to death, and bodily continuity helps us to maintain a grip on a person's identity in such cases. To be sure, complete amnesia may prevent a person P_2 at t_2 from *finding out who* he is (namely P_1), but it does not by itself prevent him from being *identical* with P_1 . And hence identity of experiential memories cannot count as a *necessary* condition of the obtaining of personal identity.

²⁷Cf. Locke, *Essay Concerning Human Understanding* II, xxvii. 9.

²⁸The most recent of these attempts, and the most thorough, is to be found in Ayers, *Locke*, vol II, pp.256 ff..

²⁹Cf. Noonan's discussion in chapters 2 and 3 of his *Personal Identity*, pp. 30-77.

Another way to interpret Locke's claim that

'... it being the same consciousness that makes a man be himself to himself, personal identity depends on that only, whether it be annexed only to one individual substance, or can be continued in a succession of several'³⁰

is to take it not as holding a memory continuity criterion, but as making a claim about the relation between a person's present memories and his past actions or experiences that these memories are *of*. An interpretation of Locke's account of personal identity along these lines brings out the logical significance of experiential memory to personal identity (suggested by Locke's remarks) in a different way. It understands Locke as claiming that first-person experiential memory is constitutive of personal identity in so far as it *establishes* the crucial 'link' that connects a person with his past. It establishes the link between his present memories and the past experiences or actions that these memories are memories *of*.

Prima facie, however, an interpretation of Locke's remarks along these lines is wide open to a well-known criticism which accuses it of being irremediably *circular*. It is circular, since, it has been pointed out, the relation in which a remembering person stands to *his* past experiences (if it can be said to be a relation at all) is *trivially* sufficient of personal identity insofar as it is part of our very *concept* of experiential memory that the remembering person is identical with the person who witnessed the event or action remembered³¹. That is to say, we would not normally count a person as experientially remembering doing or experiencing something, unless his memory was veridical, i.e. unless what he thus claims to remember actually happened. 'Remember' (like 'know' and, in some uses, 'see') is what Ryle called an 'achievement-word'³². Just as to say of somebody that he knows something is to imply that what he knows is *true*, to say of somebody that he experientially remembers something is to imply that his memory is *veridical*, i.e. both that what he remembers actually *happened*, and that it was *he himself* who witnessed or experienced it.

This may be so. However, we have already seen (cf. above sect.4.3.) that this

³⁰Cf. Locke, *Essay Concerning Human Understanding*, II, xvii, 9.

³¹The first to present this objection was of course Butler, cf. 'Of Personal Identity', p. 100. For historical detail the reader is again referred to Noonan, *Personal Identity*, pp. 68-71.

³²Cf. Ryle, *The Concept of Mind*, pp. 125/6.

restriction can easily be lifted by the introduction of a new concept of *quasi-memory*, which may be viewed as a legitimate extension of our present concept of memory. According to this new concept, a person can be said to quasi-remember doing or experiencing something without actually *himself* having done or experienced it. In Shoemaker's own words:

'Whereas someone's claim to remember a past event implies that he himself was aware of the event at the time of the occurrence, the claim to quasi-remember a past event implies only that someone or other was aware of it.'³³

That is to say, the concept of quasi-memory is arrived at by lifting what Shoemaker calls the 'previous awareness condition' from our present concept of experiential memory. In doing so, Shoemaker did not intend to imply that in order for a person to have a quasi-memory what he thus remembers need not have happened *at all*. For if so, the notion of quasi-memory could hardly be called an extension of our present notion of experiential memory, since in that case it would not be a form of *memory* at all. All that the notion of quasi-memory (as opposed to our present concept of memory) is supposed to make room for is the possibility that what a person thus quasi-remembers doing or experiencing was in fact not done or experienced by *him*, but by *someone else*³⁴.

Thus the quasi-memory theorist can, at least initially, evade the circularity-objection and may argue that his concept of quasi-memory can serve as a substantial, and that is non-circular, criterion of numerical identity of person. One crucial consequence of this lifting of the 'previous awareness condition' from our present concept of memory is of course that the veridicality of a quasi-memory (as opposed to the veridicality of a familiar non-quasi memory) could not be verified by establishing whether the remembering person was actually witness to the action or event he claims to remember. With respect to a familiar non-quasi memory, this could in principle be achieved by tracing the remembering person's *body* back to the scene or action he claims to remember. In an account of personal identity in terms of quasi-memory this aspect of our present concept of memory is typically taken to be emulated by the idea of *memory-chains*, which are conceived as *causal* chains linking present memories to the past experiences or actions that these memories are of. This feature of the quasi-memory account is quite explicit in Shoemaker. Shoemaker takes the causal requirement, according to

³³Shoemaker, 'Persons and Their Past', p. 24.

³⁴Cf. Shoemaker, 'Persons and Their Past', p.24.

which a person cannot be said to remember something unless his memory was *caused* in an appropriate way by the past event that his memory is a memory of, to be already implicit in our present concept of memory. As he puts it:

‘... a claim to remember a past event implies, not merely that the rememberer experienced such an event, but that his present memory is in some way *due to*, that it came about *because of*, a cognitive and sensory state the rememberer had at the time he experienced the event’³⁵

And he thinks that by building such a causal requirement into the concept of quasi-memory (i.e. requiring that ‘a veridical quasi-memory must not only correspond to, but also stand in an appropriate causal relationship to, a past cognitive and sensory state of someone or other’³⁶) he can specify certain relations in which two persons P_1 at t_1 and P_2 at t_2 have to stand in order to be identical with each other.

One may have qualms about this ‘reification’ of memories and about the idea of ‘memory-traces’ being causally linked to ‘past cognitive and sensory states’. There are also difficulties in spelling out the ‘appropriate way’ in which a present memory has to be caused by a past event in order to be a veridical memory *of* that event³⁷. However, I shall not pursue the matter any further in the present context, since the problems encountered by the memory-criterion and its cousin, the psychological continuity criterion, will prove to be independent of whether one accepts a causal account of memory or not.

The notion of quasi-memory is generally thought to be at the heart of the so-called *psychological continuity criterion*, but it is by no means exhaustive. We have seen above that the putative memory-continuity criterion cannot be shown to constitute a *necessary* criterion of personal identity, and the psychological continuity criterion may be viewed as a neo-Lockian attempt to remedy these shortcomings by widening the putative memory criterion into the psychological continuity criterion (our number 7 above). The psychological continuity criterion tries to capture all aspects of what might be said to constitute a person’s ‘individuality’ or ‘unique personality’, understood in a wide sense which includes not only a person’s character, but also his intellectual abili-

³⁵Shoemaker, ‘Persons and Their Past’, p. 26. Shoemaker is here following Martin and Deutscher’s well-known causal analysis of memory, cf. their article ‘Remembering’.

³⁶Ibid.

³⁷Cf. e.g. Malcolm, *Memory and Mind*, and Squire, ‘Memory Unchained’.

ties, moral qualities, beliefs, idiosyncratic fears, prejudices, predilections, and the connections that may be said to exist between an intention and a later attempt to carry it out, a desire and a later attempt to fulfill it, etc.. The psychological continuity criterion may be stated in the following way:

P_n at t_n is numerically identical with P_1 at t_1 if and only if P_n at t_n is linked by psychological continuity to P_1 at t_1 , where this is to mean that P_n at t_n is linked by the obtaining of a chain of psychological connections to P_{n-1} at t_{n-1} , who in turn is linked by the obtaining of such chains to P_{n-2} at t_{n-2} , etc..

This candidate for a criterion of personal identity certainly looks more promising. For it takes into account the crucial fact that persons conceive of themselves not only as having a certain unique past history, but also as having a, however uncertain, *future*. And one may add that, arguably, it was this wider notion of 'consciousness' -- i.e. not just experiential memory -- that Locke had in mind when he claimed that it is what 'makes a man be himself to himself', and what could be transferred from one substance to another. However, if we actually put the notion of psychological continuity to the test, we will once again find ourselves disappointed.

Our concept of personal identity seems to leave nearly unlimited room for people undergoing radical psychological changes whilst still remaining the same person. Was the old Kant, mumbling to himself and (according to de Quincey's description) absent-mindedly tying and untying the belt of his bathrobe over and over again, not one and the same person as the author of the three *Critiques*? Well, he was not dead yet. And was the mad Nietzsche who, on walks with his mother used to jump into puddles like a child, not one and the same person as the author of *Beyond Good and Evil*? Heavily mutilated persons, yes, 'not quite their old selves', one might say, but still the same persons. Our concept of personal identity *does* allow for the possibility of psychological mutilation, and in extreme cases (such as irreversible coma) it is not always clear where mutilation ends and death starts. Of course, an adherent of the psychological continuity criterion might point out against these reservations that even in the cases just cited psychological continuity does obtain, although radical changes in character, beliefs, memories, intellectual abilities may have taken place. For, he may insist, there are 'chains of mental connectedness' between the young and the old Kant, and the sane and the mad Nietzsche. And he might stress that it is just the obtaining of these connections which 'make' the young Kant one and the same person as the old and senile Kant, and the mentally sane Nietzsche one and the same person as the deranged paranoiac as which he spent the last eleven years of his life -- *not*, that is, the fact that the early and

the later Kant, and the early and the later Nietzsche, had throughout their lives only one numerically identical *living human body* ³⁸. I am not entirely convinced that we really do understand the claim that even in those cases of radical ruptures in a person's mental development psychological continuity still obtains (in whatever twisted form). However, I shall not press the point here.

I shall not press it since, it seems to me, the main aim of Locke's and various neo-Lockean theories of personal identity is not so much to prove psychological continuity to be a *necessary* condition of personal identity. The crucial point is rather to show that psychological continuity by itself (i.e. *independently* of the *spatio-temporal* continuity of a given person's *body*) can be shown to be a *sufficient* condition of the obtaining of personal identity. This is brought out by the emphasis that both Locke's and various neo-Lockean accounts have laid on demonstrating the possibility of one identical person's '*body-change*'. The possibility of a body-change is the touch-stone of the psychological continuity criterion. If it could be shown that it is *logically* possible for one numerically identical person to have or 'inhabit' different *bodies* at different times of his existence, Locke's claim that personal identity consists in 'identity of consciousness' would have found at least partial justification. For then personal identity would have been shown to be governed by a criterion of identity different from the criterion governing the identity of all other concrete particulars (namely spatio-temporal continuity). Typically, the way in which the logical possibility of body-changes is argued for is by framing thought-experiments (from Locke's prince-and-cobbler case to Shoemaker's Brown-Robinson brain-transfer to Parfit's tele-transportation stories) purporting to prove that certain extraordinary circumstances would force us to count one identical person as having *changed bodies*. It is to these thought-experiments and their alleged persuasive power that we now have to turn.

5.4.2. '*Branch-Line Cases*' and the '*Reduplication Argument*'

Often the psychological continuity criterion is argued for from the *first-person*

³⁸That -- as long as in an imaginary case we are not lured away from 'identifying ourselves' with our own bodies by the mention of an allegedly better candidate for numerical identity -- we tend to identify ourselves with our bodies even in the *total* absence of psychological continuity may be taken to have been established by Williams' second thought-experiment in his article 'The Self and the Future'. This thought-experiment will be discussed further below, cf. sect.5.5.

perspective, namely by presenting the reader with an imaginary case in which he is invited to *identify himself* with one of the protagonists. That is to say, the reader is invited to *imagine himself* waking up in a body which is not his own, and hence to regard himself (P_1) as having survived as a given psychologically continuous person P_2 . This is then taken to show that, if put under pressure, the reader would apply psychological continuity as the criterial evidence to determine if a given person (himself) has survived, and hence that psychological continuity is what personal identity 'consists in'.

Typically, these imaginary cases present the 'transfer' of psychological states as a *one-one* relation. That is to say, the 'personality'-donor's body (where the term 'personality' is here to be understood in a wide sense as denoting a set of psychological characteristics, such as memories, beliefs, character-traits, etc. which together constitute a whole '*mental history*') is typically imagined to be destroyed after the transfer, and the given person's 'personality' is imagined to be 'transferred' to *only one* new body (which, preferably, has not been *another* person's body before). This observation holds (e.g.) for Parfit's 'simple tele-transportation cases'³⁹ as well as for the people envisaged in Shoemaker's future society in which 'brain-state transfers' have become commonplace⁴⁰ (cf. the discussion below).

I do not want to deny that these cases have a certain initial intuitive appeal. However, it is interesting to note that the temptation to 'identify oneself' (i.e. to count oneself as identical) with a certain future person who is represented as *psychologically* (but not *bodily*) continuous with oneself fades immediately once the envisaged situation is modified to the effect that the 'personality'-donor's body is not imagined to be destroyed (but live on), or once the donor-person's 'personality' is imagined to be copied onto more than one recipient. Thus, with regard to Parfit's tele-transportation case, one may be tempted to identify oneself with the person regaining consciousness on Mars in the 'simple' case in which one's earth-body is imagined to be destroyed during the 'tele-transportation'. However, if the imaginary scenario is changed to the effect that I survive the procedure, and find myself talking to my Martian counterpart, the temptation clearly fades.

Parfit has called these latter imaginary cases, in which we are presented with

³⁹Cf. *Reasons and Persons*, Chapter 10, pp. 200 ff.

⁴⁰ Cf. Shoemaker in Shoemaker/Swinburne, *Personal Identity*, pp. 109f.

two persons P_2 and P_3 at t_2 , who are *equally* (or to a similar degree) psychologically continuous with a person P_1 at t_1 , ‘*branch-line cases*’⁴¹ of *personal identity*. In what follows I shall try to show two things:

I shall try to show, *firstly*, that the mere fact that our ‘intuitions’ as to imaginary cases involving a person’s putative ‘body-change’ depend on such seemingly inessential details of the presentation of the imaginary case is symptomatic of the fact that these intuitions are based on paralogistic illusions. And I shall try to show, *secondly*, that the mere possibility of such an embarrassment of riches as it is envisaged by ‘branch-line cases’ shows something crucial about the *logical* status of the psychological continuity criterion. It shows, put in a nutshell, that psychological continuity can *never* (i.e. even in cases in which psychological continuity does *not* branch) count as a sufficient criterion of *numerical* identity of person.

A case of *branching* of psychological continuity (or, as it is also sometimes put, a case of *person-fission*) was first discussed by Wiggins in the form of a brain-split case⁴², in which a given person’s brain is divided through the middle and then transplanted into two different bodies. ‘Half-brain transfers’ are a very picturesque way of imagining a given person’s psychological continuity branch. However, this particular form of presentation is not essential to the *conceptual* problems raised by a branching of psychological continuity. For, as we have seen above (sect. 5.3.), the relation between brain identity and psychological continuity is conceptually *contingent*. Therefore, instead of encumbering our discussion with certain practical limitations of imaginary cases involving half-brain transfers, we might as well regard problems of branching of psychological continuity in their purest form, namely as brought about by some such imaginary device as Williams’ *brain-state transfer device*, which is conceived to be able to scan a given brain’s total *informational state* and encode it into indefinitely many other brains⁴³. (Parfit’s tele-transportation cases may be viewed as a sub-class of such brain-state transfers).

The problems that so-called ‘branch-line cases’ pose for the view that psycho-

⁴¹This is Parfit’s expression, cf. *Reasons and Persons*, p.201.

⁴²Cf. Wiggins, *Identity and Spatio-Temporal Continuity*, pp. 51 ff.

⁴³Cf. ‘The Self and the Future’, *passim*, and also Shoemaker in Shoemaker/Swinburne, *Personal Identity*, pp. 109f.

logical continuity can count as a *sufficient* criterion of personal identity have been extensively discussed in the recent literature. They were first pointed out by Williams in connection with a well-known examination of the memory-criterion of personal identity. Williams' objection to the memory-criterion, which has become known as the '*reduplication argument*', can be briefly summarized as follows⁴⁴.

As we have seen in the previous section, in order for the memory criterion to be a substantial, i.e. non-circular criterion, it must be at least *logically possible* for a person P_2 to have experiential memories of a person P_1 's past actions or experiences without it being trivially true that P_1 and P_2 are identical. The notion of quasi-memory fulfils this requirement by making room for the possibility that two different persons have identical experiential memories without thereby (trivially) being the same person. It makes room for the possibility that P_2 at t_2 has all the same *quasi* experiential memories as P_1 at t_1 , and still P_1 and P_2 are numerically distinct persons. This conceptual leeway gained by the introduction of the notion of quasi-memory, which makes it possible for the memory-criterion to function as a non-circular criterion, however, at the very same time prevents it from being a genuinely *sufficient* criterion of *numerical* identity of person. For it opens up the possibility of there being two or more equally well qualified, but numerically distinct, candidates P_2 and P_3 for numerical identity with a person P_1 . But since P_2 and P_3 cannot both be numerically identical with P_1 (for then -- due to the transitivity of identity -- they would have to be identical with each other), nor should we conclude that either of them is identical with P_1 .

In the article in which the argument was originally presented, Williams illustrated his claim by the example of a by now famous imaginary case, which involves a contemporary -- call him Charles -- who claims to be Guy Fawkes, and whose memory-claims uniquely fit the life of Guy Fawkes. With respect to this case Williams asked whether we should view Charles' case as one of reincarnation, i.e. whether we should consider Charles as numerically identical with Guy Fawkes, or whether we should rather consider him an (extremely well-informed, but nevertheless deluded) madman. Against anybody who might be tempted to view Charles' case as one of reincarnation Williams points out that if we can imagine Fawkes' memories mysteriously transferred to Charles, we can just as easily imagine Fawkes' memories *reduplicated* in Charles' brother Rob-

⁴⁴Cf. Williams, 'Personal Identity and Individuation', esp. pp.5 ff., and Noonan's discussion of the argument in *Personal Identity*, ch.7, pp.149 ff..

ert. In that case we would have two equally good candidates for identity with Guy Fawkes. Since, however, two numerically distinct persons cannot both be identical with Guy Fawkes (since then according to the transitivity of identity they would have to be identical with each other), neither should we identify Charles with Guy Fawkes in the original case where there is no reduplication. For the identity of Charles with Guy Fawkes can only be a matter of the *intrinsic relations* that hold between the two of them, not a matter of what is true about other people (in our case, Robert). It would be absurd to claim, Williams argues, that whether a person P_2 at t_2 is identical with a person P_1 at t_1 can depend upon facts about people other than P_1 and P_2 .

Williams' original version of the 'reduplication argument' was thus framed mainly with the aim of presenting a difficulty for the more narrow memory-criterion. However, it is evident that Williams' basic insight may be equally applied to those more encompassing branch-line cases of *psychological* continuity with which we are presently concerned. Take a case in which a person's P_1 total brain-state is scanned and subsequently encoded onto two new organisms P_2 and P_3 . Since P_2 and P_3 are clearly numerically distinct, they cannot -- due to the transitivity of identity -- both be the same person as P_1 . But since the identity of P_2 with P_1 cannot depend on the non-existence of any further, equally psychologically continuous, candidate P_3 , neither -- the revised reduplication argument claims -- should we count P_2 as numerically identical with P_3 in the case in which there is *no* branching of psychological continuity (i.e. in the case in which there is only one clone psychologically continuous with P_1 and thus no duplication of a given set of psychological qualities). We should not count P_1 as identical with P_2 , because -- the argument assumes -- the question of whether P_2 is, or is not, numerically identical with P_2 cannot depend on any aspects of the case which are *extrinsic* to P_1 and P_2 , as would be the existence of a second candidate P_3 .

