The Irreducible Self.

A Consciousness-Based Account.

By
Milosz Pawlowski

Submitted to
Central European University
Department of Philosophy

In partial fulfilment of the requirements for the degree of Doctor of Philosophy

Supervisors:
Howard Robinson
Marya Schechtman

Budapest, Hungary
2009

This work contains no materials accepted for any other degrees in any other institutions. This work contains no materials previously written and/or published by another person, except where appropriate acknowledgment is made in the form of bibliographical reference.

- Milosz Pawlowski

Abstract

This work addresses primarily the questions of personal identity over time: questions about conditions under which the same person can exist at different times. Yet any theory of personal identity must be able to provide also some plausible answer to the question “What are persons?” This question organizes my discussion of personal identity. I develop a balanced methodology of inquiry into personal identity. It places metaphysics at the centre, but takes phenomenology of self-experience as the methodological starting-point and explains how practical implications of theories of persons can affect their plausibility. In accordance with this method, every major view in the field - the Psychological Theory, the Physically Based Approach, the Transience View, the No Self Theory and the Simple View – is examined from two points of view: that of metaphysics and that of intuitions (deep beliefs underlying our practices). I argue that persons are real, irreducible, ultimate components of reality. Their identity is a primitive fact that cannot be fully analyzed. This is the Simple View of persons. Its Consciousness-Based version is the best theory of personal identity we have. According to this view, a person’s survival is intimately tied with the person’s being a potential subject of a continuous flow of consciousness. Arguably the best theory of the nature of persons consistent with this view is Cartesian Dualism.

Acknowledgements

I would like to thank Howard Robinson for all I have learnt from him throughout my years in Budapest; this work would not have taken the shape without this influence. I am also profoundly grateful to Marya Schechtman for agreeing to work with a visitor from a faraway land – and one with some outlandish ideas at that. Thanks to inspiring discussion we had, the period I spent in Chicago has been most productive. I extend my thanks to University of Illinois at Chicago and to everybody who made my visit there possible (and enjoyable). I have benefited from the support and discussions with all members of the Department of Philosophy at Central European University. I am particularly grateful, however, to Istvan Bodnar for kindly offering me assistantship in his project; only thanks to this opportunity could I finalize my work.

Table of Contents

List of Abbreviations

PBA	Physically Based Approach
PT	Psychological Theory
SV	Simple View
TW-identity	trans-world identity

Chapter 1. Introduction

1.1. The questions

This work addresses questions of personal identity. The first question can be introduced as follows. Let “A” refer to a person who exists at time t_1 and “B” refer to a person who exists at a later time t_2. Suppose “A = B” is true. Now, what makes it the case that there is just one person who exists at both times and is referred to by “A” and by “B”; and what would make it the case that there are two different persons? As philosophers say, we ask about necessary and sufficient conditions of personal identity over time - we ask how to complete the sentence “A = B if and only if...”. So far, we are asking specifically about *persistence* or *diachronic identity* of persons i.e. about conditions under which the same person can exist at different times.[1] But the question can be easily generalized. Instead of stipulating that “A” and “B” each refers to a person at a different time, we can assume that they identify (or indicate or pick out) a person in two different ways. Then we can ask again how to complete the sentence “A = B if and only if...”. The question “What are the necessary and sufficient conditions for identity of persons?” is the *general question of personal identity*.[2] This question can still be generalized. I implicitly assumed that our terms “A” and “B” refer to actual persons. But it is natural to extend the question to possible persons and ask under what condition the person called “A” (actual or possible) is identical to the person called “B”

[1] The persistence question could be distinguished from the diachronic identity question as its sub-question. It is natural to take the question “What are the conditions of the persistence of any particular person over time?” as equivalent to “What kind of changes can any particular person survive?”. Now, we can give a perfectly informative answer to this question without being able to similarly answer the diachronic identity question. For we can specify a range of changes that a person *cannot* survive; and then we can say “x survives if and only if such-and-such changes do not occur”. But saying that A did not undergo any person-destroying change gives us only the *necessary* condition for A’s identity with any B. It tells us that A exists at the time that B exists, but it may not tell us *which* of the persons that exist at that time is A.

[2] A useful clarification of the logic of identity questions can be found in Noonan (2003), 88-93. I subscribe to Noonan’s general solution: reduction of identity questions to questions about kind-membership and the range of possible histories for a thing of a kind. While this is useful for dispelling confusions, it would be too cumbersome to conduct the whole discussion in these terms. I follow Noonan in continuing to use traditional terms of the debate.

(actual or possible). Since this question concerns also identity between persons in different possible worlds, we can call it the question of *trans-world identity of persons.*

These questions are aligned with other fundamental metaphysical questions. The trans-world identity question seems on reflection equivalent to the *individuation* question. This is the very abstract question "What makes an object *this* particular object as opposed to others?" Now, if we have the answer to this question - if we have a grasp on the "principle of individuation" of some thing A - then we can answer the trans-world identity question. For then we can say that A = B if and only if what makes A the very object it is the same as what makes B the very object it is.[3] In the other direction, if we can fully answer the trans-world identity question for A (i.e. if for any possible B we can say under what conditions it is identical to A), this means that we know what uniquely identifies A in any possible world. But this is nothing else than to have the principle of individuation of A.