Furthermore, Williams viewed his argument from the possibility of reduplication as bringing out the crucial importance of *bodily* identity for questions of personal identity⁴⁵. He took it to demonstrate that bodily identity is always a necessary condition of personal identity. As will become evident below, this view is finally justified, but I think that it does not follow directly from his argument. What, it seems to me, is more impor-

⁴⁵Cf. Williams, 'Personal Identity and Individuation', p.10: 'The only case in which identity and exact similarity could be distinguished... is that of the body -- 'same body' and 'exactly similar body' really do mark a difference. Thus ... the omission of the body takes away all content from the idea of personal *identity*.'

tant in the present context is to consider the more *general* implications that Williams' argument has with respect to the *logical* status of the psychological continuity criterion. That there are such more general implications was first noted by Perry who remarked that 'Williams has posed an embarrassment for any account that uses as a criterion of identity a conceivably duplicable relation'⁴⁶. And Wiggins concluded from Williams' argument that

'What we need, if *identity* is what we want to elucidate, is a criterion which will stipulate that for a relation R to be constitutive of the identity of a and b, a's having R to b must be such that objects distinct from a and b are irrelevant to whether a has R to b.'⁴⁷

Ever since this more general significance of Williams' argument had been recognized, the discussion focused on the question whether a *revised* form of the psychological continuity criterion could possibly fulfil the requirement laid down in the passage from Wiggins (the 'only x and y principle', as it has come to be known in the literature⁴⁸). There have been various attempts by adherents of the psychological continuity account to accept the challenge posed by the argument from the possibility of duplication and at the same time evade the *reductio ad absurdum* suggested by it. And there have been equally many rebuttals of these attempts.

Thus, to cite just one prominent example, *Shoemaker*, and, in a way, *Parfit*, have tried to make the psychological continuity account immune to the reduplication argument by claiming, not that personal identity consists in psychological continuity *simpliciter*, but that it consists in *non-branching* psychological continuity⁴⁹. According to this revised account, a person P_2 is identical with a person P_1 if, and only if, P_2 is psychologically continuous with P_1 , *and there is no further identity-candidate* P_3 equally psycho-

⁴⁶Perry, 'Review of Williams's *Problems of the Self*, p. 428:

⁴⁷Wiggins, *Sameness and Substance*, p. 96.

⁴⁸Cf. Noonan, *Personal Identity*, p. p.152.

⁴⁹Cf. Shoemaker in Shoemaker/Swinburne, *Personal Identity*, pp. 115-118. Cf. Parfit, *Reasons and Persons*, pp. 261-265, esp. p.263, where he says: 'Personal identity consists in R [i.e. psychological connectedness and/or continuity with the right kind of cause] holding *uniquely* -- holding between one present person and *only one* future person'.

However, Parfit's view is more complicated in so far as he also holds a 'Reductionist View' of personal identity. Parfit's reaction to the possibility of branching of psychological continuity is best understood as a *dissolution* of our present concept of personal identity; it presents special complications and hence deserves a separate treatment, which it will receive below (sect.5.7.).

logically continuous with P_1 . This modification of the original account seems to evade the duplication argument. However, as has been pointed out by Williams and Swinburne, it does so only temporarily. For the artificial and entirely *ad hoc* introduction of a uniqueness-condition into the putative psychological continuity criterion cannot make up for its crucial shortcoming. As Swinburne has sarcastically pointed out, the introduction of such a uniqueness-condition has the absurd consequence that ‘the way for a man to ensure his own survival is to ensure the non-existence of future persons too similar to himself’⁵⁰. Thus, to illustrate this claim by the example of a split-brain case in which a person’s psychological continuity is imagined to branch to two future persons, a man could guarantee his own survival by bribing one of the nurses to let one of the two half-brains die. And this is an evidently absurd consequence.

Adherents of the psychological continuity criterion have reacted to this second-level *reductio ad absurdum* of their account by claiming that ‘identity is not what matters in survival’, and that therefore the objection misses its aim⁵¹. I shall discuss this claim in more detail below in connection with Parfit’s ‘Reductionist View’ of persons. However, *prima facie* such a proposed dissociation of the notion of survival from that of numerical identity should strike us as dubious. As we have seen above (sect. 2.1.2., p.20) the notion of the survival and of the numerical identity over time of a given *substance* or *particular* are mutually interconnected, and it seems difficult to see how they could be prized apart without change of meaning.

A different strategy open to an adherent of the psychological continuity criterion would be to attack the crucial assumption on which the reduplication argument is founded, namely the ‘only x and y principle’ itself. The ‘only x and y principle’, it might be argued, expresses a requirement that could not possibly be met by any criterion which is a *continuity* criterion. It might be held that Williams’ criticism of psychological continuity as a criterion of numerical identity of person could just as well be turned against the so-called body-criterion, i.e. the criterion which invokes the spatio-temporal continuity of a living animal body. For just as we can imagine branching of psychological continuity, we can imagine that a whole body should, amoeba-like split and depart

⁵⁰ Swinburne, ‘Personal Identity’, p. 237/8. The argument had been anticipated by Williams in ‘Are Persons Bodies?’, cf. p. 78.

⁵¹ Cf. Shoemaker in Shoemaker/Swinburne, *Personal Identity*, pp. 119-121, and Parfit, *Reasons and Persons*, ch. 12, ‘Why Our Identity is not What Matters’.

in two different directions.

However, as Williams himself once pointed out, there is a crucial logical difference between the notion of spatio-temporal and the notion of psychological continuity (which was obviously modelled on the former). The possibility of a body-split, as he puts it,

'does not show... that criteria based on the continuity of material things (whether whole bodies or whole brains) are in absolutely no different case with regard to the reduplication problem than are other criteria not so based. For the reduplication problem arises if a supposed criterion of identity allows there to be two distinct items B and C, each of which satisfies the criterion in just the way it would if the other did not exist. But this is not so with bodily continuity: what is true of B when it is in the ordinary way continuous with A is just not the same as what is true of it when, together with C, it has been produced from A by fission.'⁵²

I think that Williams' observation points out the crucial flaw in the psychological continuity criterion, which, I think, has not been properly recognized in the various attempts to 'mend' the psychological continuity criterion so as to make it compatible with the 'only x and y principle'. The *crucial flaw* in the psychological continuity criterion (and the final reason why it cannot be made to comply with the 'only x and y principle') is, it seems to me, that it is not *essentially*, or *inherently individuable*. Its individuable power depends on the *contingent* fact that a given 'personality' (i.e a given set of psychological qualities) is in practice only ever instantiated once. However, in order for a given state of affairs (a given form of evidence) to be a necessary and *sufficient* criterion of numerical identity of person, it has to be inherently, *not just contingently*, individuable. The individuable power of a criterion of identity cannot depend on the existence of other, better or worse qualified, candidates for identity. For, as we saw in sect. 2.1.5., in order for a given evidential state of affairs to be a *sufficient* criterion of *numerical* identity for a given kind of particulars, it has to be able to establish the possibility of a *contrast between strict numerical and merely qualitative identity*. That is to say, it has to provide us with a way of distinguishing -- at least in principle -- between, *on the one hand*, a case in which an object x_1 at t_1 and an object x_2 at t_2 are qualitatively identical, but numerically distinct, and, *on the other hand*, a case in which an object x_1 at t_1 and an object x_2 at t_2 are not only qualitatively, but numerically identical. It has to give substance to the idea of a contrast between the *persistence* through time of one numerically identical particular, and the *consecutive existence* of two qualita-

⁵²Williams, 'Are Persons Bodies', pp. 77/8.

tively identical, but numerically distinct, particulars. Only an inherently (or essentially) individuating criterion of identity can achieve this.

What makes the notion of *spatio-temporal* continuity (as opposed to the notion of *psychological* continuity) inherently individuating, is the uniquely identifying force of an object's *spatial position in a unique and unified spatio-temporal system*. A given material object can only ever be in one place at one time, and two material objects cannot be in one and the same place at the same time. Therefore, two different material objects cannot share one and the same spatio-temporal history, and one persistent material object can only trace *one* path through space and time.

I argued above (sect. 2.1.5.) that our notion of a person is that of a particular. And as we saw in our discussion of Kant's reflections on the 'Paralogisms', I equally cannot but view *myself* as a particular. In other words, the particularity of the referent of 'I' is even *presupposed* by the 'pure point of view of consciousness. Hence, if the psychological continuity criterion is supposed to be a necessary and sufficient criterion of personal identity, it has to be able to establish the possibility of a contrast between numerical and merely qualitative identity for the kind of particulars that persons are. And that, I argue, it cannot do. It cannot provide the criterion of identity of a given *person* (which is a *particular*), but only the criterion of identity of what I called a given '*personality*' -- i.e. a set of in fact highly individuating psychological characteristics, which however in principle are always duplicable. In other words, psychological continuity can only ever be a criterion of identity of persons where this term is understood in a *type-sense*, not a *particular-sense*⁵³. But a person-type, however individuating it may be in practice, is in principle always duplicable. And this is why psychological continuity cannot provide the criterion of identity of a *particular person*.

The cited difference between the 'particular' and the 'type' understanding of the term 'person' may be clarified further by scrutinizing an argument that Shoemaker has given in support of the psychological continuity view of personal identity. In this argument Shoemaker envisages an imaginary society in which brain-state transfers have become commonplace⁵⁴. Each of the members of this society has a couple of duplicate bodies stored in some kind of body-bank, and whenever he feels it necessary to change

⁵³I am here indebted to a remark by Williams, 'Are Persons Bodies?', p. 80, and Wiggins, *Identity and Spatio-Temporal Continuity*, pp.53-55.

⁵⁴ Cf. Shoemaker in Shoemaker/Swinburne, *Personal Identity*, pp. 109f.

bodies, he has his whole 'personality' transferred to a new body. With respect to the people of this imaginary society, Shoemaker argues, *firstly*,

'there would...be a very strong case for saying that what *they* mean by 'person' is such that the BST-procedure [i.e. a brain-state transfer, which is supposed to encode a given brain's total informational state into another brain] is person-preserving (using 'person' in their sense).'

(Here, what Shoemaker means by saying that the procedure is 'person-preserving' is that the people of his imaginary society do not view the 'dis-animation' of their own old bodies, and the consequent 'animation' of their new bodies as equivalent to their own deaths and the creation of a psychologically similar creature. They view it as the continuance of one *numerically* identical being.) Shoemaker then goes on to add a *second* premise:

'But there would also be a strong reason for saying that what they mean by 'person' is what we mean by it; they call the same things persons, offer the same sorts of characterizations of what sorts of things persons are, and attach the same kinds of social consequences to judgements of personal identity.'

And he *concludes* from these two premises that

'if they are right in thinking that the BST-procedure is person-preserving, and if they mean by 'person' the same as we do, then it seems that *we* ought to regard the BST-procedure as person-preserving.'⁵⁵

It should be evident from the previous discussion where we have to disagree with Shoemaker's argument. It is Shoemaker's subscription to the *second* premise, namely that the people of his imaginary society mean by 'person' what we mean by it, where his argument is flawed. *Our* notion of a person is that of a *particular* for which there is a possible contrast between numerical and merely qualitative identity. *Their* notion is that of a *type* for which there is no such contrast. Their notion of a person is the notion of a set of characteristics so specific that *in fact* it is only ever instantiated once. It is the notion of a type which *in principle* could be instantiated indefinitely often, but *in practice* (and that is to say, *contingently*) is only ever instantiated once. Or perhaps we should say that their notion of a person is not so much that of a type as we know it, but the (paradoxical) notion of a 'concrete universal' (as Wiggins has put it⁵⁶) or an ever-developing 'individual type' -- an ontological hermaphrodite, somewhere

⁵⁵All quotations from Shoemaker in Shoemaker/Swinburne, *Personal Identity*, p. 109.

⁵⁶Cf. Wiggins, *Identity and Spatio-Temporal Continuity*, pp.53-55, and *Sameness and Substance*, pp.166-168.

between the particular and the general.

Such a conception of personality is of course not without historical predecessors. Thus, in a modification of Aristotle's theory of forms, *Aquinas* seems to have claimed that for man the form of the body (i.e. the soul) is separable from the body and capable of independent existence. According to Aquinas, the soul of man, unlike the soul of plants or animals, is an '*intellectual substance*'⁵⁷. The problem with this idea is, of course, that forms, or properties, are by their nature universal; that they can be instantiated in many *different* substances. What Aquinas wants is an *essentially particular form*, something that could logically individuate only one particular person. And that he cannot get, since duplication is in principle always possible for any, however specific, set of qualities.

Similarly, *Leibniz*' conception of 'complete notions' may be viewed as an attempt to level out the fundamental difference between the particular and the general. However, as I take it Strawson has shown convincingly⁵⁸, no concept (however 'individual' or 'complete' or specific it may be) can give sense to the idea of particularity which -- at least in our conceptual scheme -- is provided by the uniqueness of an object's position in a unified spatio-temporal system. The fact that only one entity instantiates a given concept must always remain a matter of contingency. And thence if, as I argued, our concept of a person is that of a particular, psychological continuity cannot be its sufficient criterion of numerical identity. The *contingent* uniqueness of a given *mental* history cannot replace the *necessary* uniqueness of a given *spatio-temporal* history as a criterion of identity for persons.

It is interesting to note in this context that *Cartesian dualism*, (according to which *my* identity through time consists in the persistence of a given Ego or *res cogitans*) is less misguided in this respect than the psychological continuity account of personal identity. For, by introducing a separate ontological realm of, essentially and

⁵⁷Cf. *Summa Theologiae*, Part I, quest.50, 4. For non-scholastics (like the present author) I recommend McCabe, 'The Immortality of the Soul', or Kenny, *Aquinas*, pp.43-55. In general, it seems to me that psychological continuity theorists would find a lot to sympathize with in Aquinas' reflections on the nature of angels, cf. *Summa Theologiae*, Questiones 50-64.

I owe the reference to Aquinas to Swinburne, in Shoemaker/Swinburne, *Personal Identity*, p.32.

⁵⁸Cf. Strawson's reflections on Leibniz' monadology, in chapter 4 of *Individuals*, esp. pp. 124-134.

exclusively, *thinking particulars* (Egos, or *res cogitantes*), it at least tries to provide for the possibility of a contrast between strict numerical and merely qualitative identity. According to Cartesianism, *my* Ego (the *res cogitans* which I am) would *not* be the same as yours (the *res cogitans* you are) even if they were completely qualitatively identical. Thus Cartesianism tries to introduce a *separate* realm of individuation *independent of the unified spatio-temporal framework*, -- a realm of individuation populated by entities for which there is the possibility of a contrast between merely qualitative and strict numerical identity. As I shall argue below (sect. 5.6.), the Cartesian account of personal identity fails because it cannot cite any criteria of identity for the putative substances inhabiting this realm -- a failure which is finally due to the fact that the Cartesian *res cogitantes* cannot be identified objectively (nor, I shall argue, from the first-person perspective). Thus Cartesian dualism *tries* to establish the possibility of a contrast between numerical and merely qualitative identity, but it fails in doing so.

The *psychological continuity* criterion, however, cannot even *prima facie* provide for such a contrast, and thus outright fails to fulfil the *contrast requirement* laid down in sect. 2.1.5. above. And the reason for this failure is, it seems to me, that psychological continuity can only count as a form of *qualitative* continuity characteristic of persons; it only invokes -- in principle always duplicable -- (psychological) *qualities* of a given substance (a person), and not the genuinely individuating notion of *spatial* position. This also has the further consequence that the psychological continuity criterion fails on another count (namely with respect to the *tracking-requirement* cited above, sect. 2.1.5.). For it fails to provide a *principle of tracing* a person as numerically identical through time. This can be brought out by the following observations:

Cases of a person's putative body-switch are often described as involving a '*transfer*' of a given set of mental states from one body (or brain) to another. It is not clear, however, in what sense we can understand these imaginary procedures as involving a genuine transfer of anything. Evidently, the putative '*transfer*' of mental states from one body to another cannot be understood merely as a *duplication* of psychological *qualities* highly characteristic of person P_1 in a numerically distinct entity, person P_2 . For psychological qualities are types or universals, not particulars. They can be *instantiated* (indefinitely often) in various different particulars, but they cannot (and this is a logical '*cannot*') be '*transferred*' from one particular to another. Therefore, according to this interpretation, psychological continuity could obviously *not* be understood as a criterion of *numerical* identity. One cannot (by definition) '*make*' two particulars *numerically*

identical by endowing them with the same qualities. There may be certain qualities which are highly individuating of a person being the particular person he is, but the duplication of these qualities in another person does not make these two persons numerically identical.

Some adherents of the psychological continuity criterion endorse some form of *functionalism*⁵⁹. That is to say, they endorse the view that mental states are essentially identified by their causal function in a person's mental life. Some functionalists take the position that mental states are '*essentially informational*' states, which are supposed to be 'realizable' in indefinitely many different kinds of matter. Others think that mental states, although functionally identified, are identical with certain (type or token) *states of a person's brain*.

According to the latter view, the putative 'transfer' of mental states could be understood as a transfer of certain parts of a person's brain, and this understanding would provide us with a genuine notion of 'transfer' involved in a putative body-change. However, according to this variety of functionalism, the psychological continuity criterion would collapse into a version of the so-called brain criterion. That is to say, it would be tantamount to the view that the criterion of personal identity is the spatio-temporal continuity of a given human *brain*. And we have already seen above that the view that personal identity consists in the persistence of a given person's brain cannot count as an elucidation of our *concept* of personal identity.

The former, more radical, version of functionalism (according to which mental states are 'essentially informational') prevents the psychological continuity criterion from collapsing into the brain criterion. However, it at the same time makes it impossible for the psychological criterion even initially to serve as a criterion of *numerical* identity of persons. For, just as qualities, informational states are by definition *types*, for which -- as opposed to particulars -- there is no inherent contrast between numerical and merely qualitative identity. To give an example, it would not make sense to say that the colour (surely a type) of my copy of Strawson's *Individuals* is numerically distinct from the colour of your copy, simply because the one is *this* and the other one is *that* colour. There is no such distinction between numerical and merely qualitative identity for types (as there is for particulars), simply because spatial position does not count as a principle of individuation for types (as opposed to concrete particulars). If so, however, the idea

⁵⁹Notably Shoemaker, in Shoemaker/Swinburne, *Personal Identity*, pp. 92-97.

of a *transfer* of mental states (or even a whole 'personality') from one body to another seems once again to collapse into the idea of a *duplication* of qualitatively identical states in a distinct entity. For just as it would not make sense to claim that one has transferred the colour of one's book to the colour of one's wall if one has painted the wall the colour of the book, it would not, -- according to this understanding of the terms 'mental state' and 'personality' -- make sense to say that a given person's mental states (or whole personality) have been transferred from one body to another. Mental states, just as colours, cannot be described in terms of the distinction between numerical and qualitative identity. Even if my migraine-pain always stopped when yours starts, it would not make sense to say that I have 'transferred' my pain to you. In the sense in which the two of us can have the same pain (say, a sharp pain in the left temple), you could have had it all along. And in the sense in which you cannot, you could not possibly have it⁶⁰.

Thus, as I see the situation, the only way to make sense of a genuine 'transfer' (as opposed to a mere duplication) of mental states is to adopt a fairly full-blooded version of Cartesianism, which most adherents of the psychological continuity criterion would want to refrain from. According to such a conception, mental states are conceived as entities in a second ontological realm distinct from the material world. If we endorse the idea that mental states are made of some sort of 'mind-stuff'⁶¹ -- which, according

⁶⁰ Of course we could *introduce*, we could *give* sense to, such a contrast between numerical and qualitative identity with respect to colours and mental states. We could say that *this* colour cannot be the same colour as *that* colour because this colour is the colour of *this* book, and that colour is the colour of *that*, numerically distinct, book. And this would be treating colours as particulars. However, it is worth noting that the contrast could only be established by indexing the numerical identity of colours to the numerical identity of genuine particulars (such as books), and it is evident that the introduction of such a contrast would be pointless.

Equally, we could say that Jones cannot have the same pain (or personality) as Smith since Jones' pain is *this* pain (or personality) and Smith's pain is *that* pain (or personality). But it is obvious that this introduction of a distinction between numerical and merely qualitative identity for mental states is parasitic on the genuine numerical distinctness between persons such as Smith and Jones. As Strawson has put it in *Individuals*, p.97, 'states, or experiences...owe their identity as particulars to the identity of the person whose states or experiences they are.'. Thus even if one accepted the viability of the numerical/qualitative distinction with respect to mental states, one could not hold that the numerically same mental state could in any sense be transferred from one *person* to another.

⁶¹Cf. Swinburne, in Shoemaker/Swinburne, *Personal Identity*, p. 28.

to the (unknown) laws of this second ontological realm, can be *traced* from one body to another -- we might be able to give substance to the idea that a transfer of 'personality' from one body to another, has to be understood as a body-switch of one numerically identical person. However, if a theorist is willing to go that far, he might as well go the whole way and adopt an even more full-blooded (and arguably more coherent) Cartesianism which claims that persons are essentially and exclusively *mental substances* (and that mental states are modes of these substances). According to this more radically Cartesian conception, a person's putative body-switch could be understood as a transfer of a particular mental *substance* (i.e. the simple and unanalysable referent of 'I', the Ego) from one body to another. Such a so-called Simple View will be discussed and rejected below. For the time being, I shall take it as agreed by anybody but the Cartesian that the obtaining of psychological continuity cannot be considered a sufficient condition of the obtaining of numerical identity of person.

Why then -- and this is a final question I should like to address in this context - - are we so easily tempted to accept a procedure which transfers psychological continuity from one body to another as 'person-preserving', and thus as a kind of body-change rather than the destruction of one person and the creation of a distinct person which is in all psychological respects qualitatively identical with the donor? From what source do the thought-experiments involving putative body-changes derive their persuasive power?

Partly, I think, it is the metaphor of a 'stream of consciousness' that has misled various adherents of the psychological continuity account in this context⁶². We are easily tempted to think of 'Consciousness' as a stream flowing from the past through the present into the future, -- a stream whose identity is fixed by an 'I'-thought in which a given 'Consciousness' refers to itself. And we think that wherever this stream flows there the person goes. Partly, the allure of thought-experiments involving 'body-changes' may also be due to the ease with which each of us can imagine himself (i.e. from the 'first-person perspective') 'inhabiting' a body which is different from his own. Thus each of us can quite easily imagine regaining consciousness on Mars after having entered Parfit's tele-transportation device, or waking up in another person's body. The ease with which we perform these feats of the imagination is to be attributed to the fact that, in doing so, a person does not invoke any criteria of his identity. And this tempts each of

⁶²Cf. Parfit, *Reasons and Persons*, sect.88-90.

us to 'projectively identify' himself with a future person who is in all psychological respects described as qualitatively identical with oneself. Each of us, that is, can imagine having this future person's 'I'-thoughts. And he can also extrapolate the identity of the referent of these 'I'-thoughts over the temporal gap by imagining them potentially united by 'a stream of consciousness flowing through both of them' -- by, as Nagel put it, 'specifying the temporal extension of the present subject'. It is, in other words, a misunderstanding of the 'logical identity' of the subject that gives rise to the 'intuition' that one numerically identical person could at different times of his existence inhabit several different bodies, which in turn tempts us to conclude that psychological continuity is what personal identity 'consists in'. We have already seen that this temptation lasts only as long as, in the description of the thought-experiment, the brain-state donor's body is imagined to be destroyed during the transfer, and as long as the donor's brain-state is copied onto only *one* clone. This observation alone should have made adherents of the psychological continuity criterion wary of the theoretical validity of such thought-experiments. Parfit has famously argued that, with respect to branch-line cases, 'double success can't be a failure'⁶³. I think the correct view of the matter is that double failure can't be a success.

⁶³Parfit, 'Personal Identity', p. 22.

5.4.3. *Psychological Continuity as a Criterion*

In the previous two sections I argued that psychological continuity is neither a necessary nor a sufficient condition of the obtaining of personal identity. This might be taken to imply that psychological continuity is not a *criterion* of personal identity at all, i.e. that -- as some adherents of a Simple View of personal identity have argued (cf. below, sect. 5.6.) -- psychological continuity is merely *inductive* or *symptomatic*, and hence not *meaning-constitutive*, evidence for the obtaining of personal identity. This view, however, can be shown to be wrong.

In sect. 2.1.4. I argued that the mere fact that a certain kind of qualitative identity is defeasible evidence for numerical identity does not by itself make it inductive. For there may be various forms of *qualitative* identity which are such that, if -- in a particular case -- they could not count as evidence for the numerical identity of a ϕ any more, the correctness of the identification of a given particular as a ϕ would itself become questionable. That is to say, the fact that certain qualitative continuities obtain may be a condition of our identifying a particular x as a ϕ , and thus a condition of our *reidentifying* a particular x_2 at t_2 as the *same* ϕ as x_1 at t_1 . And in this way certain kinds of qualitative continuity (namely those which are relevant to the question whether a given particular is to count as a ϕ at all) may be at least partly meaning-constitutive of statements concerning a ϕ 's numerical identity through time.

The same, I shall argue, holds for persons and personal identity. If, that is to say, in a given real or imaginary case it has become undecidable whether a given particular should count as a *person* -- since it does not fulfill our expectations concerning a person's 'typical history and continuity' -- it has become equally undecidable whether P_1 at t_1 is *the same person* as P_2 at t_2 . And in this way certain kinds of qualitative continuity (namely those which are at least partly constitutive of how we determine whether a given particular is a person at all) may be at least partly meaning-constitutive of statements concerning personal identity through time. They will then have to count as genuinely criterial, not only inductive, evidence for personal identity, although -- being merely *qualitative* continuities -- they are neither necessary nor sufficient for the obtaining of *numerical* identity of person.

It is in this sense that psychological continuity can be shown to be meaning-constitutive, not just inductive, evidence for the obtaining of personal identity. That is to say, the general obtaining of the evidential relation between psychological continuity and personal identity can be shown to be a *meaning-constitutive background-condition*

of our identity-claims with respect to persons. If, in other words, -- due to a change in the 'very general facts of nature' -- the evidential relation between psychological continuity and personal identity were to break down on a general level (such that we could not safely conclude any more from the obtaining of psychological continuity to the obtaining of personal identity) our present concept of personal identity would have broken down as well. For in such a case we could not clearly *identify* the resulting entities as *persons* any more. If, I shall argue, bodily identity and psychological continuity did not usually 'go together', our present concepts of a person and of a person's numerical identity over time would lose their applicability⁶⁴.

We have already envisaged one such change in the 'general facts of nature' with respect to Shoemaker's imaginary society (cf. the previous section). I argued that the concept of a person (and of personal identity) employed by the people inhabiting this society is different from ours. However, this is not to say that if any of us lived in Shoemaker's society, we could not *adopt* these people's notion of persons and personal identity. Nor is it to say that, if brain-state transfers became common in our world, we could not develop a very similar notion. But it is important to note that this notion would be an *extension* of our present concept of personal identity, whose logical behaviour differs in important respects. The choice of how to adapt our present concept of personal identity to the altered circumstances would to a certain extent be up to us. As Wittgenstein put it, our present concept would in such a case not have only 'one legitimate heir'⁶⁵. It would, e.g., not be clear what we should call a person's *body* in such altered circumstances. We might call the whole set of animal organisms that a given person 'uses' or 'inhabits' during his life that person's body, and thus stick to our familiar idea that a person can only have *one* body throughout his entire life -- although of course the claim that a person has only one body throughout his life would here mean something very different from what it means in our present use of the term 'person's

⁶⁴This analysis is in a way anticipated by Wittgenstein's remark that '[o]ur actual use of the phrase "the same person" and of the name of a person is based on the fact that many characteristics which we use as the criteria for identity coincide in the vast majority of cases.' (cf. *The Blue and Brown Books*, p.61). However, it seems to me that Wittgenstein underestimated the logical priority of the bodily criterion, which -- via the notion of the spatio-temporal continuity of a given living body -- first opens up the possibility of a contrast between numerical and merely qualitative identity for persons; cf. above section 2.1.5..

⁶⁵Wittgenstein, *The Blue and Brown Books*, p.62.

body'. Alternatively, we might stick to our more familiar use of the term 'person's body' and say that a person can inhabit several bodies during his life, each of which dies once a person has been 'transferred' to another body. How we decide to use our terms in such a case, which way we choose to go, is inessential, since it does not make any difference to the facts. What is essential is to note that the terms 'person' and 'person's body' have here been given a new meaning, and that the consideration of Shoemaker's imaginary case cannot contribute to an elucidation of *our present concept* of person and its position in our conceptual scheme.

One crucial feature of Shoemaker's imaginary society is that any given 'personality' (or 'person-type') is at any given time only ever realized *once*. Another is that the various realizations of a given 'person-type' in various different bodies are imagined to be a *controlled* process. What, however, if we imagined a more radical change in the general facts of nature according to which 'personalities' just started moving around from body to body in a completely unpredictable, *erratic*, manner? In this case, not only could bodily identity not be used any more as reliable evidence for personal identity; but, equally, psychological continuity could not count as a criterion for *numerical* identity of particular persons any more. For all we could trace as identical through time in such a world would be particular 'bodies' on the one hand, and (with certain difficulties) given 'person-types' on the other. But we could *not* any more trace *particular persons* as numerically identical through time. And the reason why we could not *trace* particular persons in this world is that it would not be clear whether the creatures inhabiting this world should be counted, or *identified, as persons* in the first place. In identifying a given particular as falling under a sortal concept ϕ , we do not identify it, as it were, momentarily. Rather, the identification of a given particular as a ϕ includes within it certain expectations concerning a ϕ 's *typical* history and continuity (i.e. certain *qualitative* continuities). And what is true for particulars in general, is true for persons in particular. Although it is true that we identify (and trace the identities of) persons *by* their bodies, we do not simply identify and trace them *as* bodies. The identification of a given particular *as* a person is based on the background-condition that a given 'person-type' is associated with only one body, and a given body with only one 'person-type'. And if this background-condition does not (or is imagined not to) obtain any more, then the notion of a person becomes inapplicable.

This may also be brought out by considering a world in which persons' *personalities were all the same*, in the sense that every belief, memory, character-trait, etc.

acquired by one such creature would be immediately shared by every other⁶⁶. If we had such intercommunicating experiences, if we were thus linked together by a 'group mind', psychological continuity would have lost its distinguishing force as a criterion of personal identity. Of course we could distinguish and trace the different creatures by their numerically distinct *bodies*. But it would be unclear in what sense what we thus distinguish and trace could be said to be different *persons*⁶⁷. For it would be unclear whether these creatures should be counted as persons at all. And this, I think, shows once again that psychological continuity is not just symptomatic, but genuinely *criterial*, evidence for the obtaining of personal identity. If it could not generally be used as reliable evidence for personal identity -- and that is to say, if psychological continuity and bodily identity did not usually go together -- our present concepts of a person and of personal identity would become inapplicable.

5.5. 'Body-Switches' and the First-Person Perspective

In the previous section I argued that the contingent fact that bodily identity and psychological continuity usually go together is one of the background-conditions of the applicability both of the concept of a person and of the concept of numerical identity with respect to persons. This claim can also be substantiated in a different way, namely by considering cases in which these two criteria for personal identity actually conflict. This, I shall argue, is the case in thought-experiments involving putative *body-switches*.

The imaginary cases discussed so far only involved putative *body-changes*. That is to say, they envisaged the 'disanimation' of what used to be a person's body followed by the consequent 'animation' of a *new* body (via the dubious idea of a 'transfer' of psychological qualities from one body to another). In saying that these thought-experiments envisaged the 'animation' of a *new* body, I mean that it was implicitly assumed in these thought-experiments that the body to be 'informed' or 'animated' with a given 'personality' had not been *another person's* body before. That is to say, it was assumed that the body which a given person was imagined to 'change in' for his old one had not before been 'inhabited' by any other person. It was conceived as a 'mere body' without any psychological qualities. Typically, in the respective thought-experiments this is

⁶⁶I once saw a film entitled *The Midwich Cuckoos* which envisaged just this possibility.

⁶⁷Cf. Strawson's discussion of the topic in *Individuals*, pp 112-114.

achieved by imagining the new body to be created basically out of nothing at the moment of 'animation' (as in Parfit's teletransportation case⁶⁸) or to be artificially created and stored in a frozen or coma-like state in a kind of body-bank (cf. Shoemaker's imaginary people described above).

I have already argued (sect.5.4.2.) that the only sense in which we can interpret these so-called 'brain-state transfers' as 'person-preserving' is in a *type* or *universal*, not a particular, sense of the term 'person'. That is to say, we can conceive the 'animation' of a new body with a given 'personality' as the creation of a person who is in all psychological respects *qualitatively* identical with the 'personality-donor'. But we cannot understand it as *one* (numerically identical) person's body-change. For our concept of a person is that of a *particular*, not a type or universal. And hence -- unless we opt for a full-blooded Cartesianism -- the 'transfer' or duplication of psychological qualities cannot be understood as one *person's* body-change. Equally, since the use of the pronoun 'I' presupposes -- though it cannot by itself give substance to -- the *particularity* of its referent, I cannot conceive of these imaginary brain-state transfers as cases of *myself* having changed bodies. Thus brain-state transfers are irrelevant to the fate of the 'brain-state' or 'personality'-donor. Depending on what is imagined to happen to him after the 'transfer', he either dies or survives. (And, equally, if *I* am the brain-state donor, I either die or survive). Thus imaginary cases envisaging putative 'body-changes' are *conceptually decidable* in a straightforward manner.

The cases I should like to consider now are imaginary cases of putative body-switches (as I shall call them in order to distinguish them from the previously discussed body-changes). In body-switches (as opposed to body-changes) the body which a given person is imagined to 'change in' for his old one has typically been 'inhabited' by another person before. That is to say, the body to be 'informed' or 'animated' with a given 'personality' has typically been *another person's* body. My aim in discussing these putative body-switches will be twofold:

Firstly, I shall try to show that -- as opposed to the previously discussed body-changes -- imaginary cases of body-switching pose *conceptually undecidable* questions of personal identity. That is to say, they present us with situations in which the question 'Is P_1 at t_1 the same person as P_2 at t_2 ?' has lost its ability to pose a Yes/No-question to reality. And this, I shall argue, is due to the fact that in thought-experiments envisaging

⁶⁸Cf. Parfit, *Reasons and Person*, ch.10, pp.199-219.

putative body-switches one of the *background-conditions* of the applicability of the notion of personal identity (namely that the criteria of bodily identity and psychological continuity do *not* conflict) is imagined *not* to obtain.

Secondly, I shall rebut an objection to this analysis. This objection claims that the analysis of putative body-switches as conceptually undecidable must be flawed, since it *cannot be accepted from a first-person perspective*. I shall argue that the objection rests on a paralogistic misconception concerning the validity of the testimony of the first-person perspective in matters of personal identity.

1. A case of putative body-switch has been discussed by Bernard Williams in his article 'The Self and the Future'. In this article Williams presents us with a dilemma arising from a juxtaposition of two puzzle-cases:

The first case presents us with two persons, A and B, whose 'personalities' are going to be exchanged via a brain-state transfer device. From this procedure what Williams calls the A-body person (i.e. the person who now 'inhabits' the body which used to be A's) emerges with B's personality, and the B-body person emerges with A's personality. Before being subjected to their respective 'brain-zaps' both A and B are told that either the A-body person or the B-body person is going to receive a substantial sum of money, whereas the other person is going to be tortured. Both A and B are then given the choice of which of the two resulting persons (the A-body person or the B-body person) is going to receive the money, and which of the two persons is going to be tortured. Williams suggests that -- provided their choice is influenced by purely egoistic motives -- the two protagonists would each choose the money to be given to the person with his own personality and the other person's body, and the torture to be performed on the person with his own body and the other person's personality. That is to say, he suggests that it would be rational for both A and B to view the experiment as a case of body-change rather than 'mind-swap'. Both A and B, it is suggested, and everybody who would find himself in their situation, would be wise to 'identify himself' with the future person who (according to the description of the case) is *psychologically continuous* with him, rather than with the future person who is going to have his present *body*. Williams takes this first imaginary case to suggest that bodily identity is not a necessary condition of personal identity, i.e. that -- given the situation is as clear-cut as presented in his thought-experiment -- we would accept psychological continuity as the (necessary and) sufficient criterion of personal identity.

The second, apparently quite different, case (which is deliberately presented in

the first-person mode) presents the reader with the following scenario: the reader is invited to imagine finding *himself* in the hands of a mad scientist. He is told by the scientist that he is going to be tortured tomorrow, but that before then he is going to be connected to the brain-state transfer device, by means of which his whole personality will be erased from his brain and another person's (say his neighbour's) personality (or brain-state) will be 'stored onto' his brain. Williams suggests that, if one found oneself in such a situation, one could hardly console oneself with the thought that at least *one-self* was not going to be tortured, but rather somebody else (i.e. the person with whose personality one's former body has been 'informed'). Rather, as he puts it:

'Fear, surely, would be the appropriate reaction: and not because one did not know what was going to happen, but because in one vital respect at least one did know what was going to happen -- torture, which one can indeed expect to happen to oneself, and to be preceded by certain mental derangements as well.' (p.52⁶⁹)

This, of course, suggests that in the second case one identifies oneself with the future person 'inhabiting' one's present body, i.e. one applies the bodily criterion to answer the respective identity-question.

The *dilemma* which Williams sees embodied in these two different cases arises from a reflection on the fact that the second case is merely one side, differently described, of the first case. For, assuming *I* found myself in A's situation as described in the first case, although in both cases what is imagined to happen to *me* is exactly the same, in the first case, I would be inclined to identify myself with a future person who will be psychologically continuous with me, whereas in the second case I would be inclined to identify myself with a future person who will have my present body. If so, however, we are faced with a conundrum. As Williams puts it:

'...it looks as though there are two presentations of the imagined experiment and the choice associated with it, each of which carries conviction, and which lead to contrary conclusions.... Following from all that, I am not in the least clear which option it would be wise to take if one were presented with them before the experiment. I find that rather disturbing.' (p. 61f.).

Williams intentions in setting up the cited dilemma are not entirely clear. Sometimes he seems to take the dilemma to show that there is some sort of antinomy at the very heart of our notion of personal identity; sometimes one gets the impression that Williams (himself the main champion of the bodily criterion) tries to present a *reductio ad absurdum* of the psychological continuity criterion. It is obvious, however, that he

⁶⁹All references in the text are to Williams, 'The Self and the Future'.

believes that the juxtaposition of his two puzzle-cases presents us with a genuine dilemma. I think we already have all the elements we need in order to defuse Williams' dilemma. We just have to bring them in the right relation to each other.

With respect to the first horn of Williams' putative dilemma, we have already seen that -- barring a full-blooded Cartesianism -- there is no sense in which one numerically identical particular person can change bodies. Thus, faced with Williams' choice, A would be ill-advised to choose the B-body person (who is conceived to be psychologically continuous with him) to be given the money and the A-body person to be tortured. For the B-body person, whoever he is, is *not* A. However, with respect to Williams' first case, it seems equally implausible -- and this is the difficulty posed by 'body-switches' as opposed to 'body-changes' -- to describe the situation after the brain-state transfer by claiming that the A-body person is (identical with) A. If the A-body person's brain had been left, as it were, 'blank', we might have been able to describe the situation by saying that the brain-state transfer has reduced A to a 'vegetable'; that, in other words, A is still alive, though alive in a coma-like state. Our concept of personal identity *can* account for cases where psychological continuity, as it were, *breaks off*. Although somebody in an irreversible coma is clearly a borderline-case for being counted as a person, we can make room for such events in our notion of a person. We can make room for the fact that the psychological criterion points, as it were, *nowhere*.

However, as things stand, with A's former body having been 'informed' with B's personality, it seems as implausible to say that the A-body person is A as it would be clearly false to say that he is B. All that is left of A, and hence all that is left for us to 'latch onto' in identifying the A-body person with A, is what used to be A's living animal body. However, whereas A's falling into a coma is a possibility provided for by our concept of a person as a rare kind of 'typical continuity' that a person's life can display, A's becoming completely and, most of all, *systematically* deluded about his own identity (about *who* he is) is not thus provided for. And this is just what we would have to say if we identified the A-body person with A. We would have to say that the brain-state transfer has turned A into a madman, who -- due to having B's brain-state copied onto his brain -- has come to think that he is B. Now there are various forms of madness which push our concept of personal identity to its limits, but which can still be accommodated by it. However, they are not nearly as radical as the scenario envisaged in Williams' total 'brain zap'. And this is why, it seems to me, the A-body person's predicament is not one of those unusual continuities which, although they push the

concept of personal identity to its limits, can still be accommodated by it. Here, I think, those limits have been transgressed. The changes happening to A which resulted in the existence of the A-body person are not by any stretch of the imagination 'continuities' which could be called 'merely unusual' events in a person's life. And that is to say, they are simply *not provided for* by our concept of a person. And this is why it would be equally implausible to describe the result of the brain-state transfer by saying that A survives as a (completely deranged) A-body person, as to say that A has died. (For his body is, after all, still alive). The A-body person is neither clearly A, nor clearly not A (though he is clearly *not* B). The truth is that the resulting situation simply cannot any more be described in terms of our notion of personal identity. The A-body person is neither identical with nor distinct from with A. The concept of personal identity has here lost its application. It has become undecidable whether it applies or not.