The second fundamental metaphysical question related to identity questions is the question "What is X"? In our context, we will ask "What is a person?" or "What am I?". I take this question to be about the *essence* of the thing; and I take "essence" to be a general name for a thing's *necessary* properties. Now, there are three particularly important types of necessary properties. We ask about them in the following sub-questions of the what-is-x question:

- What is the *ontological category* of X? Is X a substance, or property or event etc.?
- What is the *fundamental nature* of X? Is it physical, or biological, or psychological etc.?
- What is the *ontological structure* of X? Is it a simple thing or is it composed or constituted or constructed by or from other things?

[3] It may be possible to re-raise the trans-world identity question for the individuating factors. But this does not change the fact that they do give an answer to trans-world identity question about A and B.

The essence question is not simply equivalent to any identity question; but the relation is intimate. First, if we specify the necessary properties of the thing, we thereby automatically give necessary conditions of its identity; and vice versa: necessary conditions of identity automatically translate into necessary properties of a thing (though they may not be properties of any of the three important types I enumerated above). On the other hand, any informative specification of identity conditions will give some answers at least to the sub-question about the nature.[4] So much for logical relations. But there are no less important methodological connections. First, the essence question seems *explanatorily prior* to identity questions. It is because a thing is what it is, that it has this sort of identity conditions and not others. This should incline us to say that a theory giving only diachronic identity conditions of persons without suggesting what they are is unsatisfactory.[5] For it does not give us a metaphysical explanation of why persons have such identity conditions. Secondly, we may say, in the spirit of Plato, that before we discuss conditions of persistence of persons, we should first be able to say *what* it is that we are talking about. Hence I propose a methodological postulate:

(P) Any theory of personal identity must be able to provide some plausible answer to the question what are persons. If no provided answer is plausible, the theory is implausible.

The main topic of my work is the diachronic identity question about persons. But the "What am I?" question will be the guiding question structuring the discussion. Given that answers to the essence question have immediate consequences for the trans-world identity question, the latter will also figure prominently in my work.

[4] Specification of identity conditions is not limited to saying that some factors are necessary or sufficient. Saying that some factors are *irrelevant* also counts.

[5] Cf. Thomson (1997), 204f.

1.2. The aims

My aim is to defend the Simple View with regard to diachronic identity question. This position could better be called the Primitive Identity View. The central claims of this theory are:

(SV1) There are persons.
(SV2) It is impossible to provide a non-circular criterion of diachronic identity of persons i.e. to specify a relation R s.t. it would be true that for any persons A, B "A = B iff R(A, B)" and the holding of R of A and B could be described without presupposing that A and B are a single person.[6]

Persons exist, but their identity is a primitive fact that cannot be fully analyzed. This tenet sets the Simple View against Complex Views which attempt to provide an analysis of personal identity in terms of physical or psychological factors: persistence of the body or of psychological systems. This disagreement becomes intelligible when we turn to the question "What are persons?". On Complex Views, persons are wholly *composed* of familiar natural objects. They are "nothing over and above" these lower-level objects. This means that their persistence conditions can be analyzed in terms of those lower-level objects: a person persist if and only if a whole composed of such-and-such objects exists. But this also means that the existence of persons can be in some sense *reduced* to the existence of lower-level objects. Now, Simple View theorists maintain that persons cannot be so reduced and that they are among the ultimate components of reality. And identity of ultimate components of reality cannot be fully analyzed in terms of other objects. It is a basic, primitive fact.[7] Thus the Simple View is less mysterious than it is often made. It need not rest on the appeal to unique features of persons. It rests on the recognition that persons are real, irreducible, ultimate

[6] Noonan construes Simple View as denying the possibility of "any informative specification of constraints on the possible history of a person"; *op. cit.*, 97. He mentions the option of admitting the specification of necessary, but not sufficient conditions of identity. But my definition of the Simple View is weaker still. It admits specifying *some* necessary conditions and even *some* sufficient conditions, but denies the possibility of informatively specifying a condition which would be both necessary and sufficient. The position I will develop has these features.

[7] Cf. Lowe (2003), 91.

components of reality - as real and ultimate as, say, elementary particles are thought to be. The unanalyzability of their identity is the consequence of this status. Now, being an ultimate component of reality may naturally be thought to entail *mereological simplicity*.[8] Indeed, many Simple View theorists profess the view that persons are mereologically simple. But it is possible to conjoin Simple View, as we defined it, with the thesis that persons do have parts (so the name "Simple View" is potentially misleading). We need to carefully distinguish several ontological issues:

- Reality of persons: Realism vs Eliminativism
- Identity conditions: Primitivism vs Analyzability
- Ontological status: Anti-Reductionism vs Reductionism
- Mereological structure: Mereological Simplicity vs Mereological Complexity

Primitivism and Anti-Reductionism are mutually entailing. Analyzability entails Reductionism but not vice versa.[9] Mereological Simplicity entails Primitivism and Anti-Reductionism, but not vice versa. On this scheme, Simple View is Realist, Primitivist and Anti-Reductionist. Its main opponent, the Complex View, is Realist, Analyzabilist, Reductionist and therefore embraces Mereological Complexity.