The *reason* for this conceptual undecidability is of course that Williams' case imagines one of the *background-conditions* of our identity-discourse concerning persons (and of our practice of *naming particular* persons) *not* to obtain. The notion of a person, I argued above, is based on the contingent fact that the persistence of a given animal body and the continuity of a given 'personality' go together. Our concept of a person provides for the possibility that the psychological continuity criterion points, as it were, nowhere. But it does not provide for the possibility that the psychological continuity criterion and the bodily criterion point in *opposite* directions. In such a case we have lost any hold on the notion of a particular which could be called a person.

Of course, in saying that the situation envisaged by Williams' thought-experiment cannot be described in terms of the notion of personal identity, I do not want to deny that it may still be described in other terms. We may say that the A-body person is a creature who has (what used to be) A's living animal body and is psychologically continuous with (what used to be) B. And, equally, we can describe the B-body person as a creature who has what used to be B's living animal body and is psychologically continuous with A. The facts of the matter are clear enough. But they cannot be described in terms of the notions of persons and personal identity. No rules have been laid down in our concept of a person to deal with such a situation. Here the concept has been torn apart into two different directions. The bounds of sense have been crossed.

And that the bounds of sense have been crossed may also be taken to explain the *wavering of our intuitions* with respect to the two aspects of Williams' case, which mislead Williams to view this case as posing a genuine dilemma. Thus Williams' puta-

tive conundrum should be understood, not as posing a dilemma, but rather as a particularly picturesque illustration of the claim made above (sect.5.1.) that, once the bounds of sense have been transgressed, our intuitions start to waver and become increasingly unreliable. In other words, the fact that -- as Williams shows -- our intuitions begin to waver, could be taken as additional evidence that in the envisaged scenario it has become undecidable whether our concept of personal identity applies or not.

2. Williams anticipates, and rejects, this analysis. In what follows I shall try to show that Williams' argument for rejecting it rests on a paralogistic misconception.

In fact, Williams distinguishes, and rejects, two possible objections to his case being a genuine conundrum. The *first* objection states that here we are dealing with cases in which the question whether a given future person P_2 is identical with a given present person P_1 has become '*conceptually undecidable*'. But this objection, Williams says, cannot be right, because -- although, he says, it may seem intellectually comforting from a third-person perspective -- from the *first-person perspective* it is simply unacceptable. In his own words:

'...this well-known piece of advice [i.e. to regard the question whether the A-body person is going to be identical with A is conceptually undecidable], sensible as it is in many cases, seems in the present case to involve an extraordinary difficulty. It may intellectually comfort observers of A's situation; but what is he supposed to make of it? To be told that a future situation is a borderline one for its being myself that is hurt, that it is conceptually undecidable whether it will be me or not, is something which, it seems, I can do nothing with; because, in particular, it seems to have no comprehensible representation in my expectations and the emotions that go with them' (p.58).

The *second* objection, which Williams calls the 'close friend and neighbour' (p.61) of the first, is that whether the A-body or the B-body person is going to be identical with A, is a matter of '*conventionalist decision*'; that, in other words, A's future identity is up to what people will choose to say about the matter. And if they find the B-body person a suitable candidate for identity with A, then (whatever people might be inclined to say in the second case in which there is *no* suitable candidate for identity with A other than his body) we should take this decision of theirs as settling the matter. Again, Williams invokes the first-person perspective in rebutting this objection to viewing the two cases as posing a genuine conundrum. He says:

'This line of talk is the sort of thing indeed appropriate to lawyers deciding the ownership of some property which has undergone some bewildering set of transformations... But as a line to deal with a person's fears or expectations about his own future, it seems to have no sense at all' (p.61)

Williams is certainly right in saying that -- with respect to the claim that the

identity of A is a matter of 'conventionalist decision' -- if one found oneself in a situation in which one's future identity would be, as he puts it, 'conceptually shadowed' (p.60), it would be absurd to let one's emotional response to this future situation be determined by what people would, or would be inclined to, say about the case. It would be absurd to either sigh with relief or fall into despair depending on whatever linguistic decisions other people would make. If I were told that my body was going to be tortured tomorrow, but that before that my whole 'personality' was going to be erased from my brain, my worries about this future event would not be laid to rest simply by being told that my relatives, or in fact all members of my language-community, would *not call* this body (or the person 'inhabiting' this body) RS. And, alternatively, if I was told that people *will call* this person RS, ought I then to cringe in terror at the prospect of being tortured? Certainly not. What people will decide to say about the case cannot make the slightest difference to my fears about *my* future. However, from this it does not follow that, as Williams seems to think, the first-person perspective effectively precludes the possibility of 'conceptually undecidable' cases of personal identity. This can be shown by applying the lessons we learnt in our discussion of Kant's analysis of the paralogistic illusion in sections 4.1. and 4.2. above.

There we saw that, although the pure 'point of view of consciousness' can provide for the idea of the *particularity* of the entity thus referred to (and thus for the idea of a possible contrast between numerical and merely qualitative identity), it cannot by itself fill it in. It cannot by itself fill it in because (cf. sect. 3.4.) a person's reference to himself-as-subject cannot be said to be a case of *identifying* reference. The first-person perspective, in other words, is *not inherently identifying*, -- although it presupposes the *identifiability* of the particular *non-identifyingly* referred to by 'I' used as-subject from a *third-person* perspective. A person's reference to himself-as-subject, we have seen, does not draw on any criteria of identity of the particular thus referred to. Thence, far from being able to provide the thinker with any criteria of his identity, a person's self-reference by 'I' depends for its *determinacy* (and, that is, for its ability to pin down, or single out, a given particular) on the *objective* identifiability of the referent of 'I' *as a particular*. The identity of the referent of 'I' can only be determined from an objective point of view, since only the objective perspective can provide criteria for its identity. The first-person perspective is thus *essentially* dependent on the objective perspective to fill in, or give substance to, the idea of the particularity of the referent of 'I'.

I have argued in section 5.4. that -- barring a full-blooded Cartesianism -- only

the spatio-temporal continuity of a person's *body* can give substance to the idea of the particularity (and hence the possibility of a contrast between numerical and merely qualitative identity) of the referent of 'I'. I also argued that the notion of personal identity depends for its applicability on the general 'coincidence' of bodily identity and psychological continuity. That is to say, I argued that the (admittedly contingent) fact that the criterion of bodily identity and the criterion of psychological continuity point in the same direction, is a background-condition of our identity-discourse on persons. If, in a given imaginary case, this background-condition is imagined *not* to obtain, then the concept of numerical identity of person has lost its applicability. The statement 'P₁ at t₁ is identical with P₂ at t₂' has then lost its ability to represent a given possible state of affairs which could be said to either obtain, or not -- *tertium non datur*.

If so, however, and if -- as I argued above -- the first-person perspective is essentially dependent (in the specified respects) on the third-person perspective, then the question 'Will it be *me* who...?' (or 'Will this future person be *me*?') has here *equally* lost its ability to determine reality save for a Yes or No. It has equally lost its ability to represent a possible state of affairs which must either obtain, or not obtain -- *tertium non datur*. For, as we have seen, the ability of the 'I' (when used as-subject) to pin down, or single out, a certain determinate particular depends essentially on the identifiability (as a particular) of the referent of 'I' from an objective point of view. Thence if, in a given imaginary case, the identity of the 'I'-thinker has become 'conceptually shadowed', then so has the referent of his 'I'-thoughts. (For the first-person perspective is not autonomous from the third-person perspective with respect to the determinacy of its referent.) This, however, is exactly what happens in Williams' imaginary experiment. In Williams' imaginary case one of the background-conditions of our identity-discourse concerning persons is imagined not to obtain. And therefore the question 'Is P₂ at t₂ identical with P₁ at t₁?' has here lost its ability to pose a Yes/No question to reality. And therefore the question 'Will it be *I* who...?' has here equally lost its ability to pose a Yes/No-question to reality.

Williams' main argument against viewing his imaginary case as conceptually undecidable is that one cannot represent this possibility to oneself; that, as he puts it, 'to be told that it is conceptually undecidable whether it will be me or not, is something which... I can do nothing with' (p. 58); that 'there is an obstinate bafflement in mirroring in my expectations a situation in which it is conceptually indecidable whether I occur.' (p.59). However, it should be clear by now that this refusal to accept the possi-

bility of conceptually undecidable cases of personal identity is based on the paralogistic idea that 'I' used as-subject can single out a given entity *independently* of the objective perspective. It is based on the Nagelian 'sensation of picking out something whose identity over time is well defined, just by concentrating on [one's] present experience and specifying the temporal extension of its subject'⁷⁰.

I hope to have shown that this 'sensation' is a delusion based on a misconception of the purely 'logical' identity of the subject. Far from providing an insuperable counterargument to the claim that there can be conceptually undecidable cases of personal identity, the fact that one cannot 'represent to *oneself*' the idea that one's future identity will be conceptually shadowed, that one cannot 'mirror it in one's expectations', is in way just what was to be expected. For there is no reason to suppose that one should be able to come up with an emotional response to a situation that one cannot really describe, and thence envisage, in terms of one's own conceptual scheme. The use of 'I' as-subject, we have seen, presupposes the applicability of the notion of numerical identity to the referent of 'I'. Thence it is not astonishing that the first-person perspective cannot 'represent' a situation in which this notion has become inapplicable.

5.6. *The Simple View*

The Simple View of personal identity is a syndrome of various theses concerning self-reference and personal identity, which support each other like bricks in an arch. I have already tried to pluck out one of these bricks above (cf. sect.4.1.), where I scrutinized and rejected the thesis that a person's reference to himself-as-subject is a case of *identifying* reference to a necessarily simple and unanalysable substance. In the following I should like to discuss a thesis that is often presented as a direct consequence of the above one, namely the thesis that personal identity through time is a '*simple and unanalysable fact*' which is distinct from any observable or experienceable fact that might be cited as evidence for it⁷¹. Such a thesis was probably first explicitly propounded by

⁷⁰Nagel, 'Subjective and Objective', p.200.

⁷¹Cf. Swinburne in Swinburne/Shoemaker, *Personal Identity*, p. 20: 'The simple view claims explicitly that personal identity is one thing and the extent of similarity in matter and apparent memory another.' Cf. also op. cit., p.21: 'As I shall understand it, the simple view is simply the view that the truth about personal identity is not analysable in terms of the fallible empirical evidence for it of brain and memory continuity;'.

McTaggart who argued that, since self-hood is a simple and unanalysable property (i.e. since what it is for a given set of properties to be *mine* cannot be further analysed) the property of being *the same self* or person through time could not be analysed either⁷². However, insofar as this view is held to imply that, as it has been put, ‘personal identity cannot admit of degrees’⁷³, or that there cannot be any borderline-cases (and thence *empty* identity-questions) of personal identity, it goes back as far as Reid and Butler, whom contemporary adherents of the Simple View often cite as their distant predecessors (cf. the passages cited in sect.2.2., p.41).

It is not easy to separate the true from the misconceived in this context. It seems to me that there are several different understandings of the claim that personal identity is a ‘simple and unanalysable fact’, which should be carefully distinguished. First of all, I think that the view that personal identity cannot ‘admit of degrees’ (sometimes referred to as the ‘determinacy-thesis’⁷⁴) is trivially true. As has justly been pointed out by Mackie⁷⁵, identity *never* admits of degrees. A given object x_1 either is, or is not, identical with a given object x_2 . It is only with respect to the application of the *criteria* of identity for particular *kinds* of entities that matters of realization by degree come to bear upon questions of identity. Cases in which a sub-average realization of our criterion of identity for an entity of a given kind makes it doubtful whether we should apply the concept of identity or not may be called *borderline-cases* of identity, but they are not cases in which ‘identity admits of degrees’.

Sometimes the thesis that personal identity over time is an ultimate unanalysable

Cf. also Maddell, *The Identity of the Self*, p.124: ‘The conclusion to which one is compelled is that personal identity through time is unanalysable; what unites my experiences [...] is just their being mine’.

A similar conclusion, though for different reasons, has been reached by Lowe, *Kinds of Being*, p.137, where he says that personal identity ‘consists in nothing but itself’ and that it is ‘primitive and ungrounded’.

⁷²Cf. McTaggart, *The Nature of Existence*, vol.2, ch.36.

⁷³Cf. Maddell, *The Identity of the Self*, p.122.

⁷⁴Cf. Noonan, *Personal Identity*, pp. 128-130.

⁷⁵cf. Mackie, *Problems from Locke*, p. 148: ‘The concept of identity is not one which can be made more or less stringent; identity is an all-or-nothing affair. Where there is room for relaxation and indeterminacy is in the individual concept to which identity is applied.’

fact is taken to mean that what it is for a temporally dispersed series of experiences to be one and all those of a given person cannot be analysed in purely 'objective' terms, i.e. in terms which do not make use of the notion of a self or person as the *bearer* or *owner* of these experiences⁷⁶. In Maddell's words:

'...there can be nothing which unites a series of experiences such that they are all the experiences of the one person, except their being, unanalysably, mine (yours, his)'⁷⁷.

Again, in one understanding, this version of the unanalysability-thesis seems to me trivially true. The unanalysability of what it is for an experience to be mine is simply due to the fact that I do not apply any criteria in order to check whether a past, present, or future experience was, is, or will be mine. That, to cite Kant once again, 'in the whole time in which I am conscious of myself, I am conscious of this time as belonging to the unity of myself'⁷⁸, is trivial consequence of the 'logical identity' of the subject.

Often, however, the unanalysability-thesis is taken to mean something more substantial than the trivial truth just stated (though it may well be that this stronger claim is derived by a logical *non-sequitur* from the first⁷⁹): namely that personal identity is itself criterionless. This stronger version of the unanalysability-thesis amounts to the claim that there is no genuinely criterial, and thence meaning-constitutive, evidence for personal identity; that all evidence which can be cited in favour of, or against, the obtaining of personal identity is purely inductive, or symptomatic. And then the unanalysability-thesis is taken to imply that any claim to the effect that a person P_2 at t_2 is numerically identical with a person P_1 at t_1 is *logically independent* of any claim to the effect that P_2 at t_2 is psychologically continuous with P_1 at t_1 , or that P_2 at t_2 has the numerically same body as P_1 at t_1 . Bodily identity and psychological continuity, according to this version of the unanalysability-thesis, may be called criteria of personal identity, but only in the sense of purely *inductive*, and thence *not* meaning-constitutive evi-

⁷⁶Cf. Maddell, *The Identity of the Self*, pp. 134-136.

⁷⁷Maddell, *The Identity of the Self*, p.135.

⁷⁸Kant, *Critique of Pure Reason*, A362.

⁷⁹Cf. Strawson's critique of Chisholm in 'Reply to Chisholm', p. 186.

dence (and that is, not in the sense invoked in this essay)⁸⁰. From this understanding of the unanalysability-thesis it follows that a person P_1 at t_1 can survive as P_2 at t_2 without either being psychologically or bodily continuous with P_1 in any way whatsoever; and, vice versa, that a person P_1 at t_1 may be both psychologically as well as bodily continuous with P_2 at t_2 but still not be the very same person.

This version of the unanalysability-thesis is often supported by thought-experiments allegedly showing that a person could survive any, however radical, physical as well as psychological change whilst remaining the numerically same person. Thus McGinn has claimed that he could imagine surviving a combined personality-cum-body-switch if only he remained conscious throughout the whole procedure⁸¹. And Swinburne reminds us repeatedly of the imaginability of disembodied existence or of a person's wading through the waters of Lethe and being given a new body, whilst still remaining the numerically same person⁸². I already cautioned above that in view of paralogistic illusions to which the 'logical identity of the subject' gives rise we should not attach too much significance to the alleged imaginability of these situations. Just as I can imagine being a different person from the one I am, or even (following Nagel) being no particular person at all (without it therefore being logically possible that RS be a different person from the one he is, or that RS be no particular person at all) I may be able to imagine losing all my (physical as well as psychological) properties (without it therefore being a genuine logical possibility that the particular person that I am, RS, can lose all his characteristic properties and still remain the numerically same entity⁸³). Just as a person's ability to perform various feats of transference does not show that this particular person could have been anybody else than who he is, a person's ability to imagine

⁸⁰Cf. Chisholm's distinction between what he calls 'truth-conditions' and 'criteria', in *Person and Object*, p. 112; Swinburne's criticism of what he calls the 'empiricist view', in Shoemaker/Swinburne, *Personal Identity*, pp. 35-48; and Maddell, *The Identity of the Self*, pp.117-123.

⁸¹Cf. McGinn, *The Character of Mind*, pp.111-117.

⁸²Cf. Swinburne, in Swinburne/Shoemaker *Personal Identity*, p.25. Cf. Chisholm, 'The Loose and Popular and Strict and Philosophical Senses of identity', p.138: 'my future experiences need not be linked by any of our present criteria of identity to my present self'.

⁸³And in fact it is not quite clear that we really can consistently imagine ourselves disembodied, cf. Strawson, *Individuals*, pp. 115/6.

surviving various procedures (even his own physical death) does not show that this person could survive these procedures.

The claim that the obtaining of personal identity is logically independent of the obtaining of *any* kind of evidence for personal identity has allowed the supporters of the Simple View to present putative solutions to various of the puzzle-cases discussed in the recent literature on the topic. Thus with respect to fission-cases in which a person P_1 's psychological continuity is supposed to branch to two successor persons P_2 and P_3 , Chisholm and Swinburne have claimed that we (including P_2 and P_3) may not be in a position to *find out* which of the two successors P_2 and P_3 is identical with P_1 (or whether in fact any of the two is identical with P_1), but that this does not prevent the statement ' P_1 is identical with P_2 ' from being either true or false. That is to say, they have claimed that although in these imaginary situations we may not be able to find the truth of the matter, there still has to be a *fact of the matter* as to personal identity⁸⁴. Equally, with respect to a temporally reversed situation, it has been argued that a person who found himself the victim of Williams' imaginary mad scientist would not be in a position to *know* which of the two persons emerging from the mad scientist's experiment would be him (whether he would be identical with the person who will have his present body, or whether he would be identical with the person who will be psychologically continuous with him). But he would have to be either one or the other (or in fact neither). For either he is going to be tortured, or he is not. *Tertium non datur*. 'No half-tortures will be laid on', as Swinburne has ironically put it⁸⁵.

I shall start by examining the suggestion that the relation between *all* forms of evidence for personal identity, and personal identity itself, is purely symptomatic, or inductive. This suggestion is *prima facie* baffling. For it is essential to a relation's being inductive that it was established *in experience*, which in turn implies that the states of

⁸⁴Cf. Chisholm, *Person and Object*, pp. 111-113; and equally Swinburne, in Swinburne/Shoemaker, *Personal Identity*, p.20, where he claims that in a case of fission in which the two successor-persons are equally continuous with their predecessor, in at least one case the evidence for personal identity must be misleading.

⁸⁵Swinburne, 'Personal Identity', p. 237/8. Cf. also Chisholm, *Person and Object*, p.111: '...one's question 'Will that person be I?' or 'Will I be he?' always has a definite answer... There is no possibility whatever that *you* could be *both* the person on the right and the person on the left. Moreover, there *is* a possibility that you would be *one or the other* of those two persons. And finally, *you* could be one of those persons and yet have no memory at all of your present existence.'

affairs inductively related have to be *identifiable independently* of each other. Now this is clearly the case for the various forms of *purely inductive* evidence for personal identity discussed above. The meaning of any statement that P_1 at t_1 is numerically identical with P_2 at t_2 is *logically independent* of any statement that P_1 at t_1 has the same blood-group (or the same fingerprints, or the same genetical structure, or even, we saw, the same brain) as P_2 at t_2 . However, as I have tried to show above, this logical independence does not hold for other forms of evidence such as psychological continuity and bodily identity. For if, I argued, we imagined the evidential relation between psychological continuity and personal identity, or between bodily identity and personal identity, distorted, then our present concept of personal identity would have become inapplicable. And from this I concluded that psychological continuity and bodily identity have to be regarded as genuinely criterial (and that is *meaning-constitutive*, not just inductive) evidence for personal identity.

Another way of making the very same point would be to say that there cannot be *any more direct access* to a person's identity over time than via the continuance of his personality and the persistence of his body. If we had such an access to the obtaining of personal identity, which could be said to be more direct than, or at least *independent* of, the obtaining of psychological continuity or bodily identity, then the thesis that all criteria for personal identity are merely inductive, and thence not meaning-constitutive, could be true. And in fact I think it is exactly the idea of such a more direct and evidence-independent access to personal identity on which the Simple View is based, and from which it derives its strength and initial plausibility. This is the idea of a *special first-person access* that every person has to *his own identity* over time -- an idea that stems from, and is fed by, the first-person perspective. This is in fact quite explicit in Swinburne and Maddell. Thus Swinburne claims that

'the continuing existence of a person over a very short period of time is something which can often be experienced by that person'⁸⁶.

And he goes on to justify this claim by the following Jamesian reflections on the 'stream of consciousness:

'...what the subject of experience is conscious of is of those experiences as his, that is as having a common subject. For the subject in claiming that not merely his present experiences of the world but the experiences which stream into it are his, is claiming that one person experiences them....So my experience of continuing change is

⁸⁶Swinburne in Shoemaker/Swinburne, *Personal Identity*, p.42.

the experience that my experiences of certain small changes are experienced in succession by a common subject.⁸⁷

Maddell avoids the Cartesian 'objectivisation' of the referent of 'I' as an (essentially thinking) Ego. As opposed to Swinburne, he realizes that the introduction of an (essentially mental) Ego as the referent of 'I', and as the bearer of all *my* experiences, cannot solve the problem it is introduced to solve. For the suggestion that what makes all my past, present and future experiences mine is their being had by a given Ego immediately raises the question what makes this given Ego mine, i.e. what determines that I am *this* Ego, rather than another one⁸⁸. But Maddell does not seem to see that this exposition of the Cartesian fallacy at the same time should prevent him from interpreting a person's ability to refer to himself as the subject of a given mental history (without the application of any criteria of his identity) as a *special access* to what this identity consists in (namely simply 'being myself'). He does not realize, in other words, that the idea that a person has a special access to his own identity is a paralogistic illusion based on a misinterpretation of the purely 'logical identity of the subject' as the numerically identical persistence of a certain particular person. To quote Kant once again: the fact that 'the 'I' which accompanies, and indeed with complete identity, all representations at all times in *my* consciousness' must not be misunderstood as 'the objective permanence of myself'⁸⁹.

Another way in which to represent the Simple View's paralogistic invoking of the first-person perspective would be to interpret it as holding that the only constitutive criterion for a person's P_1 being numerically identical with a person P_2 is, trivially, but necessarily and sufficiently, P_1 's being identical with P_2 . Identity is (trivially, but necessarily) 'what makes' P_1 identical with P_2 . Now this statement can be understood simply as a blunt refusal to explain the meaning of our identity-claims with respect to persons in terms of the persistence-conditions of this particular kind of entities. But it can also be understood as the postulation of an *essentially private criterion of personal identity*, which can only be applied by the person whose identity is in question, namely the

⁸⁷Swinburne, in Shoemaker/Swinburne, *Personal Identity*, p. 43f.

⁸⁸Cf. Maddell, *The Identity of the Self*, pp. 135-137.

⁸⁹Kant, *Critique of Pure Reason*, A362/3.

criterion expressed by 'being I'⁹⁰. *Being me* is supposed to be what 'constitutes' *my* identity over time, and it is what still gives sense to personal identity-claims in situations (such as branching psychological continuity or putative body-switches) in which our usual criteria for personal identity have become inapplicable. *Being me* is supposed to be what gives sense to the question 'Will it be me, or not?' and thus what endows this question with the ability to determine reality save for a Yes or No. As I have argued above, there is no such essentially private criterion of identity for persons. A person's reference to himself-as-subject is not an identifying reference, it is not *inherently* individuating (i.e. via its 'intentional content'). Its individuating power depends on the fact that the entity referred to by 'I' is also identifiable *objectively* as the particular entity it is. And that is to say, it is identifiable *independently* of the act of reference in which it refers to itself. Once this background-condition of the use of 'I' is removed, once the entity referring to itself cannot be clearly identified -- once, we might say, the identity of the self-referring entity has become 'conceptually shadowed' -- the use of 'I' has lost its individuating power. Thus 'being I' (or 'being *the same* I') is a form of evidence that is *in principle inapplicable*, from the third-person perspective (trivially), but *also from the first-person perspective*. As such, it cannot fulfil the *applicability-requirement* laid down in section 2.1.5..

And this is just another way of saying that the putative criterion number (8) cited in the table above ('identity of mind') is a spurious form of evidence. We do not have any identity-conditions for Selves (or Egos or Minds or Subjects) as such. We have no idea of how the notions of numerical identity and difference are supposed to be applied to such pseudo-entities. But if so, then, as Strawson has pointed out, 'we have no clear concept at all of such items'⁹¹. The notions of a Self, Mind, or Ego are only misguided attempts at reification (or 'objectivisation') of the unity of apperception. The identity of a given Mind or Ego is dependent on the particular *embodied* person whose Mind or Ego it is. It is a person's body and its spatio-temporal history that is the final *principium individuationis* for persons. If we try to sever the notion of a Self or Ego from the individuating background provided by a person's *body* and its spatio-temporal continuity, the notion of numerical identity becomes inapplicable to it. That is to say,

⁹⁰Cf. Chisholm, *Person and Object*, p.33, and Maddell, *The Identity of the Self*, pp.122-134.

⁹¹Strawson, 'Self, Mind, and Body', p. 174/5.

we cannot make sense of the difference between *one* such Self and *another* such Self any more. The first-person perspective, although it presupposes the applicability of the notion of numerical identity to the referent of 'I', cannot by itself provide the materials for its applicability. The Cartesian approach, which tries to represent the Self as essentially and exclusively identifiable as the thinker of a certain thought, cannot succeed since it cannot guarantee the *singularity* of the thinker thus singled out. That is to say, it cannot by itself establish the *contrast* between one such Cartesian Ego and a thousand Cartesian Egos thinking a certain thought⁹². And, in an analogous way, the Simple View's desperate attempt to introduce an essentially private criterion of personal identity in order to account for those imaginary cases in which the bodily criterion has become inapplicable, must be viewed to have failed. The alleged criterion of 'being I' (or, in its objectified version, 'being the same (particular) Self') cannot replace the *genuinely individuating* bodily criterion. It remains an empty wheel; nothing turns on it.

To sum up: By severing personal identity claims from their meaning-constitutive criteria, the Simple View at the same time *deprives them of their sense*, i.e. of their ability to pose a Yes/No-question to reality. If, as Chisholm and Swinburne maintain, personal identity could *obtain* in the *absence* of all of its customary evidence (inductive and criterial), and *not obtain* in the *presence* of all its evidence, then such claims would have lost their meaning. For then we would not simply find ourselves in a position in which (under certain unusual circumstances) we could not *know* (or find out) whether a given personal identity claim is true or not (e.g. which of the two fission products is identical with the original person). We would not even know any more *what to count* as its truth or falsehood (even if we could know). And, as I argued above, if we could not even know any more *what to count* as its truth or falsehood, we would not really understand what it means.

5.7. Parfit's Reductionist View

Parfit's Reductionist View consists in a combination of two closely related theses. The *first* thesis is, as he puts it,

'that the fact of a person's identity over time just consists in the holding of

⁹²Cf. Strawson's argument 'Self, Mind, and Body', p. 174; cf. also Anscombe, 'The First Person', p. 58.

certain more particular facts' (p.210)⁹³.

The *second* thesis is

'that these facts can be described without either presupposing the identity of this person, or explicitly claiming that the experiences in this person's life are had by this person, or even explicitly claiming that this person exists. These facts can be described in an *impersonal* way' (p.210).

The following, necessarily brief, remarks are intended to throw some doubts on both of these claims and to contrast Parfit's views on personal identity with the account suggested in the present treatise.

Parfit's reasons for propounding the *first* thesis have to be understood on the background of his opposition to the view (which he calls the 'Simple View') that the obtaining of personal identity consists in a 'further fact' over and above the obtaining of what he calls '*relation R*', i.e. 'psychological connectedness and/or continuity with the right kind of cause' (cf. p.262f.). Personal *identity* can, according to Parfit, be said to consist in the *unique* holding of relation R. But the uniqueness of the obtaining of this relation is not, as Parfit puts it, 'what matters in survival' (cf. p.263). It is important to note in this context that, as opposed to other adherents of an account of personal identity in terms of psychological continuity, Parfit does not take the holding of relation R to be a *criterion* of identity (at least not in the sense in which the notion has been invoked in this essay). For Parfit does not think that any criterion of personal identity can meet two requirements: namely (1) that the obtaining of personal identity between two identity candidates P_1 at t_1 and P_2 at t_2 depends only on 'intrinsic features' of the relations between them, and (2) that the obtaining of identity does not depend on 'trivial facts'⁹⁴. Parfit's 'reductionist' claim is rather that even in the cases in which we *would* assert the obtaining of personal identity the holding of relation R is all that personal identity can be said to 'consist in'.

Unfortunately, Parfit never makes quite clear what this reduction of one kind of facts (namely putative 'identity-facts') to another kind of 'more particular' facts (call them 'R-facts') is supposed to amount to. For he never indicates whether the reduction of 'identity-facts' to 'R-facts' is to be understood as a *meaning-reduction*, or whether it is to be understood in some more substantial, genuinely *metaphysical*, sense. If the

⁹³All references in the text of this section are to Parfit, *Reasons and Persons*.

⁹⁴Cf. Parfit, *Reasons and Persons*, sect.91, pp. 266-273.

reduction of personal identity to the holding of relation R is supposed to be a meaning-reduction in terms of necessary and sufficient conditions, then Parfit's claim that personal identity 'consists in' the holding of relation R seems to be clearly false. For, as we have seen above, psychological continuity is *neither a necessary nor a sufficient* condition of the obtaining of personal identity. If, however, Parfit's reductionism is to be understood in a different, genuinely metaphysical, sense, then Parfit owes us an explanation of what this sense is. *Prima facie*, a fact is simply whatever is the case if a given proposition is true (or false). And propositions are individuated in terms of the *meaning* of given sentences. If so, the claim that the obtaining of one fact p 'consists in' the obtaining of certain other facts q and r *either* amounts to saying that 'p' really *means* the same as 'q and r' -- in which case nothing is really 'reduced' (but the meaning of a statement is *explained*). *Or* it means that whenever 'p' is *true*, then so is 'q and r' -- in which case we might just have expressed the obtaining of an inductive relation, according to which, whenever a person P_1 is identical with a person P_2 , then P_1 and P_2 are also R-related. And this is not necessarily true. (For, as we have seen, the holding of relation R does *not imply* the obtaining of personal identity, nor vice versa.) Thus, it seems to me, as long as we do not know what his claim that personal identity consists in 'the holding of certain more particular facts' amounts to, we do not really understand Parfit's 'Reductionist View'.

I shall not belabour the problem here. For one, problems of reductionism are notoriously difficult, and this is not the place to venture into a long-winded discussion of these issues. More importantly, however, it seems to me that Parfit's 'reductionist' theses concerning personal identity should not so much be understood as making *ontological* claims about what is, or is not, the *case* in matters of personal identity, but as claims about *rationality*, i.e. as claims concerning what it would be *rational* for a person to do and wish for if he found himself faced with 'unprecedented' circumstances. Thus Parfit's imaginary cases are often less concerned with the question whether in a given case identity obtains, but with how it would be rational for a person to act, and what it would be *rational* for a person to wish for, if he found himself in the imaginary situation. To be sure, Parfit often tries to support his 'revisionary' theses concerning rationality by making claims about the ontology of personal identity. However, this seems to me a misleading feature of his account.

For the question whether in a given imaginary case personal identity *obtains* has to be distinguished from the question whether *one's own* continued existence is '*just as*

good' (cf. p.272) as one's own death combined with the continued existence of a psychologically continuous (and thus *qualitatively* identical), but *numerically* distinct, person. Whether the future existence of a qualitatively identical person would be 'just as good as' my, RS's, continued existence is a question I have not addressed so far, and shall not address here. (I am inclined to think that it would not be as good for *me*). The claim I tried to disprove in section 5.4.2. is simply that psychological continuity by itself can be said to be a sufficient condition of the obtaining of personal identity. Parfit typically does not distinguish between these two questions. Thus he often characterizes the future existence of a person who is imagined to be psychologically continuous with oneself as *one's own survival*. But it is not at all clear that the notions of a given particular's *survival* and its *numerical identity through time* can be prized apart so easily. After all, as Wiggins has pointed out⁹⁵, what one wishes for in survival is not that *somebody* (*x*) with all one's qualities be around, but that *oneself* with all one's qualities be around (or at least with the more desirable ones). It is a further question whether the special concern that human beings typically display with regard to their own continued existence is, in any sense, *rationaly founded* - a question which should be sharply distinguished from the question whether in a given case personal identity obtains.

Perhaps Parfit's insistence on the insignificance of strict numerical identity of person can be understood in terms of the second aspect of his reductionism about persons, namely his view that reality can be described in an '*impersonal way*'. This view is developed in sections 81 and 88 of *Reasons and Persons*. By claiming that the facts that personal identity consists in can be described in an impersonal way, Parfit does not only intend to deny any (Nagelian) doubts about our ability to give an all-encompassing *objective* description of reality⁹⁶. That is to say, Parfit's claim is not only that there are no 'essentially subjective' facts such as 'I am NN', or 'It is now 1992'. It is much stronger than that. For Parfit wants to hold that, in order to explain what personal identity consists in, we would not even have to refer to persons as a given kind of *particulars*. His claim is that a person's 'unity of consciousness' at a given time, and its numerical identity through time, can be explained *without essential reference* to this person, namely simply by referring to a given person's mental states and by specifying certain rela-

⁹⁵Wiggins, 'The Concern to Survive', esp. p.418.

⁹⁶Cf. his discussion of Nagel, in *Reasons and Persons*, p.252.

tions between them (namely 'co-consciousness' in the case of synchronic, and 'relation R' in the case of diachronic identity). We could, as he puts it,

'refer to and describe different thoughts, and describe the relations between them, without ascribing these thoughts to thinkers' (p.226).

In saying this Parfit does *not* want to deny that persons may figure essentially in the *content* of our thoughts, i.e. that persons *must* be mentioned in describing the contents of our thoughts, desires, etc.. But he *does* think that we do not necessarily have to *ascribe* thoughts to persons⁹⁷.

Now sometimes this Parfitian claim seems to mean something relatively weak (and arguably true), namely that we could describe what personal identity 'consists in' without making essential to a (Maddellian or Chisholmian) Self or Cartesian Ego. Sometimes, however, it seems to mean something much stronger, namely that we can conceive of experiences as *not necessarily owned by a person*; that, in other words, the fact that we ascribe thoughts to thinkers is simply due to the 'way we talk'⁹⁸; that persons are 'grammatical fictions'. This stronger claim is brought out by Parfit's likening of mental states to events, and by his comparison of persons to clubs or nations (cf. e.g. pp.213 ff.). Just as people are neither ontologically nor conceptually dependent on nations, so experiences -- Parfit seems to suggest -- are neither ontologically nor conceptually dependent on people.

Thus this stronger version of Parfit's second reductionist thesis that reality can be described in an 'impersonal way' seems to amount to the claim that 'mental states' could in principle lead an independent existence, i.e. that they are not necessarily states of a *particular* or *substance* (such as a person or a human being). Again, however, I am not quite sure how to understand this claim. Of course, nothing in 'the nature of things' forces us to talk about reality in a certain way. Nothing, that is, prevents us from abandoning our practice of ascribing mental states to the persons who have them. But it is not at all clear what deep metaphysical truth would be served by doing so. Perhaps, Parfit's claim that reality could be described in an 'impersonal way' is best understood as an over-reaction to the Simple View's misconceived interpretation of self-reference

⁹⁷ Cf. Parfit, *Reasons and Persons*, p.226. Parfit sometimes also makes this point by stating that persons are not 'separately existing entities' (e.g. p.251) apart from their brains and bodies (such as Cartesian Egos would be).

⁹⁸Cf. various passages in *Reasons and Persons*, esp. pp. 223, 225, 226, 251, 341.

and personal identity -- an over-reaction which of course is not without historical precedence. For just as *Hume* went too far in fending off Reid's and Butler's (and Descartes') views of the self by claiming that there was no such thing at all and that

'[t]he true idea of the human mind, is to consider it as a system of different perceptions or different existences, which are link'd together by the relation of cause and effect, and mutually produce, destroy, influence, and modify each other'⁹⁹,

Parfit has gone too far in fending off the Simple View's mistaken account of personal identity. And once again, I think, it is a misconception concerning self-reference that lies at the bottom of these confusions. Hume certainly did express an insight in his ironic remark that 'whenever I enter most intimately into what I call *myself* [...] I can never catch myself... without a perception'¹⁰⁰. Above I tried to restate, and clarify, it by saying that, from the first-person perspective, self-reference cannot be said to involve an *identification* (nor a *tracing* through time) of a given *particular*. But this, of course, is not to say that there are no such particulars as persons at all, and that 'mental events' should be considered as leading a life of their own. *I* do not have to single out a given particular in order to predicate of *myself* that I am in a given mental state. But this does not imply that *my* mental states are not owned by anything at all, -- that they could in principle lead an ontologically independent existence. As Kant correctly realized, I cannot conceive of myself as just a collection (or 'bundle') of 'experiences'. The first-person perspective itself, we saw above, presupposes the *particularity of the referent of 'I'* -- although it cannot by itself fill it in. And, that is to say, it presupposes the applicability of the distinction between numerical and merely qualitative identity to this entity. It presupposes, to put it bluntly, that *I am a particular*, -- not a type, or a set of interrelated experiences¹⁰¹. Mental states are, as Shoemaker has put it, essentially 'adjectival'¹⁰². Or, as Strawson pointed out long ago against the 'no-ownership view'

'particular states of consciousness ... cannot be identifyingly referred to except as the states or experiences *of* some entity. [They] ... *owe* their identity as particulars to

⁹⁹Hume, *A Treatise of Human Nature*, Book I, Part IV, p.266.

¹⁰⁰Op. cit., p.252.

¹⁰¹Cassam has recently made this point by showing that 'the content of a particular "I"-thought cannot be characterized without ascribing it, or drawing upon the possibility of ascribing it, to a particular person or subject'; cf, his forthcoming 'Reductionism and the First Person'.

¹⁰²Shoemaker, 'Critical notice: Parfit's *Reasons and Persons*', p.447.

the identity of the person whose states or experiences they are.’¹⁰³.

I shall not here attempt to repeat Strawson’s criticism of the no-ownership view of persons, or try to sketch the inconsistencies it necessarily runs into. As things stand in our conceptual scheme, we do not only ascribe thoughts to ‘thinkers’, but we ascribe them to the very same things to which we also ascribe physical attributes, namely persons. Ultimately, this practice is founded in the fact that we ascribe mental properties to persons on the grounds of their manifesting, or expressing, these properties in their *bodily* behaviour. (I shall have more to say about this topic in the next section). Parfit recognizes the possibility of such a ‘middle position’ between the Simple and the Reductionist View (cf. p.225), but postpones discussion of it. This is unfortunate. For I think a consideration of this position might have made him doubt some of the more ‘revisi-onary’ conclusions he draws. Both the Simple and the Reductionist View of personal identity, each in its own way, imbalance the ‘scales of grammar’. The way to restore their equilibrium is here (as always) not to heap more weights on either side, but simply to remove them all.