While Simple View denies the possibility of *full* analysis of personal identity, it allows for placing *some* constraints on histories of persons. I will defend three such constraints:

(CS) If a person's consciousness flows on continuously, the person persists.

[8] The doctrine of mereological simplicity concerns only *substantial* parts i.e. parts which are substances. It does not concern "logical" or "spatial" parts. Cf. Lowe (1996), 36.

[9] Cf. Noonan (2003), 97. Reductionism may be purely ontological and not conceptual; hence Analyzability is not entailed.

(PCN) Persons essentially possess the potential for consciousness.

(PCCN) Persons essentially possess the potential for *continuous* consciousness.

(CS) says that a person cannot go out of existence while her consciousness continues. Assuming that consciousness cannot be shared with other persons, (CS) provides an informative *sufficient* condition of identity: if A and B are linked by continuous consciousness, then A = B.[10] Continuity of consciousness does not seem, however, to provide a necessary condition of personal identity. We think we can survive periods of unconsciousness. But it seems necessary that we retain at least the potential to be conscious - otherwise, for all intents and purposes, we are dead. This, I think simply follows from the concept of person. Hence (PCN) is a necessary condition of personal identity. The last necessary condition, (PCCN) is a shorthand for more complicated clause (C) which will be formulated and defended in chapter 7.

So much for diachronic identity. In accordance with my methodological postulate (P), we should now ask: "What, then, are persons according to the Simple View?". Our first three clauses say that persons are at least *loci* of potentially continuous consciousness. What else? With regard to ontological categories, I embrace a Substantialist version of the Simple View:

(SUB) Persons are substances.

(SUBJ) Persons are subjects of mental states ("thinking things").[11]

[10] Admission of consciousness-sharing brings complications, since at a later time there may be two different persons linked by continuous consciousness to A. Thus (CS) may only be a sufficient condition of the *persistence* of A.

[11] Regarding persons as substances contributes to removing the air of mystery from the Simple View. It is not mysterious that if a substance does not undergo certain changes, it continues to exist. To still ask "But why does it continue to exist?" is to misunderstand the metaphysics of substance. The concept of substance is the concept of something which can exist and continue on its own (a *continuant*); and it is this concept which Aristotle found to make all change intelligible. This may be questioned. But *within* metaphysics of substance the fact that some substances have necessary but not sufficient informative conditions of persistence is nothing strange.

When it comes to the nature and structure of persons, Simple View is surprisingly promiscuous. Every option can, with some effort, be married with it.[12] Since my main topic is diachronic identity, I do not propose to solve these questions. I will tentatively argue that the Cartesian Dualist version of the Simple View is the most plausible view overall. I define Cartesian Dualism in the following way:

(D) The *only* essential property of persons is being "thinking things".

(MS) Persons are mereologically simple.

(NC) Persons are not constituted by other things or properties of other things.

To sum up. My aim is to defend a version of the Simple View: (SV1) & (SV2) & (CS) & (PCN) & (PCCN) & (SUB) & (SUBJ). It can be called a Consciousness-Based Substantialist Simple View. I will also tentatively argue that Cartesian Dualism is the best account of what persons are: (D) & (MS) & (NC).

1.3. Consciousness-Based Approach. Self and consciousness.

The approach I take in this work can be called Consciousness-Based.[13] Proponents of the Consciousness-Based Approach hold that consciousness is central to the inquiry into personal identity in two ways. First, consciousness and unity of consciousness should figure prominently in the substantial account of identity and nature of persons. Secondly, phenomenology is the methodological starting point and a central part of the inquiry into personal identity. There are two areas which need to be phenomenologically investigated above all: experiences of the self and the unity of consciousness. Both issues need a comment.

[12] See ch. 10, 251.

[13] J. Foster, B. Dainton, G. Strawson and D. Zahavi are among recent proponents of this approach.

In this work, I treat ‘person’ and ‘self’ as synonyms. There are few slightly different nominal definitions of ‘person’. I favour Boethius’ definition:

(DB) person $=_{\text{DF}}$ individual substance of rational nature[14]

There is also Locke’s popular definition:

(DL) person $=_{\text{DF}}$ thinking intelligent being, that has reason and reflection and can consider itself as itself, the same thinking thing in different times and places[15]

The trouble with Locke’s definition is that it builds persistence in time, and arguably strong requirement of self-awareness, into the definition of ‘person’. Boethius’ definition, while conveniently vague and non-committal, builds substantiality into the definition. If we want to have a maximally general non-partisan nominal definition, we can say:

(DFP) person $=_{\text{DF}}$ individual something of rational nature

If there is any doubt what I mean by ‘person’, I refer the reader to (DFP).

Concepts can be stretched, narrowed, re-defined and abused. It often happens in contemporary discussions of personal identity that disputants seem to talk about entirely different things.[16] But it is a fact that we have self-experience. It also seems that the core of our concept “person” is derived from self-experience. Or, at any rate, we *want* to have a concept based on self-experience. We would not regard a concept which is wholly un-related to our self-experience as the concept of person that we want to have. And it is reasonable to

[14] Boethius, *De persona et duabus naturis*, 3: “Persona est naturae rationalis individua substantia”

[15] Locke, *An Essay Concerning Human Understanding*, II, 27, 9.