5.8. *Persons and Their Bodies*

In the previous sections I argued for the significance of a person’s body for personal identity mainly *via negativa*, namely by drawing attention to the shortcomings of the psychological continuity criterion as a necessary and sufficient condition of personal identity. However, the logical priority of the bodily criterion can also be demonstrated in a number of other ways which have been ignored so far. One such way is by reflection on the manner in which mental states and traits are ascribed to human beings and other kinds of animals. In what follows I shall try to show that such a reflection can uncover certain *logical* limits on our ability to accept certain kinds of ‘person-ality’ (understood in the wide sense introduced above, i.e. as a set of vaguely defined psychological attributes) as embodied in certain kinds of bodies. This reflection will also shed some light on a further issue which has not been addressed so far, namely the question what is to count as a *person’s body*. For, although we have seen that a person’s numerical identity through time can only be said to consist in the persistence of its *living body*, we have not yet said anything about what such a body must be *like*. That is to say, we have not yet said anything about whether a person’s body has to be necessarily

¹⁰³Cf Strawson, *Individuals*, p.97.

human, or at least animal, or whether it may be of a non-human, and perhaps even non-animal, shape and constitution. As we shall see, this question is in turn closely related to the question whether persons are necessarily human beings (or at least necessarily animals), or whether there could be non-human (and perhaps even non-animal) persons. In exploring the first issue, namely the question of what is to count as a person's body, I also hope to give an at least initial answer to the second question. In doing so, I shall try to elucidate some of the logical connections between our concepts of a person, a human being, and of an animal. It will appear that, since these notions are by no means sharply defined, clear-cut answers are here not always possible. However, I think it can be shown that for something to count as a person's body, it would first have to be an animal body, and hence for something to count as a person, it would first have to be an animal.

5.8.1. *Logical Limits of Embodiment*

The observation that there are certain *logical limits* on our readiness to count a given 'personality' as embodied in a given type of body was first made, although not elaborated, by Williams in the context of a discussion of Locke's imaginary case of the body-change between the emperor and the peasant. There Williams remarked that such a body-change may be inconceivable since

'the emperor's body might include the sort of face that just could not express the peasant's morose suspiciousness, the peasant's a face no expression of which could be taken for one of fastidious arrogance. These could's are not just empirical -- such expressions on these features might be unthinkable.'¹⁰⁴

It seems evident that this remark can constitute an argument against the possibility of a putative body-switch *only* if we also accept a further assumption, namely the assumption that a being that *in principle* cannot *express*, or *manifest*, a certain mental state or trait, cannot possibly *have* it. However, I think there is a good case to be made for such an 'expressive' account of the mind, which may also have a certain explanatory power as to the logical limits of one's ability to *imagine oneself* embodied in a body different from one's own.

It was of course Wittgenstein who held the view that statements to the effect that a being *manifests* a certain kind of behaviour expressive of a given mental property are at least partly meaning-constitutive of statements to the effect that this being *has* the

¹⁰⁴Cf. Williams 'Personal Identity and Individuation', p. 12:

given mental property. And this implies that a being's ability to (at least in principle) manifest (or express) a certain mental state or character-trait is at least partly constitutive of its having it¹⁰⁵. For Wittgenstein, this was a necessary consequence of the way mental properties are ascribed. We apply psychological predicates to beings on the basis of their behaviour, and the behaviour (including, in the case of human beings, *linguistic* behaviour) that beings display, manifests (or expresses) their 'mentality' (understood as a set of, vaguely defined, psychological attributes). The relationship between a being's behavioural capacities and the mentality expressed by its behaviour is, Wittgenstein maintained, *criterial*, not inductive. It is thence definitive of what psychological predicates it makes *sense* to attribute to certain kinds of beings, and thus what psychological attributes it is *logically* possible for a being to have. Thus, one might say, it is logically possible for a baby to wish for his or her mother to come to the cradle, since she can express this wish by (e.g.) crying. The baby's crying is a logically good reason for us to attribute this wish to the baby; it is a kind of behaviour that could *count* as the expression of such a wish. But a baby cannot (logically) wish for her mother to invest money in a life-insurance. For nothing in the baby's behaviour could justify our ascription of this wish to her. Nothing in the baby's behaviour could possibly count as the expression of such a wish, since it is a wish available only to beings capable of *linguistic* behaviour of a rather sophisticated kind. That is to say, it would not just be false, it would not really make sense to ascribe this wish (i.e. the wish for her mother to invest in a life-insurance) to the baby. For if it were false, then a possibility would be excluded by our saying that the baby does *not* wish for her mother to invest in a life-insurance. And in that case we would have to be able to specify this possibility, i.e. we would have to be able to specify what in the baby's behaviour could count as her *not* wishing for her mother to invest in a life-insurance. And that we cannot do.

Similarly, a *dog* can (logically) believe -- perhaps mistakenly -- that the cat went up the tree, because it can *manifest* this belief by (e.g.) barking at the tree. The dog's barking at the tree *justifies* our ascription of this belief to it; and the dog's finally turning away growling justifies our saying that it has changed its mind about the present location of the cat. But nothing in the behaviour of a *gold-fish* could possibly justify our saying that it does, or does not, believe that the cat went up the tree. Nothing in its

¹⁰⁵For a survey, cf. Hacker, *Wittgenstein. Meaning and Mind*, pp. 192-200, and pp. 224-253, and Malcolm, *Nothing is Hidden*, pp. 133-154.

behaviour could possibly *count* as the expression of this belief. And thence it would not just be false, it would not make sense to say that the gold-fish either does, or does not, believe that the cat went up the tree.

I think that recognition of the criterial, non-inductive, character of the relationship between the psychological attributes and their manifestations can account for, and draw our attention to, certain logical constraints on our readiness to count a certain 'personality' (or 'person-type') as embodied in a certain (kind of) body. At the same time, however, I do not think that these constraints are as narrow as Williams' remark might be taken to suggest. For Williams' remark seems to suggest that a given person's personality could only ever be 'realized' or embodied in a body which is very similar to that person's own body. And this, I think, would mean drawing the bounds of sense too 'near' to reality.

Citing Wittgenstein's dictum that 'the human body is the best picture of the human soul'¹⁰⁶, Williams claims that no expression in the peasant's face could be taken as one of 'fastidious arrogance'. And this may well be true. But perhaps, one might object, there are other behavioural ways in which the king's personality could 'shine through' the peasant's body. For there are other ways of manifesting fastidious arrogance than by displaying it in one's facial expression. There is, one might point out, no reason to assume that a given mental attribute can be manifested in only one specific way.

Let us consider a different case once envisaged by Quinton¹⁰⁷: One might admit that there is a certain, non-empirical, difficulty in imagining the personality of the 85-year old Churchill embodied in the body of a five year old girl. What, one might ask, in the behaviour of a five year old girl could possibly count as a typically Churchillian gesture? How could she give expression to Churchill's portentous views in her thin girlish voice? How would she manifest the fits of deepest depression that Churchill was allegedly prone to? These are difficulties. However, I think they are not insuperable. For, it seems to me, most of them could be overcome by the simple fact that the five year old girl can *speak*. In other words, where those difficulties can be overcome at all, they can be overcome by the fact that the girl, by displaying the appropriate *linguistic beha-*

¹⁰⁶Cf. Williams, 'Personal Identity and Individuation', p.12; Wittgenstein's dictum is to be found in *Philosophical Investigations*, II, iv, p.178.

¹⁰⁷Cf. Quinton, 'The Soul', p.60.

viour, can give expression to certain traits of Churchill's personality which she could not possibly express by displaying more narrowly defined *bodily* behaviour. This is significant; and, far from constituting an objection to the 'expressive' account of the mind sketched above, I think it can be properly explained only in terms of such an account.

Wittgenstein has taught us that linguistic behaviour is grafted upon, replaces, and extends, natural or instinctive behaviour¹⁰⁸. And I think it is the fact that linguistic behaviour can *replace* (narrowly defined) bodily behaviour which makes it possible for us to imagine certain personalities embodied in *prima facie* very unlikely kinds of bodies, and thus pushes the bounds of sense a bit further outwards in this context. This can be brought out by comparing the case of persons (or human beings) with the case of other kinds of animals which are not endowed with the gift of speech.

We could not imagine a chimpanzee's 'mentality' (understood as a set of, vaguely defined, psychological attributes characteristic of it) embodied in a rhinoceros. For nothing in the rhino's behaviour could count as giving expression to the chimp's mentality. But we *can* imagine Churchill's personality embodied in a five year old girl, because even a five year old girl, if she can speak, can give expression to typically Churchillian views, demand a cigar, brood with depression, etc.. What makes it possible for us to imagine Churchill's personality embodied in a five year old girl, is of course partly the fact that both Churchill and the girl are *human*, which provides them with a common behavioural background. It is the fact that they share a common *biological nature* (i.e. common interests, needs, desires, purposes, perceptual and intellectual capacities, etc.) and perhaps, more broadly, a common '*form of life*' (which may include culturally determined behavioural constancies). The fact that both Churchill and the girl share their common human nature is, one might say, the background against which the ascription of linguistic capacities first 'gets a hold'. However, it seems to me, we need not be overly restrictive about the specifically *human* nature in this context. As long as we imagine a given *animal* body capable of speech, it would not even have to be human in order to qualify as a candidate for 'housing' Churchill's personality. A bull-terrier (for example) may not be able to manifest Churchill's depression by his drooping head, or his irritation by the glint in his eye. However, assuming the bull-terrier was gifted with rational speech, it could, it seems to me, make good for certain lacks in its bodily

¹⁰⁸Cf. Malcolm, *Nothing is Hidden*, ch. 8: 'Language As Expressive Behaviour', pp.133-154.

expressive repertoire by displaying appropriate linguistic behaviour. Of course this purely linguistic behaviour would have to be backed up by a *basic range of bodily expressive behaviour*. A body which was to qualify for housing Churchill's personality would have to be of a kind that is capable of manifesting, and hence having consciousness (in general) and sensations, intentions, emotions, etc. (in particular). And this is to say it would have to be able to move about voluntarily, give bodily (not just linguistic) expression to various sorts of sensations, stretch out for something it wants, etc.. This implies that, at a minimum, such a body would have to have equivalents of human members, eyes, and mouth. (Connoisseurs of science fiction literature will easily recognize that this is typically the way in which extraterrestrial creatures are 'personalized'). The bodies of higher animals clearly fulfil these requirements. Of course, the lower down we descend on the evolutionary ladder, the more difficult it becomes to accept a being's linguistic behaviour as expressive of a certain personality. The lower we descend, the more our imagination loses its grip. However, it seems that at least in the case of the bodies of higher primates, the required bodily background constituted by bodily expressive behaviour is provided by our shared animal (or mammal) nature.

Furthermore, the fact that linguistic behaviour can replace narrowly defined bodily behaviour may also be part of the reason that it is so easy to *imagine oneself* inhabiting (i.e. embodied in) a body which is different from one's own, or even leading a completely disembodied existence. *Prima facie*, these at times astonishing feats of the imagination seem difficult to explain. For if, as was suggested above, it is *logically* impossible for a being to have a certain mental state or trait if it could not at least in principle *express* or manifest it, it seems difficult to see how one could imagine *oneself* embodied in the body of such a being. For, it seems, one would then have to imagine a logical impossibility. And this, arguably, one cannot do. However, once again it is the fact that linguistic behaviour can replace nearly all forms of bodily behaviour which pushes the bounds of sense a little bit further outwards here, and makes it possible for us to imagine oneself embodied in a body different from one's own -- and not only in a different body of (roughly) the same kind, but even in a completely *different kind* of body. Thus, it seems to me, one can easily imagine oneself embodied in a chimpanzee's or a dog's, and perhaps even an insect's, body -- although I would not recommend this as a regular exercise. In such cases of *subjective* imagination (cf. sect. 4.4.), one does not even have to postulate that the body in which one imagines *oneself* embodied is in any way capable of *publicly* expressing or manifesting one's personality (as one would

have to if one tried (objectively) to imagine Churchill's personality embodied in such a body). That is to say, one need not assume that the chimp-body or insect-body in which one imagines *oneself* embodied is in any way capable of public speech. And I think the reason why one can dispense with this requirement in the case of subjective (as opposed to objective) imagination is that what one imagines in such a case is finding oneself *with all one's linguistic abilities* embodied in such a body. That is to say, one imagines oneself as a chimpanzee who can at least *silently* (i.e. *in foro interno*) speak to himself. One imagines oneself as a chimpanzee who can have thoughts, beliefs, sensations, and character-traits which can only be had by a being endowed with certain linguistic capacities (although the chimpanzee in whose body one imagines oneself may not be able to express these thoughts, beliefs, etc. in any way). Otherwise, i.e. if one did *not* imagine the chimpanzee as silently reasoning with himself, one would not imagine *oneself* jumping about in a chimpanzee's body, but simply a chimpanzee jumping about. (And the fact that, in imagining oneself embodied in a body different from one's own, one cannot abstract from one's own linguistic capacities, may also be the reason why all the familiar stories (e.g. Ovid's) of the metamorphosis of what used to be a human being into some sort of animal -- if they want to lay claim to any intelligibility at all -- have to describe this animal as harbouring at least some remnants of the person's former 'inner life' -- although the beings into which they have turned may not any more be able to express it¹⁰⁹.)

5.8.2. *Persons, Animals, and Their Bodies*

The previous reflections on what is to count as a person's body may also help us in assessing the thesis that the concept of a person is parasitic on the concept of a human being in such a way that the only kinds of beings we could possibly accept as persons are in fact members of the species *homo sapiens*¹¹⁰. As I understand it, the

¹⁰⁹Cf. Wollheim's discussion of Ovid's *Metamorphoses*, in *The Thread of Life*, pp.4-10, (esp. the reference to Io to whom *verba animo desunt*).

¹¹⁰Cf. the recently voiced doubts whether the concept of a person is to count as a genuine sortal concept at all, Williams, *Ethics and the Limits of Philosophy*, p.114, and Morton, 'Why There is no Concept of a Person'.

Similarly, Snowdon (cf. his 'Persons, Animals, and Ourselves') has recently suggested that, in trying to find out what *one's own* numerical identity over time consists in, one should focus not so much on the dubious notion of *personal* identity, but rather on the more familiar, and more perspicuous, *natural kind concept* of a human

claim that *only* human beings could qualify for personhood (i.e. that persons are *necessarily* human), has a double thrust, depending on what kinds of beings are supposed to be excluded from personhood. It can be understood as claiming that only humans (*as opposed to other kinds of animals*) could possibly be persons, or it can be understood as claiming that only humans (*as opposed to artificially created beings*) could be persons. Let us first discuss the claim that *only human beings* as opposed to other kinds of animals could possibly qualify for personhood.

1. Now it is certainly true that the only kind of person any of us has ever actually encountered in experience was a member of the species *homo sapiens*. It is equally true that human beings are our paradigms for the notion of a person. However, the truth of neither of these two observations can by itself be taken to imply the stronger claim that the *possibility* of non-human animal persons is *conceptually* excluded. (Another way of understanding the claim that persons are *necessarily* human beings would be to take it as stating an *a posteriori* or Kripkean necessity according to which only beings equipped with the specifically *human* central nervous system could possibly be capable of language and hence self-conscious thought. I cannot see what good reasons could be adduced in favour of such a view, and I shall therefore ignore it in what follows.) Thus the way in which I should here like to understand the claim that only human beings

being (i.e. the animal species *homo sapiens*). For, since anything that is a human being is *essentially* a human being, we should look at the persistence-conditions specified by the sortal concept 'human being' if we want to find out what personal identity consists in. Equally, since *I* am *essentially* a human being, *my* numerical identity over time can be taken to consist in whatever human being identity over time will be found to consist in. (Cf. also Johnston, 'Human Beings' pp.75-80).

As should be evident from sect. 5.4., I agree with this 'animalist' or ('human being-ist') approach in that it makes no sense to say (as Locke famously did, cf. *Essay Concerning Human Understanding* II, xxvii, 20/1) that a person (a 'self', as Locke puts it, i.e. a self-conscious being) could survive a given real or imaginary procedure *as the same person*, but not *as the same human being* ('the same man'). The fact that -- apart from being members of the species *homo sapiens* -- we are also persons (i.e. beings with certain, vaguely defined, capacities for self-consciousness and rationality) does not endow us with '*special*' criteria of identity different from those specified by the notion of a human being. *However*, I do not think that the 'animalist' (or 'human being-ist') approach to questions of personal identity can by itself solve, or rather dissolve, the problems posed by those questions. And I think the reason for this shortcoming is that it perceives the questions posed by personal identity as problems concerning a special, ontologically problematic, kind of entities -- rather than as problems arising from certain systematic *delusions* about their own identity that members of this kind (i.e. persons or human beings) are prone to.

could be persons (and, I think, the only way in which it can be understood) is to take it as stating that only a human *body* could possibly count as 'housing' a person, i.e. that only a human body could *display* the kinds of behaviour that would justify the ascription of personhood (requiring mastery of a language, self-consciousness, rationality, etc.) to the being who 'inhabits' this body.

One possible attempt at substantiating this claim would be to point out that, as we argued above, higher psychological capacities (such as mastery of a language, self-consciousness, rationality, etc.) can be ascribed to a being only given an already established background of lower mental capacities, such as perception, sensation, emotion, volition, intention, belief, etc.. That is to say, it would not make sense to ascribe rationality and self-consciousness to creatures which do not first fulfil the more basic requirements for 'having a mind'¹¹¹. Citing Wittgenstein's famous remark that

'Only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious.'¹¹²,

one may be tempted to conclude that only human beings could fulfil these requirements. However, I think that this resort to an otherwise most respectable maxim would here be misguided. Wittgenstein's remark excludes Cartesian selves, parts of bodies (such as the brain), and 'mere' bodies themselves as possible candidates for consciousness (and thus, *a fortiori*, for rationality and self-consciousness). However, it by no means excludes other kinds of animals. Wittgenstein does not spell out the *relevant respects* in which a being has to resemble a human being in order to qualify for the cited psychological attributes. Nevertheless, I think we all have a vague idea what these relevant respects are. In order for it to make sense to say of a being that it has sensations, is conscious or unconscious, etc., the being has to be able to express these mental attributes in some way; it has to have equivalents of human eyes and ears: it has to be able to move about; it has to be able to avoid obstacles when awake; it has to be able to react to sounds when conscious, etc.. In short, it has to have just the kinds of capacities that distinguish the higher from the lower animals. Since higher animals resemble human beings in the relevant respects, there is no reason to deny that they are 'endowed with minds and mental states'.

¹¹¹Cf. Hacker, *Wittgenstein. Meaning and Mind*, pp.164-171.

¹¹²Wittgenstein, *Philosophical Investigations*, §281.

Of course, even the higher animals do not resemble human beings in the further respect which makes human beings capable of self-consciousness and rationality, namely the capacity for displaying *linguistic* behaviour, the mastery of a language. However, as I already argued above, these further requirements that a being has to fulfil in order to qualify for personhood are not so narrowly defined that only a *human* body could possibly fulfill them. There is no good reason why we should deny personhood to some super-parrot or super-chimpanzee which, by whatever fluke of nature, has acquired human language, and is capable of conversing intelligently on various literary, moral, political, and philosophical topics, displays emotions in his speech and his actions, remembers fondly (or not so fondly) the days of his youth, displays preferences and aversions, etc.. Of course a super-parrot, as opposed to (e.g.) a super-chimpanzee, could not express certain emotions by his bodily behaviour as humans can. (We can easily understand the 'body-language' of chimps or even dogs. We know when they are sulking and when they are merry. But what would it be for a parrot to display 'sulking-behaviour'?). Once again, however, the fact that linguistic behaviour can replace naturally expressive behaviour can fill the gap here. Assuming it was endowed with a Proustian versatility of language, our super-parrot could make good for the shortcomings in its naturally expressive behavioural repertoire by replacing bodily with linguistic behaviour.