[16] Cf. Strawson (1999), 99-101.

suggest that we should want to discuss a self-experience-based concept before everything else: for all further conceptual constructions will depend on their intelligibility on this concept. Now, these thoughts can suggest the following reasoning. It is surely absurd to say "This thing satisfies the nominal definition of 'person', but it *is not* a person" or "This thing does not satisfy the nominal definition of 'person', but it *is* a person". Given that the core of our concept of the self is essentially derived from self-experience, we should say that it is likewise absurd to say "This thing is accurately represented in a self-experience, but it is not a self" and "This thing is unlike anything represented in self-experience but it is a self". In this spirit, Strawson makes two methodological postulates:

> (E1) If there is such a thing as the self, then some SMS [sense of the mental self] is an accurate representation of something that exists,
> (E2) If some SMS is an accurate representation of something that exists, then there is such a thing as the self.[17]

Phenomenology of the self is thus methodologically prior to and imposes substantial constraints on metaphysical theories. I am very sympathetic to Strawson's approach, but I prefer a more balanced methodology. There are many questions to be asked about Strawson's postulates: what is the range of self-experiences considered? can there be selves which can have self-experience but never attain it? can self-experience involve mistakes? how accurate does it have to be to count as successful self-experience? which parts of self-experience are central and essential? Phenomenological consideration would no doubt play a role in answering such questions, but almost certainly they would under-determine the answers.[18]

[17] Strawson (1997), 341.

[18] Strawson stipulates that he discusses a special form of self-experience; and a special sense of 'self'. This makes his postulates more plausible, but does not remove all questions. Furthermore, it is an open question how Strawson's sense relates to other senses: is it more basic, independent or parasitic upon other senses? Strawson's project cannot be fully extricated from the broader framework of personal identity inquiry.

Our answers would be largely dictated by decisions with regard to *concepts*. Phenomenology alone does not provide a ground for metaphysical theorizing. It has to be coupled with conceptual analyzis. Nonetheless, we need to start with describing self-experience to understand what lies at the core of our concepts; and it is fair to postulate that philosophical theory should not disregard experience without providing good reasons. How, then, is the self presented in self-experiences which provide the core of the concept "person"? What *kind of thing* is the self as presented in self-experience? Strawson provides a concise and compelling characterization. The self is ordinarily experienced as:

(1) a *subject of experience*, a conscious feeler and thinker

(2) a *thing*, in some interestingly robust sense

(3) a *mental* thing, in some sense

(4) a thing that is *single* at any given time, and during any unified or hiatus-free period of experience

(5) a *persisting* thing, a thing that continues to exist across hiatuses in experience

(6) an *agent*

(7) something that has a certain character or *personality*.[19]

Before encountering Strawson's work, I have drawn my own list. I find the two remarkably similar. I suggest that the self is figured in self-experience as something that:

(1) *has* conscious mental life (experiences, thoughts, desires etc.)

(2) *has* a body

(3) exists *here* and *now*

[19] Strawson (1999), 106.

(4) has a *past* and a *future*

(5) can be *self-aware*

(6) *thinks*

(7) can *speak*

(8) is *active*

(9) is *free*, i.e. can do one thing or some other thing

(10) is partly *elusive*; the ordinary self-experience is felt to leave something out.

Although there may be slight divergences between the lists drawn by different people, it seems that there is a reasonably well-delimited common core to human self-experiences. But it is an open question whether all the experienced features are essential to selves; and whether they are experienced or conceived as such. Nor is it obvious which and how many features can be missing from experience without compromising its status as self-experience. Such questions are not tractable by phenomenology alone. But if we take phenomenology to provide only an important point of reference for classical inquiry into personal identity, we do not need to address these questions. We do not need to force uniformization between slightly divergent lists. Rather, it will be one of the tasks of the comprehensive theories of persons to take a stance with regard to self-experience, to explain its diverse features and its relation to our concepts.

The decision to start with the common *generic* self-experience is nonetheless a weighty methodological decision. Self-experience may also include our circumstantial or individual features. And while there is the question “What *kind* of thing I am?” there is also the question “What *distinguishes* me from other persons?”. To a philosopher, this may sound like a *bona fide* individuation question. And this question is highly relevant to personal identity questions. So why should the question, and experiences purporting to provide the answer, be disregarded? The answer is that it is dubious whether the ordinary experiences and questions really are about individuation. What may really concern us are questions like "What makes me different and unique (rather than a clone or an anonymous member of mass-society)?" or "What particular sort of person I am now?" The features figuring in the answers will likely not be our *necessary* properties, nor will they be conceived as such. It would be dogmatic to brush off these questions as wholly irrelevant to personal identity inquiry; but it would be naive to think that their relevance is straightforward. Given these doubts, non-generic self-experiences cannot provide a secure starting-point for our inquiry.