I do not think that Wittgenstein's opaque remark that 'if a lion could talk, we could not understand him'¹¹³ could constitute a serious counterargument to this analysis. Presumably, what Wittgenstein had in mind in this passage was that linguistic behaviour cannot replace bodily behaviour, as it were, *in vacuo*. It can replace bodily behaviour *only* given a background of certain shared interests, wants, desires, etc. which we and the lion have in common. Wittgenstein seems to have thought that such a common ground cannot be established in the case of a lion. But this claim seems wrong, even by Wittgenstein's own standards. In the case of a lion, the common background of shared interests, wants, desires, etc., *is* provided by our shared *animal* nature. Both the lion and we have feelings of hunger, have sexual desires, raise our young, are familiar with the pleasures and vicissitudes of having a metabolism, spend a lot of time sleeping, take pleasure in lazing in the sun, etc.. In a very basic sense, lions and we *do* share a common biological nature; we do share a common 'form of life'. Thence I see no reason to deny personhood to a super-lion which has acquired a language, and pro-

¹¹³Wittgenstein, *Philosophical Investigations* II, p.223.

ceeding from that, rationality.

The conclusion which I take to be suggested by these admittedly unsystematic remarks is that we should view the concept of a person as the concept of an animal with certain specifiable capacities -- the paradigm for these capacities being the capacities characteristic of human beings. This view converges, although from a different direction, with Wiggins's so-called 'animal attribute view' of persons, and it may prove useful in this context to point out the similarities and the dissimilarities between Wiggins' account and the account suggested here. According to Wiggins' view the concept of a person is a *functional restriction* of the natural kind concept *animal*. In Wiggins' own words:

'[The concept *person* is] a concept whose defining mark can be given in terms of a natural kind determinable, say *animal*, plus what may be called a functional or (as I shall prefer to say) systemic component. Perhaps x is a person if and only if x is an animal falling under the extension of a kind whose typical members perceive, feel, remember, imagine, desire, make projects, move themselves at will, speak, carry out projects, acquire a character as they age, are happy or miserable, are susceptible to concern for members of their own or like species, etc... [and here Wiggins adds that the dots signify that this list of qualities characteristic of human beings is in principle open-ended], conceive of themselves as perceiving, feeling, remembering, imagining, desiring... [and here follows another in principle open-ended list]'¹¹⁴.

Wiggins further explains his 'animal attribute view' by stating:

"According to this view, a person is any animal that is such by its kind as to have the biological capacities to enjoy fully the psychological attributes enumerated; and whether or not a given animal qualifies is left to be a strictly empirical matter."¹¹⁵

And he concludes from this that the concept of a person is not a natural kind concept. Rather, he claims:

'Every person would belong to some natural kind that determined a sound Leibnizian principle of identity through change for some one kind of person (human-person, dolphin-person, parrot-person or whatever). There would be no one real essence of person as such; but every person could still have the real essence of a certain kind of animal. Indirectly, this would be the real essence in *virtue of which* he was a person.'¹¹⁶

Although I agree with Wiggins that our concept of a person is not a natural kind concept, and that the psychological attributes necessary to personhood are drawn from the *human* paradigm and are not clearly definable, I am more doubtful about the explanatory value of Wiggins' notion of a real essence and its ability to determine what he

¹¹⁴Wiggins, *Sameness and Substance*, p.171.

¹¹⁵Wiggins, *Sameness and Substance*, p.172.

¹¹⁶*Ibid.*

calls 'a sound Leibnizian principle'. I suggested above that we should follow Strawson in his more moderate view that our natural kind concepts do not really express more than 'cluster[s] of readinesses for variation *between* individual members of the kind and of expectations of typical continuities and typical modifications *in* individual members of the kind'¹¹⁷. Wiggins seems to think that even our pre-scientific natural kind concepts are determined by the unknown 'real essences' of the natural kinds specified by those concepts. As I tried to show above (sect.2.1.4.), it is not quite clear what role this notion is supposed to play in Wiggins' account. However, Wiggins' account of natural kind concepts in terms of real essences seems particularly inappropriate with respect to the case of a rational super-parrot which is endowed with an abnormally (for a parrot) large brain. What would be this animal's 'real essence in virtue of which' it is a person? Could we even still count it as a member of the species *parrot*? Of course, nobody would want to deny that if a given non-human animal were to develop the capacities which qualify it for personhood, it would have developed these capacities *in virtue of* certain physiological idiosyncracies (such as enlarged brain-size) untypical of its species. But, it seems to me, the very fact that they are *untypical* of the species should give us a good reason not to make them part of what fixes the concept of this species.

2. Perhaps the difference between Wiggins' account and the account suggested here (a difference which is by no means substantial) can be brought out more clearly by considering the second, and arguably weaker, version of the claim that persons are necessarily human beings, namely that persons are necessarily *animals* (as opposed to *artificially created beings*). This claim is sometimes put by saying that a *machine* could not possibly be a person. However, as has recently been pointed out by Hanfling¹¹⁸, this is an unhappy way of putting it, since machines are *by definition* neither conscious nor unconscious, let alone persons. The only intelligible way to pose the question is to ask whether artificially created beings, beings who are not made of flesh and blood as humans and mammals are, could possibly qualify for personhood. I think Hanfling shows convincingly that -- as long as we imagine these artefacts to look sufficiently like us (they have to have a *face*) and behave sufficiently like us in their various expressions of pain, emotion, desire, belief, intentions, character-traits, etc., and as long as we

¹¹⁷Strawson, 'Entity and Identity', p.210f (cf. sect. 2.1.1., p.13).

¹¹⁸Cf. Hanfling, 'Machines as Persons?', pp.25-27.

furthermore imagine them to be acquainted with what might be called the *basso continuo* of life (such feelings of hunger, sexual desire, etc.) -- we would, as he puts it, 'find it impossible not to treat these beings as persons'¹¹⁹. We would have to count them as persons in spite of the fact that they were not made of flesh and blood as we are.

This result might be taken as a demonstration that there *can* be persons who are *not* animals. However, to put it this way would be misleading. The OED defines 'animal' as

'a living being: a member of the higher of the two series of organized beings of which the typical forms are endowed with life, sensation, and voluntary motion, but of which the lower forms are hardly distinguishable from the lowest vegetable forms by any more certain marks than their evident relation to other animal forms'.

Hanfling's artificial person would clearly fall under this definition. For the very same facts that would justify our counting these beings as persons would *at the same time* also justify our counting them as animals. If, as we saw above, it makes sense to say of a certain being that it can think or is capable of linguistic behaviour, rationality, and self-consciousness, then -- due to what might be called the 'holism of the mental' -- it must *a fortiori* make sense to say that it has certain wants, desires, drives, and other attributes characteristic of animals. For anything to be a person, it must first be an animal. The respective grammars of 'animal' and 'person' are synchronized. In order for a being to qualify for personhood, it must first qualify for being an animal¹²⁰.

And in this sense the claim that all persons are animals could be said to be *conceptually* true. Of course, Hanfling's artificial persons are not animals in a strictly biological sense. It is not clear in which sense they can be said to have a metabolism, or whether they procreate. If not, certain rather crucial aspects of animal (and human) life would be unintelligible to them. However, I think one should not put too much emphasis on these (strictly biological) dissimilarities, as Wiggins seems to do. According to Wiggins, Hanfling's artificial beings could not be called persons since they do not fall under any *natural* kind concept specifying a given animal's real essence. However, this

¹¹⁹Op.cit., p. 31.

¹²⁰Which is the reason why a desktop computer, however intelligently programmed, cannot (logically) think. The ascription of thought and self-consciousness to a certain entity makes sense only on the background of a fairly rich and highly differentiated behavioural repertoire; cf. for a discussion of these issues, Hacker, *Meaning and Mind*, pp.160-170.

is where our paths diverge. What, it seems to me, we should say of these artificial beings is that they are a *hitherto unknown* form of animals; that they are animals with bodies of a very unusual chemical composition and a very unusual origin. This claim is not as outrageous as it may seem at first thought. Animals have all sorts of different ways of coming into existence, and they are made of all sorts of different kinds of stuff. There does not seem to be any a priori reason for assuming that only bodies constituted of certain kinds of ('organic') materials could justify the ascription of psychological predicates. To put it bluntly, there does not seem to be a good reason to be chauvinistic about carbon and water.

3. The conclusions suggested by the above discussion confirm a result we already arrived at in the previous sections, namely that persons do not differ from animals in that persons, but not animals, could change bodies. The identity of persons is not governed by a *special* necessary and sufficient criterion, namely psychological continuity, which is different from the criterion governing the identities of all other kinds of particulars, including animals. In truth, there is really only one necessary and sufficient criterion of numerical identity over time for all particulars, and that is spatio-temporal continuity. (Though of course how this governing criterion is applied is determined by what *kind* of particular it is applied to, the spatio-temporal continuity of a piece of rock being something very different from the spatio-temporal continuity of a living organism).

Persons, we have seen, are a functionally definable subclass of animals (or living beings), and their identities are governed by exactly the same criterion as those of other kinds of animals, namely the spatio-temporal continuity of their bodies. What -- amongst other things -- sets persons apart from other kinds of animals is (a certain kind of) self-consciousness, i.e. their ability to refer to themselves as-subject, and that is, without application of any criteria of their identity¹²¹. This ability -- an ability which may be said to be a necessary feature of personhood -- gives rise to certain *systematic illusions* that persons have about their own identities. But their proneness to these illusions does not make persons *substantially* different from the rest of the animal kingdom. What makes persons different from the rest of the animal kingdom, what sets them apart, is

¹²¹And, arguably, the ability to imagine having a different body from one's own is intimately connected with this ability. It is, in a way, just the other side of the same coin.

in a way just their proneness to being deluded about what constitutes their numerical identity over time.

Put differently, the problem of personal identity is not so much a problem concerning a particularly mysterious kind of beings but a problem arising from the illusions that members of this particular kind are prone to once they start to reflect about their identity. Fido, the terrier, cannot ask himself the question 'Will it be *me* who...?' before being connected to the brain-state transfer device which will transfer his brain-state onto Labox's, the St. Bernard's, brain. And I cannot 'projectively identify' myself with Fido, i.e. I cannot ask the question for him. For if I did, I would not really imagine being Fido, but rather imagine being myself (with all my linguistic capacities) embodied in Fido's body. And this is why Fido's numerical identity over time does not seem as profound a problem to us as a person's identity over time. But the profundity of personal identity is, we have seen, a spurious one.

Thus in a sense, there is really (as Strawson once remarked) only one 'common criterion for the identity of all substantial things'¹²², and that is spatio-temporal continuity. Space, one might say, is the only true 'individuator' when it comes to the numerical identity over time of particulars. Equally, however, it is worth noting that if Strawson's *Individuals*-claim that persons are a kind of *basic* particulars was supposed to imply that they are governed by a specific criterion of identity different from those of other kinds of particulars (as it is sometimes understood¹²³, although I am inclined to doubt the correctness of this interpretation), then this claim would have been wrong. And it would then not be without a certain 'irony of truth' that Strawson's *Individuals* definition of the concept of a person as

'the concept of a type of entity such that *both* predicates ascribing states of consciousness *and* predicates ascribing corporeal characteristics...are equally applicable to an individual entity of that type'¹²⁴,

is really the definition of an animal¹²⁵. If anything, *animals*, not persons (which are only a special, functionally defined, sub-set of animals) should count as a second class of 'basic particulars' next to material bodies.

¹²²Strawson, 'Entity and Identity', p.209.

¹²³Cf. Lowe, *Kinds of Being*, p.129, note 40.

¹²⁴ Strawson, *Individuals*, p.104.

¹²⁵This was first pointed out by Williams, in 'Are Persons Bodies?', p. 66.

It is a further question whether this implies -- as Williams has suggested¹²⁶ -- that persons are (identical with) their bodies. I shall not go into the matter at any length. *Prima facie*, the thesis that persons are bodies seems to me to have a mainly negative - and, that is, anti-Cartesian -- use. Persons certainly do not have bodies as they have cars; they cannot simply change them in for a new one. Equally, persons are not *constituted* of their bodies plus *something else* (such as a Cartesian *res cogitans*). But this is not to say that -- in a sense that would have to be explained first -- persons are 'reducible' to their bodies. It is persons (and other kinds of animals) that are conscious and unconscious, perceive, feel, hold beliefs, etc., not their bodies. Persons (and at least all the higher animals) *have* bodies, but they are not identical with them. When they die, they leave their bodies behind¹²⁷. (But again, this is not to say that persons are something substantially distinct from their bodies.) Equally, the conditions under which we would count a 'merely' material body (such as a piece of rock) as numerically identical over time are certainly very different from the conditions under which we would count a living organism (such as a dog or a human being) as numerically identical over time. And what constitutes this difference is just what constitutes the difference between a living and a dead animal body. Furthermore, the logical difference between the concept of a person (and of an animal) and the concept of a material body is also evident from the fact that, as we saw in sect.5.4., the concept of a person is, as it were, more '*brittle*' than the concept of a material body. We can imagine changes in the background-conditions of the concept of a person which render the concept of a person inapplicable to a given situation, but which still leave us with a hold on a person's body (although we cannot decide any more *whose* body it is). Thus, put in a nutshell, what seems to me the main value of the claim that persons are bodies is that persons are not something over and above (i.e. *substantially distinct* from) their living bodies; that -- given that bodily identity and psychological continuity go together -- the identity of a person does not consist in anything more than the numerically identical persistence of this person's body; and that the 'natural assumption' that *my* identity consists in something different from the identity of my body is based on a paralogistic illusion.

¹²⁶Cf. his 'Are Persons Bodies?'. *passim*.

¹²⁷Cf. for a survey of the 'grammar' governing a human being's (or an animal's) relation to its body, Hacker, *Wittgenstein. Meaning and Mind*, pp.244-253.

6. OUTLOOK

Let us sum up briefly what we have achieved: We started out by scrutinizing the epistemological idiosyncracies of a person's reference to himself-as-subject which, I tried to show, make us susceptible to certain systematic illusions with regard to the nature of persons and of personal identity over time. What lies at the heart of these illusions is, it became apparent, the mistaken view that a person's capacity to refer to himself *without* invoking any criteria of his identity can still be said to be a form of self-*identification*. Kant's distinction between a purely '*logical*' and an ontological, or *substantial*, understanding of the 'identity of the subject' was shown to be central to a dissolution of the illusion as well as to a solution of various philosophical conundrums concerning (particular kinds of) memory and imagination, and indexicality. Finally, in *Chapter 5*, we examined how various paralogistic illusions arising from a misconceived view of self-reference have influenced the interpretation of several thought-experiments concerning personal identity. The view of persons and their identity that emerged from this discussion differed sharply from both a 'Simple View' and a 'Reductionist View' of personal identity.

Thus, in a way, this essay has been one long meditation on Kant's 'Third Paralogism of Pure Reason' (although, of course, it was by no means an interpretation of Kant's actual text). Kant characterized the 'paralogistic illusion' as '*natural and inevitable*', i.e. as a

'a natural and unavoidable dialectic of pure reason -- not one in which a bungler might entangle himself through lack of knowledge, or one which some sophist has artificially invented to confuse thinking people, but one inseparable from human reason, and which, even after its deceptiveness has been exposed, will not cease to play tricks with reason and continually entrap it into momentary aberrations ever and again calling for correction'¹.

Judging from this and other passages, Kant seems to have thought that the 'paralogistic' illusion is not just based on various sorts of *linguistic* confusion (on 'misunderstandings of the grammar of our language'), but that it comes with being a *rational, self-conscious* creature as such; that it is, as he also puts it, a 'transcendental illusion'², which follows its own 'logic of illusion'³. One may have doubts about this

¹Kant, *Critique of Pure Reason*, B354f.

²Op.cit., A295 / B351.

³Cf.op.cit., A293 / B349.

assertion and about Kant's ability to uncover the hidden 'logic' of this 'unavoidable dialectic'. One may have doubts, that is, that the susceptibility to the paralogistic illusion is a necessary feature of all rational creatures, and not just a consequence of certain accidental structures of our language. Frankly, I do not know how to answer this question. But perhaps the mentioned opposition of language and 'human reason' does not harbour a real difference after all. Rationality and self-consciousness can only be attributed to creatures which have certain, fairly complex, *linguistic* capacities. And there may be good reasons for thinking that any language that we could recognize as such must display a certain very *fundamental structure*, which at least in the Indo-European language-group is captured by the distinction between *subject* and *predicate*. That is to say, one might argue that any language, which is supposed to be recognizable as such, has to be able to provide for the distinction between a referring and a predicating element (the 'basic combination'⁴). And one might then suggest that it is the application of this basic structure to ourselves (in 'first person psychological statements') which gives rise to the paralogistic illusion. This seems to me a promising prospect. However, I shall not pursue the matter any further in the present context.

In these final pages I should also like to counterbalance an impression which I may have given by exclusively focusing on the potential *illusory* qualities of a person's capacity to refer to himself-as-subject, namely the impression that there are no other philosophically significant aspects of this capacity. This, however, would be far from true. For I think that the capacity to refer to oneself-as-subject is what - in one sense of the word - makes us distinctively human. This can be brought out by examining its logical interrelations with other philosophically interesting human capacities.

One such capacity is the capacity for *detachment* that human beings (or persons) typically have⁵. The ability to refer to oneself without invoking any criteria of one's identity, also enables each of us to *detach* himself, if only in the imagination, from the particular identity of the kind of thing he is. In other words, via the ability to refer to himself-as-subject, a person can gain a 'distance' from his own criteria of identity. By detaching himself from the identity of the particular individual he is, each of us can gain a distance from this particular individual with its desires, wants, interests, its birth, and

⁴Cf. Strawson, *Subject and Predicate in Logic and Grammar*, ch.1.

⁵Cf. Nagel, *The View from Nowhere*, passim, but esp. chs. I and XI.

its death.

This capacity for self-detachment is, I think, an important fact about human beings. It is arguably a crucial precondition of the capacity for *moral* reasoning (as Kant certainly thought). For, in making it possible to gain a distance from oneself, this capacity to view oneself 'objectively' enables each of us to view himself as just one amongst others, whose aims and purposes have no objectively justifiable priority over others' aims and purposes.

The capacity for self-detachment may also be the primary intuitive source of various kinds of *monistic philosophies* (such as Spinoza's and Fichte's) and *religions* (such as Zen-Buddhism) which advocate the transcendence of individuality in some form of 'delimitation' of consciousness. According to these conceptions, the possibility to distance oneself from one's individuality (i.e. to lead one's life *sub specie aeterni*, to lead it in the light of the truth that 'All is One'), is an original aspect of personal life. It is planted in us as a potentiality that is essential to human life in particular, and perhaps to personhood in general.

Finally, the characteristically human capacity for self-detachment may also find an elucidatory application in an analysis of the phenomenology of *aesthetic experience*. Schopenhauer's reflections on this topic in *The World as Will and Idea*⁶, which characterize a person's contemplation of a work of art as a process of 'delimitation' which results in the person's temporary 'stepping out' of the chain of causality and recognizing himself as 'the pure knowing subject, the clear vision of the world'⁷, certainly deserve an 'austere' interpretation along these lines.

Last, but not least, it seems to me that the phenomenon of *humour*, arguably one of the greatest achievements of humanity, cannot be adequately understood without reference to the capacity to view the (too often absurd) events that happen to oneself purely 'objectively', as it were, from a distance - 'from no particular point of view', as

⁶Cf. Schopenhauer, *The World as Will and Representation*, Third Book, esp. §§ 30-38. Cf. e.g. the following passage: '[In aesthetic contemplation] the subject ... loses himself [in the contemplated object], i.e. forgets even his individuality, his will, and only continues to exist as the pure subject, the clear mirror of the object...; and therefore, he who is sunk in this perception is no longer individual, for in such perception the individual has lost himself; but he is *pure*, will-less, painless, timeless *subject of knowledge*.' (p.230).

⁷Op.cit., p.240.

Nagel would put it.