I turn now to consciousness. By 'consciousness' I always mean *phenomenal* consciousness. To quote Tye, "A mental state may be said to be *P*-conscious [phenomenally conscious] if, and only if, it has a phenomenal character (or a subjective 'feel')"; "*P*-consciousness is integral to experiences and feelings generally [...] Wherever there is a feeling or experience, there must be *P*-consciousness, some phenomenology that the relevant state has."[20] To add another popular idiom, a state is phenomenally conscious when there is "something it is like to be in it".[21] Two brief comments. First, the scope. I agree with Dainton that consciousness is not limited to most familiar cases of feelings and perception, but encompasses also such things as fringe-experiences and understanding-experiences (and, I would add, volition-experiences).[22] Secondly, absoluteness. Phenomenal consciousness is usually thought to be all-or-nothing. And for good reasons. It is difficult to understand how there could be a middle ground between there being *some* experiences (no matter how "dim" or un-informative) and there being *no* experiences. This looks rather like trying to find a middle-ground between existence and non-existence. Of course, somebody who seriously claimed that phenomenal consciousness can be gradual would either have to question the

[20] Tye (2003), 7f. Tye provides a lucid discussion of different kinds of consciousness and its unity, *op. cit.*, 1-15.

[21] Cf. Dainton (2008), 29.

[22] Mangan (2001); Dainton (2000), 11-14; Dainton (2005), 7f; Dainton (2008), 43f.

intelligibility of the whole “what it is like” idea or would have to advance claims about the nature of the phenomenal which implied that there is a common scale on which the clearly phenomenal and the non-phenomenal are located. Now, the first option looks rather like an *elimination* of the phenomenal than a defense of its gradualness. The second option is not much better. If there is the nature of the phenomenal to be discovered beyond what it seems to be, then the phenomenal could be something quite different from what it seems to be. This is a bizarre idea indeed (notwithstanding its past popularity among Materialists). Moreover, we would be required to conceive that there can be a middle ground between something having a qualitative character and something not having it. This would be like conceiving a middle ground *in the quality dimension* between being red and not being coloured at all. I do not find this intelligible. For these reasons, I feel entitled to think that phenomenal consciousness is all-or-nothing. I am also inclined to suppose that those who seriously argue for gradualness of consciousness talk about different kinds of consciousness than the phenomenal; and about these they may be entirely right. Thus I excuse myself for not discussing the issue in more detail.[23]

There is the phenomenon that some experiences are *experienced together*. When this occurs, we say that consciousness is *unified* or that we have a case of *co-consciousness* between experiences. Co-consciousness is a *structural* relation. It has to do with the structure of consciousness and *not* with the inter-content relations.[24] It is not concerned with different contents meaningfully “fitting” together, or with cross-references between contents. Secondly, co-consciousness is a *phenomenal* feature of consciousness and as such it is in principle directly observable through introspection. Thirdly, co-consciousness is a *short-term* relation. It links temporally close contents. This relation is used to account e.g. for awareness

[23] I respond, however, to one argument against absoluteness advanced by Unger in ch. 2, 50ff.

[24] Cf. Strawson (1997), 358: “these constancies and steadinesses of development in the *contents* of one’s consciousness may seem like fundamental characteristics of the *operations* of one’s consciousness, although they are not”.

of movement. I use the term *continuity* for talking about long-term diachronic unity of consciousness or its stream-like character. *Unity of consciousness* is the most general term, covering all sorts of co-consciousness and structural continuity of consciousness. I take no stance with regard to the metaphysics of the unity of consciousness; whether it consists in the singleness of (complex) experience, or is founded upon primitive relations of co-consciousness between experiences, or is founded upon (or identical to) con-subjectivity of experiences. None of these conceptions is presupposed by my arguments.

I want to finish this section by mentioning two new books which came into my hands when most of my work has already been done. These are Barry Dainton's *The Phenomenal Self* and Galen Strawson's *Selves: An Essay in Revisionary Metaphysics*.[25] I have decided to add references to these works, but I do not discuss them. Still, I would like to offer some comments. Strawson's book provides more detailed arguments than his earlier articles, but the essence of Strawson's methodology and views remains largely constant. Thus I think that my discussion of his position is not rendered obsolete. My approach to personal identity turned out to be quite similar in many respects to Dainton's approach. The common idea is that the persistence of the self is to be explained in terms of persistence of the potential for continuous consciousness. Out of my central theses, only (SV2), characteristic of the Simple View, would not be accepted by Dainton. Still, there are important differences between my project and Dainton's. First, Dainton understands potential for consciousness in terms of powers to *produce* consciousness. I remain non-committal with regard to powers-metaphysics. I just hold that selves have the potential to be conscious. I doubt whether powers to produce consciousness provide an adequate analysis of this potential. Intuitively, the power, if we want to use this word, to *have* or to *undergo* experiences is something quite different from the power to *cause* experiences. The fact that something causes experiences

[25] I thank Galen Strawson for sending me an advanced draft of the book.

does not mean that it has them. This is a facet of a broader disagreement about the metaphysics of subjects. Dainton takes subjects to be *constituted* by powers of material systems such as the brain. His theory is a version of what I call Constitution Psychological Theory. I will argue extensively that this general conception of subjects is flawed in more than one way. These arguments apply to Dainton's theory as well.

1.4. Method and strategy of argument

Debates of the 90's have led to the displacement of thought-experiments from their central role in the philosophy of personal identity and to subsequent divergence of methodological paths. Some have focused on pure ontology, making puzzle-solving capacities the main test for the success of theories.[26] Others postulate that philosophy of the self should start from phenomenology of self-experience. Finally, some approach matters of personal identity predominantly through investigation of our everyday concerns, practices and ethics. These are the three main approaches nowadays. Ideally, a work on personal identity should deliver a package-deal: cover phenomenology, ontology and ethics and consistently combine different perspectives. In Chapter 2, I provide an overview of a comprehensive, balanced methodology of personal identity inquiry. But in practice, it is hard to treat everything at equal depth. Although I give a slight methodological priority to phenomenology, the centre stage in this work is taken by ontology. I try, however, to consider every theory from two perspectives: the perspective of pure ontology and the perspective of intuitions. My discussion of intuitions is inevitably selective and is always focused on the intuitions *motivating* a given theory and on its most *counter*-intuitive aspects.

My general strategy is to make a *cumulative* case for theoretical *preferability* (highest plausibility) of the Simple View. This strategy has two aspects. On the one hand, I develop four themes favourable to my view:

[26] Hudson (2007), 217.

(i) Intuitions about numerical identity of persons are sparse. Nothing in our shared experience, thinking and attitudes requires us to believe anything inconsistent with Simple View.

(ii) Seeing ourselves as subjects is central to our self-understanding.

(iii) Our grasp on personal identity consists in the grasp on potential continuity of consciousness.

(iv) Consciousness has features uncongenial to Reductionism.

These considerations support Consciousness-Based Simple View against Reductionist approaches in general; but they could be outweighed by other arguments. Thus the second aspect of my strategy is the comparative assessment of the plausibility of particular theories. Sometimes it is possible to argue that one theory *collapses* into another under pressure. In order to carry out the comprehensive comparison, I introduce my own classification of theories with the intention of covering the whole logical field. The ontological structure of persons is at the heart of my classification.

The structure of the work reflects my strategy. In Chapter 2, while laying methodological foundations, I comment on some popular misconceptions, making first forays in defense of SV in the area of intuitions and consciousness. Considerations of Psychological Theories in ch. 3 leads to selection of Parfitian Reductionism as the best rival of SV within this camp. The next chapter goes back to intuitions: I argue that the favoured intuitions of PT-theorists do not support their view and do not count against SV. Chapter 5 is devoted to the assessment of the Physically Based Approach (PBA); One Realizer theory will stand as the most plausible PBA view. Yet PBA in general cannot claim to be supported by intuitions, as I argue in ch. 6. The next odd chapter is, as usual, concerned with metaphysics; this time with metaphysics of consciousness. Themes (iii) and (iv) are developed. I then argue that Fission provides a general argument *against* Reductionism. Consideration of Fission provides a handy introduction to the topic of anticipation considered in ch. 8. There I go back to Parfitian Reductionism and argue for its collapse into Eliminativism. I suggest that similar arguments may work for any form of Reductionism. Then in ch. 9 I come to views denying existence to persisting persons: Transience and No-Self views. I argue that the latter view is a methodological non-starter. Galen Strawson's Transience View is less preferable than Cartesian Dualism both in the intuitive and ontological dimensions. Chapter 10 synthetically presents the support for Simple View. I briefly argue for the preferability of Cartesian Dualism and defend it against some common challenges. Overall conclusions are contained in ch. 11.

1.5. Limitations of this work

Lastly, I want to mention some limitations of my project. First, I consider only *mainstream* ontological approaches to personal identity. I do not discuss partial or vague or relative identity. These bring very special problems which are treated by other authors in more depth than I could reach here. Secondly, I work in the framework of the metaphysics of *things* (substances). This is justified on the grounds that self-experience present us to ourselves as things. At the same time I am quite liberal about the limits of substance-metaphysics. For instance, I do not take collapsing substance into another category as abandonment of substance-metaphysics.

The classification of theories focused on the ontological structure of persons has the consequence of pushing other differences between theories into the background. Thus when I discuss, for instance, Psychological Theories, I devote little attention to the question which particular psychological relations are used in the accounts of personal identity. The greatest loss this procedure generated is the lack of a systematic discussion of the Narrative

Approach to personal identity.[27] The Narrative Approach is hard to pin down. I tend to regard it as a peculiar version of the Psychological Theory.[28] As such, it faces the same ontological problems. It may, however, be able to deal better than the classical Psychological Theory with my arguments concerning intuitions.[29] It is possible that some Narrative theories involve acceptance of special ontology of persons or relative identity etc. In such cases, the disclaimer about the limited scope of my ontological investigation applies. Finally, there is an interesting option of taking Narrativism in non-metaphysical way. Narrativism could be said to investigate one sense of "person" and "personal identity" - perhaps predominant in our everyday experience, thinking and language - while leaving matters of ontological categorization and structure largely open. From the metaphysical perspective, Narrativists could be talking about *personae* or phase sortals, while classical theories of personal identity talk about the substance which is a person.[30] In such case, there need not be any conflict between the Simple View and Narrativism.[31]

[27] MacIntyre (1985), Taylor (1989) and Schechtman (1996) are among the classical Narrative accounts in analytical philosophy. The seminal work of Wollheim (1984) is an important source of the Narrativist ideas. Zahavi (2005), 99-132 usefully compares the Narrativist conception of the self with some other conceptions. Strawson (2005b) offers a critical discussion of Narrativism.

[28] Cf. Schechtman (2001), 239, 257. Slors (2001) provides a clear example of such theory.