There are also further aspects of a person's identity over time which - it seems to me - deserve, and allow of, philosophical treatment, although they have not been touched upon in the present treatise. These are aspects which have typically been dealt with by 'existentialist' philosophers, although they should by no means remain their prerogative. One such aspect of personal life that these philosophers have latched onto is brought out by the intuitive observation that there is a sense in which persons are not just 'there' as physical objects and even animals are. In a way, a person actively *creates himself*. A person, in other words, cannot just view himself as an object to which things simply 'happen'. In leading my life, in making *choices*, I *create* the product called *my life*. Following a remark by Kierkegaard, Wollheim has to express this point by saying that life is 'understood backwards', but 'lived forwards'⁸. And he has tried to analyse the complex relations that obtain between a person's *future* 'life-projects' and his *past* existence as it reveals itself in self-examination. This seems to me fertile ground for future philosophical analysis, and a field where we are bound to encounter more genuine problems of 'personal identity' than the ones envisaged in the present essay - problems which are inextricably interwoven with the idea of personal freedom.

Some philosophers, such as Heidegger and Sartre, have suggested that, in order to account for the sense in which a person is not 'just there' but of his own creation, we must replace the *ontological* category of a person or human being with some more fundamental 'existential' category which takes into account this crucial difference between human existence and other forms of being (Heidegger's 'Dasein', and Sartre's 'être pour-soi'). However, this seems to me an over-reaction. In making choices, in creating the product called 'my life', I create the *particular* thing that I am. But the *kind* of thing I am is not of my own making. That I am a member of the particular ontological *kind* 'person' is rather a precondition of my being able to create the particular individual I choose to be. But I do not choose to be a member of this kind. I am what I am: a member of a given animal species whose typical members display certain capacities which set them apart from the rest of the animal kingdom. One of these capacities, I have tried to show, systematically misleads them about their true nature and their own identity over time. But it is also essential to their being what they are.

74873 words precisely

⁸Cf. Wollheim, *The Thread of Life*, passim. For the passage from Kierkegaard, cf. p.1.

BIBLIOGRAPHY

- Albritton, R.: 'On Wittgenstein's Use of the Term 'Criterion'', *Journal of Philosophy*, vol.56, 1959; pp. 845-857.
- Anscombe, G.E.M.: 'The First Person'. In *Mind and Language*, ed. S. Guttenplan. Oxford: Clarendon Press, 1975.
- Aquinas, Thomas: *Summa Theologiae*, Blackfriars, Eyre and Spottiswood, London.
- Ayers, M.: *Locke. Volume II: Ontology*, Routledge, London, 1991.
- Baker G.P. / Hacker P.M.S.: *Wittgenstein: Rules, Grammar, and Necessity Volume 2 of an Analytical Commentary on the Philosophical Investigations*, Blackwell, Oxford, 1985.
- Bennett, J.: *Kant's Analytic*, Cambridge: Cambridge UP, 1966.
- Burge, T.: 'Sinning against Frege', *Philosophical Review*, vol.88, 1977, pp. 398-432.
- Bishop Butler: 'Of Personal Identity', *The Works of Bishop Butler*, ed. by J. H. Bernard, London, 1900, vol.II.
- Cassam, Q.: 'Kant and Reductionism', in: *The Review of Metaphysics*, 1989, pp. 72-106.
- 'Reductionism and First-Person Thinking', forthcoming, in: Charles D. / Lennon, K.: *Reductionism, Realism, and Explanation*, OUP, Oxford, 1992.
- Dummett, M.A.E.: 'A Defence of McTaggart's Proof of the Unreality of Time' repr. *Truth and Other Enigmas*. Cambridge, Massachusetts: Harvard UP, 1978, pp.351-357.
- Castaneda, H.N.: 'He: A Study in the Logic of Self-Consciousness', *Ratio* Vol. 8 (1966).
- 'Indicators and Quasi-Indicators', *American Philosophical Quarterly*, 4, 1967; pp. 85-100.
- 'On the Logic of Attributions of Self-Knowledge to Others', *Journal of Philosophy*, vol. 65, 1968; pp. 439-456.
- 'On the Phenomeno-Logic of the I', in *Akten des XIV.Internationalen Kongresses für Philosophie*, vol 3, Vienna 1969, pp. 260-266.
- 'The Self and the I-Guises', in *Theorie der Subjektivität*, K. Cramer, & al. eds., Frankfurt: Suhrkamp, 1987.
- Chisholm, R.: *Person and Object. A Metaphysical Study*, La Salle, Illinois, 1966

- *The First Person. An Essay on Reference and Intentionality*. Sussex: Harvester, 1981.
- 'Identity Through Time', in H.E. Kiefer / M. K. Munitz, eds, *Language, Belief, and Metaphysics*. Albany, SUNY Press; pp. 163-182.
- 'Reply to Strawson', in the same volume, pp. 188/9.

- Dennett, D.: 'Where am I?' in Dennett, D.: *Brainstorms*, Montgomery, Vt.: Bradford Books, 1978.

- Diamond, C.: 'The Importance of Being Human', in: Cockburn, D. (ed.), *Human Beings*, Cambridge University Press, 1991, pp.35-63.

- Evans, G.: *The Varieties of Reference*, Oxford: Oxford UP, 1982.
- 'Understanding Demonstratives', repr. in his *Collected Papers*, Oxford: Oxford UP, 1985, pp. 290-318.

- Frankfurt, H.: 'Freedom of the Will and the Concept of a Person', *Journal of Philosophy* 68, 1971, pp.5-20.

- Frege, G.: 'The Thought a Logical Inquiry'. In *Philosophical Logic*, ed. P.F. Strawson, Oxford: Oxford UP, 1967.

- Geach, P.T.: 'On Beliefs about Oneself', repr. in *Logic Matters*, Oxford: Blackwell; 1972; 28-9.

- Geach, P.: *Truth, Love, and Immortality*, Hutchinson, London 1979.
- *Reference and Generality*, 3rd ed., Ithaca, Cornell University Press, 1980.

- Guyer, P.: *The Claims of Reason*, Cambridge, Cambridge University Press, 1987.

- Hacker, P.M.S.: *Insight and Illusion* Rev.ed., Oxford: Clarendon Press, 1986.
- *Appearance and Reality*, Blackwell, Oxford, 1987.
- *Wittgenstein: Meaning and Mind. Volume 3 of an Analytical Commentary on the Philosophical Investigations*, Blackwell, Oxford, 1990.

- Hanfling, O.: 'Machines as Persons?' in: Cockburn, D. (ed.), *Human Beings*, Cambridge University Press, 1991, pp.25-35.

- Henrich, D.: *Identitaet und Objektivitaet*, Carl Winter, Heidelberg, 1976.
- 'The Identity of the Subject in the Transcendental Deduction', in: Schaper, E./ Vossenkuhl, W. (eds.): *Reading Kant. New Perspectives on Transcendental Arguments and Critical Philosophy*, Blackwell, Oxford, 1989, pp. 250-280.
- 'Self-Consciousness', in: *Man and World*, vol.4, 1971, pp.3-28.

- Hume, D.: *A Treatise of Human Nature*, ed. L.A. Selby-Bigge, Oxford: Oxford UP, 1960..
- Husserl, E.: *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*, Tübingen: Niemeyer, 1980.
 -- *Logical Investigations*, Routledge, London 1970.
- Johnston, M.: 'Human Beings', *The Journal of Philosophy*, vol. LXXXIV, 1987, pp. 59-83.
- Kant, I.: *Critique of Pure Reason*, transl. by N. Kemp-Smith, London: Macmillan 1985.
 -- *Critique of Judgement*, transl. by J.C. Meredith, Oxford, OUP, 1928.
- Kenny, A.J.P.: 'Criterion' in: Edwards, P., *Encyclopedia of Philosophy*, vol.2, pp. 258-261.
 -- *Aquinas*, Oxford, Oxford University Press, 1980.
- Kripke, S.: *Naming and Necessity*, Blackwell, Oxford, 1980.
- Künne, W.: 'Sinn, Indexikalität, und propositionaler Gehalt' in *Grazer Philosophische Studien*, vol.18, 1982; pp. 41-74.
- Lewis, D.: 'Attitudes de Dicto and De Se', *Philosophical Review*, vol. 88, 1979, pp. 513-543.
- Locke, J.: *An Essay Concerning Human Understanding*, ed. A.D. Woozley, London: Fontana, 1964.
- Lowe, E.J.: *Kinds of Being. A Study of Individuation, Identity, and the Logic of Sortal Terms*, Blackwell, Oxford, 1989.
- Mackie, J.L.: 'The Transcendental 'I'', in *Philosophical Subjects: Essays Presented to P.F. Strawson*, ed. Z. van Straaten, Oxford, 1982.
 -- *Problems from Locke*, Blackwell, Oxford 1976.
- Madell, G.: *The Identity of the Self*, Edinburgh: Edinburgh UP, 1981.
 -- 'Personal Identity and the Idea of a Human Being', in: Cockburn, D.(ed.), *Human Beings*, Cambridge University Press, 1991, pp.127-143.
- Martin, C.B. / Deutscher M.: 'Remembering', in *Philosophical Review* 75, 1966, pp.170 ff.
- Malcolm, N.: 'Whether 'I' is a Referring Expression', in: C. Diamond / J. Teichmann, eds., *Intention and Intentionality. Essays in Honour of G.E.M. Anscombe*, Ithaca: Cornell UP, 1979, pp.15-24.
 -- *Nothing is Hidden*, Blackwell, Oxford, 1986.
 -- *Memory and Mind*, Cornell University Press, 1977
- Mellor, D.H.: *Real Time*, Cambridge: Cambridge UP, 1981.

- 'I and Now', in: *Arist. Soc. Proceedings*, 1988/89, 1st ed. 89, pp.79-94.
- McCabe, H.: 'The Immortality of the Soul', in: *Aquinas. A Collection of Critical Essays*, Macmillan, 1969, pp.297ff.
- McDowell, J.: 'Criteria, Defeasibility, and Knowledge', *Proceedings of the British Academy*, vol. LXVIII, 1982, pp. 477f..
- McGinn, C.: *The Character of Mind*, Oxford: Oxford UP, 1982.
- *The Subjective View. Secondary Qualities and Indexical Thoughts*, Oxford: Clarendon Press, 1983.
- McTaggart, J.: *The Nature of Existence*, Cambridge: Cambridge UP, 1927; vol.2.
- Moore, A.W.: 'Points of View', *The Philosophical Quarterly* vol.37, no.146, 1987, pp. 1-20.
- Morton, A.: 'Why there is no Concept of a Person', in: Gill, C. (ed.), *The Person and the Human Mind. Issues in Ancient and Modern Philosophy*, Clarendon Press: Oxford 1990, pp. 39-59.
- Nagel, T.: *The View from Nowhere*, Oxford, Oxford UP 1986.
- 'Subjective and Objective' repr. in *Mortal Questions*, Cambridge: Cambridge UP, 1979, pp. 196-215.
- Noonan, H.W.: *Personal Identity*, London, Routledge 1989.
- Nozick, R.: *Philosophical Explanations*, Cambridge, Mass.: Harvard UP, 1981.
- Parfit, D.: *Reasons and Persons* Oxford: Oxford UP, Pbk., 1986.
- 'Personal Identity', *Philosophical Review*, vol. 80, 1971, pp. 3-27.
- Peacocke, C.: *Sense and Content*, Oxford: Oxford UP, 1982.
- Pears, D.: *The False Prison*, Oxford, OUP 1988.
- Perry, J.: 'Frege on Demonstratives'. *Philosophical Review* 86, 1977, pp. 474-497.
- 'The Problem of the Essential Indexical' *Nous*, 13, 1979; pp. 3-21.
- 'Review of Williams's *Problems of the Self*', in: *Journal of Philosophy*, 73, 1976, pp.416-428.
- Prior, A.: 'On Spurious Egocentricity', in: *Papers on Time and Tense*, Clarendon Press Oxford 1969, pp.15-26.
- Powell, C.Th.: *Kant's Theory of Self-Consciousness*, Clarendon Press Oxford, 1990.

- Putnam, H.: 'Is Semantics Possible?', repr. in: Putnam, *Mind, Language, and Reality. Philosophical Papers*, vol.2, Cambridge UP, Cambridge, 1975.
- Quine, W.V.: reviewing Milton K. Munitz, (ed.) *Identity and Individuation*, *Journal of Philosophy*, vol. 69, 1972, p.490.
- Quinton, A.: 'The Soul', repr. in: Perry, J.(ed.): *Personal Identity*, Berkeley, University of California Press, 1975, pp.53-73.
- Reichenbach, H.: *Elements of Symbolic Logic*, New York, 1947.
- Reid, T.: *Essays on the Intellectual Powers of Man*, ed. by A.D. Woozley, London 1941.
- Rovane, C.A.: 'The Epistemology of First Person Reference' *Journal of Philosophy* vol. 84, 1987, pp. 147-167.
- 'Branching Self-Consciousness', *The Philosophical Review*, vol. XCIX, 1990, pp.355-395.
- Russell, B.: *Theory of Knowledge, the 1913 Manuscript*, Allen and Unwin, London, 1984.
- 'On the Experience of Time', *Monist*, vol.25, 1915, pp. 212-233.
- Ryle, G.: *The Concept of Mind*, Penguin Books, London, 1983.
- Sacks, O.: *Awakenings*, London: Picador, 1982.
- Schechtman, M.: 'Personhood and Personal Identity', *Journal of Philosophy* 1990, pp.71-92.
- Schopenhauer, A.: *The World as Will and Idea*, transl. from German by Haldane R.B. / Kemp J., vol I, Routledge Ltd., London, 10th impression, 1957.
- Shoemaker, S.: *Self-Knowledge and Self-Identity*, Ithaca: Cornell UP, 1963.
- 'Self-Reference and Self-Awareness', *Journal of Philosophy*, vol. 65, 1968, pp. 555-567.
- 'Persons and Their Past', repr. in: Shoemaker, S., *Identity, Cause, and Mind. Philosophical Essays*, Cambridge: Cambridge UP, 1982, pp. 19-49.
- 'On an Argument for Dualism', repr. in *Identity, Cause, and Mind*, pp. 287-309.
- 'Personal Identity: a Materialist's Account' in: S. Shoemaker /R.Swinburne (eds.), *Personal Identity*, Oxford: Blackwell, 1984.
- 'Introspection and the Self', in: *Midwest Studies in Philosophy*, vol. X, 1986, pp. 101-120.
- 'Wiggins on Identity', in: Munitz, M.K. (ed.): *Identity and*

- Individuation*, New York University Press, New York, 1971, pp.103-113.
- Smith, P.: 'Human Persons', in: Gill, C. (ed.), *The Person and the Human Mind. Issues in Ancient and Modern Philosophy*, Clarendon Press: Oxford 1990, pp. 60-81.
- Snowdon, P.: 'Persons, Animals, and Ourselves', in: Gill, C. (ed.), *The Person and the Human Mind. Issues in Ancient and Modern Philosophy*, Clarendon Press: Oxford 1990, pp. 83-107.
- 'Personal Identity and Brain Transplants', in: Cockburn, D.(ed.), *Human Beings*, Cambridge University Press, 1991, pp.109-127.
- Squires, R.: 'Memory Unchained', *The Philosophical Review*, 78, 1969, pp.178-196.
- Strawson, P.F.: *Individuals*. London: Methuen, 1959.
- *The Bounds of Sense*, London: Methuen, 1966.
- 'Singular Terms and Predication', repr. in: *Logico-Linguistic Papers*, London: Methuen, 1971; pp. 60-75.
- 'On Referring', reprinted in: *Logico Linguistic Papers*, pp. 1-27.
- 'Identifying Reference and Truth-Values', repr. in *Logico-Linguistic Papers*, pp. 75-95.
- 'Self, Mind, and Body', reprinted in: *Freedom and Resentment and Other Essays*, Methuen, London, 1974, pp. 169-178.
- 'Entity and Identity', in: H.D.Lewis (ed.), *Contemporary British Philosophy*, London: Allen and Unwin, pp. 193-219.
- 'Chisholm on Identity Through Time', in: *Language, Belief, and Metaphysics*, eds. Howard E. Kiefer and Milton K. Munitz, Albany, State University of New York Press 1970.
- 'Kant's Paralogisms: Self-Consciousness and the "Outside Observer"' in: Cramer K. et al. (eds.), *Theorie der Subjektivitaet*, Suhrkamp, Frankfurt, 1987, pp. 203-220.
- *Subject and Predicate in Logic and Grammar*, Methuen, London, 1974.
- Stroud, B.: 'Wittgenstein on Logical Necessity', *Philosophical Review*, LXXIV, 1965, pp.504-518.
- Swinburne, R.: 'Personal Identity: the Dualist Theory' in S. Shoemaker / R. Swinburne, eds., *Personal Identity*, Oxford: Blackwell, 1984.
- 'Personal Identity', *Proceedings of the Aristotelian Society* 74, 1973-74, pp. 231-247.
- 'Persons and Personal Identity', in: H.D.Lewis (ed.), *Contemporary British Philosophy*, London: Allen and Unwin, pp. 221-237.
- Tugendhat, E.: *Selbstbewußtsein und Selbstbestimmung*, Frankfurt: Suhrkamp, 1979.

Unger, P.: 'Consciousness and Self-Identity', in: *Midwest Studies in Philosophy*, vol. X, 1986, pp. 62-100.

Vendler, Z.: *The Matter of Minds*, Oxford: Oxford UP, 1984.

Vesey, G.: *Personal Identity*, Macmillan, London, 1974.

Walker, R.C.S.: *Kant*, London: Routledge and Kegan Paul, 1982.

Wiggins, D.: *Identity and Spatio-Temporal Continuity*, Oxford: Blackwell, 1967.

-- *Sameness and Substance*, Oxford: Blackwell, 1980.

-- 'The Individuation of Things and Places', *Proc. Arist. Soc.*, suppl. vol.37 (1963), pp. 122-146.

-- 'The Concern to Survive', in: *Midwest Studies in Philosophy*, vol 4, 1979, pp.417-422.

-- 'The Person as Object of Science, as Subject of Experience, and as Locus of Value', in: Peacocke, A. / Gillett, (eds.), *Persons and Personality*, Oxford, Blackwell, 1987.

Williams, B.: 'Personal Identity and Individuation', repr. in: Williams, B.: *Problems of the Self*, Cambridge: Cambridge UP, 1973; pp. 1-19.

-- 'Bodily Continuity and Personal Identity', repr. in *Problems of the Self*, pp. 19-26.

-- 'Imagination and the Self', repr. in: *Problems of the Self*, pp. 26-46.

-- 'The Self and the Future', repr. in *Problems of the Self*, pp. 46-64.

-- 'Are Persons Bodies?', repr. in *Problems of the Self*, pp. 64-82.

-- *Ethics and the Limits of Philosophy*, London, Fontana, 1985.

-- *Descartes: The Project of Pure Enquiry*, Penguin, Harmondsworth, 1978.

Wilkerson, T.E.: *Minds, Brains, and People*, Clarendon Press, Oxford, 1974.

Wilkes, K.: *Real People*, Oxford: Blackwell, 1988.

Wittgenstein, L.: *Tractatus Logico-Philosophicus*, Frankfurt: Suhrkamp, 1971.

-- *The Blue and Brown Books*, Oxford: Blackwell, 1958.

-- *Philosophical Investigations*, trans. G.E.M. Anscombe. Oxford: Blackwell, 1958.

-- Zettel, ed. G.E.M. Anscombe / G.H. von Wright, tr. G.E.M. Anscombe, Blackwell, Oxford, 1980.

-- *Remarks on the Foundations of Mathematics*, ed. G.E.M. Anscombe, G.H. von Wright, R. Rhees, tr. G.E.M. Anscombe, revised edition, Blackwell, Oxford, 1978.

-- *Remarks on the Philosophy of Psychology*, vol.II, ed. G.H. von Wright / H. Nyman, tr. C.G.Luckhardt / M.A.E. Aue, Blackwell, Oxford, 1980.

Wollheim, R.: *The Thread of Life*, Cambridge: Cambridge UP, 1984.

-- 'Memory, Experiential Memory, and Personal Identity' in: G.

McDonald, ed., *Perception and Identity*, London, 1970.