[29] Dainton's Consciousness-Based version of PT is altogether untouched by these arguments, since they are made precisely from the consciousness-based perspective.

[30] Rist (2002), 61-94 and Zahavi (2005), 128-130 develop this kind of concordantist approach.

[31] My arguments against PT in ch. 4 are influenced by Narrativism taken in this way.

Chapter 2. Methodology

2.1. Plausibility of theories of personal identity. Strategies of argumentation.

What do personal identity theorists talk about? We find the following kinds of items:

1. The *experiences* of selves (including experiences of *value*)

2. The *words* ‘self’, ‘person’.

3. The *concepts* which these words express. This concept, besides having a content, has a *role* in the system of our concepts, in our thinking and our practices. This role is partly relative to our *interests*.

4. The *beliefs* we have about the selves. Some of which may be *deep,* but not true by the definition of ‘person’. Beliefs may be beliefs about values, in which case they express experiences of value or evaluative attitudes.

5. The *referents* of ‘self’ and ‘person’ (if they exist).

The philosopher has three basic tasks:

a) to faithfully describe the experience (phenomenology)

b) to analyze the concepts and to propose a definition, if possible (conceptual analyzis, semantics)

c) to say whether the referents exist and what they are: what is their nature, what are their conditions of identity (metaphysics)

Theories of personal identity belong to metaphysics. This is not to deny that they may be arrived at through, or supported by, phenomenology and conceptual analyzis.

The bulk of the work of the personal identity theorist consists in *a priori* argumentation (mostly the attempts to reduce rival theories to absurdity). Consistency of the theory and validity of the arguments are primary standards for evaluating the theory. But beyond the basic level of logical cogency there is the level of *plausibility* of philosophical theories.

A theory of personal identity is a special theory in the field of metaphysics. To large extent problems about personal identity are the same as problems about identity as such. The plausibility of the theory of personal identity is therefore directly dependent on the plausibility of its general metaphysical framework. It is not easy to say on what grounds theories in general metaphysics are found plausible or implausible. Beyond logical coherence, which is the dominant consideration, the reasons have to do with its intelligibility, agreement with our ordinary thinking and parsimony of the posited ontology.

Apart from the question about the plausibility of the general metaphysical framework of a theory of personal identity, we can ask how plausible it is as a theory about *persons*. This is the “personal” bit of “personal identity”. We need to have a closer look at the relation between theories of personal identity and our concepts, experiences, beliefs and attitudes.

2.1.1. Concepts

First, it is obvious that the theory cannot tell us that persons are objects which have totally different properties than that ascribed to it by our concept of person. For that would be just nonsense.[32] It is, however, possible to claim that while there is no thing which has *all* the relevant properties, there is something in the world which has almost all the relevant properties. It can be argued that such objects deserve to be called ‘persons’ and that our concept should be modified so as to allow for that.

[32] I take it that ‘person’ and ‘self’ are not natural-kind terms, but terms which apply to things only because these things satisfy certain descriptions. It is a further issue (not to be conflated with the one just mentioned) whether ‘person’ is a substance sortal or merely a phase sortal.

Although ‘person’ is not a natural-kind term, it is importantly related to one such term, namely ‘human being’. The relation is certainly not analytical.[33] But mentally healthy adult human beings provide the *paradigm* example of persons. A theory which would deny that things we refer to using the term ‘human beings’ are persons seems wildly implausible.[34] But why? The concepts have certain roles in our thinking and practices. And often we form and use concepts because of some special interests of ours. Now, suppose that a theorist would offer an analyzis of the concept “person” as well as a metaphysical account such that it follows from them that human beings would not be persons. And suppose we could not prove him wrong. What would we do? We should be happy to *give* him the concept “person”. Even if it turned out that in fact this was our concept, it was not the concept we *wanted* to have. For the concept we want to have should apply to human beings. And we would proceed to try and form such concept, “person*” and just forget the useless concept “person”. We would shift our interest to discussions of personal* identity accordingly. The fact that a theory denies that human beings are persons does not by itself show that it is *wrong*. It means that it is by and large a *non-starter*. It would not show anything important or interesting for us. But it could not be altogether ignored either. For it could be right about our concepts like “person” and “identity”, and if so, it would force us to make changes in our conceptual system. If such changes could not consistently and plausibly be made, then the theory would have to be reconsidered and taken seriously. It would, after all, show something very important.[35]

[33] There is no reference to human beings in the meaning of ‘person’. And it is not a part of the meaning of ‘human being’ that this being satisfy the conditions of personhood. There is a general agreement that there could be non-human persons, say, Martians. They could have the concept “person”. But they could lack the concept “human being”. On the other hand, many people deny that fetuses and babies are persons, while happily admitting that they are human beings, and it does not seem that they commit a logical mistake.

[34] We should be careful in judging whether a theory denies that human beings are persons. For example, a Platonist claiming that an immaterial soul is a person could be accused of doing so. It is possible, however, to hold that our ordinary beliefs about what human beings are mistaken (since ‘human being’ is a natural-kind term, this claim is not absurd). A Platonist could say that ‘man’ actually refers to an immaterial soul, which is a person. So the theory does not deny that human beings are persons.

[35] These remarks apply to the No-Self Theory. It could be claimed to be a non-starter, since there are human beings, and so there are obvious instances of “person” or at least “person*”. Some No-Self Theorists, namely early Unger, duly deny that human beings exist; Unger (1979*a*); Unger (1979*b*). Unger also claims that our

If a theory can be rejected in the way explained above, I will say that it can be rejected on *pragmatic grounds*. The same analysis applies, *mutatis mutandis*, to several other demands which are often made on theories of personal identity. First, at least “I” and “you” engaged in the communication of the theory - the reader and the writer - should come out as persons. It could be thought that a theory denying this would be straightforwardly self-defeating. Since I can refer to myself, and so be self-aware, and I can understand the theory and arguments, I obviously meet standard requirements for counting as a person. But if somebody claims that ‘I’ does not refer, or that ‘I do not exist’ can be said truly, then perhaps she can avoid this objection. In this case her theory would be disqualified only on the pragmatic grounds. Secondly, I should be able to know that I am a person. If a theory entails radical scepticism with regard to such claims, it cannot be taken seriously. Thirdly, our ordinary methods of re-identifying persons should prove to be reasonable and to some extent justified.[36] Fourthly, there are various postulates of preserving and/or justifying our *attitudes* and *moral practices*. The status of such demands is more complex than in the cases considered so far. We would be extremely unwilling to give up on the function performed by the concept “person” in our classificatory and epistemological interests. But our actual attitudes and moral practices are open to change. We can come to regard them as unjustified and abandon them. Or even if we cannot abandon them, we may come to regard the picture of the world which they presuppose as an unavoidable and perhaps practically necessary illusion. Hence the fact that a theory of personal identity does not square with our actual attitudes and practices cannot be used to disqualify it. Considerations of this nature can play an important role, however, in the absence of conclusive theoretical reasons for choosing one of the theories. If two or more theories differed in the matters of conceptual analyzis, we

concept “person” is incoherent, Unger (1979*c*). Giles comes close to this when he argues that *all* realist answers to Fission are implausible on logical grounds, Giles (1997), 68-73. I maintain, however, that to vindicate his position the No Self theorist should claim that a coherent concept “person*” *could not* be framed.

[36] More on this demand in section 2.2. below.

could legitimately choose the one which fits in better with our interests. Even if disagreement seemed to concern only substantial metaphysical claims, we could think that if we were forced to accept an uncongenial metaphysics, this may be because something is wrong with our concepts and we would have to modify them. So we can accept the more congenial metaphysics because it suits the concepts we want to have, regardless of whether these are in fact our actual concepts or modified concepts. Finally, if we were fairly certain that a metaphysical disagreement could not plausibly be solved by any such fiddling with concepts, we could openly make the choice on the grounds of practical preference for preserving our actual attitudes.[37] Since disagreement about conclusiveness of philosophical arguments is bound to arise, the considerations of coherence with our attitudes and morality will play an important role in supporting or criticizing any theory. We need to add one final refinement. What can plausibly be demanded of a theory is to conform to *rational* or *plausible* or at least *tenable* attitudes and morality. Since these are what we want to have, whether or not we actually do. Even if a theory does not tally well with our actual morality, it can fit with a morality we *should* have. If independent reasons for a revision of our attitudes and morality can be given, the theory avoids criticism and gains plausibility.[38]

How much weight one is willing to give to interest-related considerations largely depends on one's approach to metaphysics. If one thinks that metaphysics uncovers "real unities" and real metaphysical relations, one can think that such considerations have heuristic value but one may as well treat them as basically irrelevant. The more one thinks in terms of choosing consistent ways of "cutting up" the world (or, conversely, of the ways its components can be arranged into units) the more weight our interests and concerns gain for metaphysics.

[37] See Haksar for a sustained defense of legitimacy of appealing to practical reason in deciding which theory to accept in the absence of conclusive proofs Haksar (1991) 59-86. Strictly speaking, such reasons for choosing which theory to believe do not have anything to do with *plausibility*.

[38] Parfit's *Reasons and Persons* is a *locus classicus* for this strategy.

2.1.2. Experiences

Experiences of the self may reveal many more properties than that which are included in the concept "self". Any theory of personal identity should explain the occurrence of experiences of the self. If the theory is in accordance with our experiences i.e. if the explanation provided by the theory makes these experiences (in most cases) veridical, this provides a strong support for the theory. It is an acceptable move, however, to claim that these experiences are mis-representing reality and to explain how we fall into error. It is incumbent on the proponent of this kind of theory to provide strong arguments for regarding our experiences as mistaken. Unless we are given such arguments, we are entitled to assume that our experiences are veridical and hence to regard the theory which does not agree with them as implausible.[39]

2.1.3. Beliefs

Our beliefs about persons are only partly derived from our experience of our selves. First, we may have tendencies to *think* about our selves in the past and the future in certain ways; we also think about ourselves as existing during the periods of unconsciousness. Our beliefs about what happens at such times obviously cannot be based on direct experience. Secondly, our involvement with morality, religion, spiritualism, metaphysics and science makes us hold all sorts of beliefs about what persons are, what their position in reality is and what their value is. Now, some of the beliefs not directly based on experience could belong in the class of so-called *deep* beliefs. If such beliefs were apparently largely independent of any developed theory they could be called "intuitions" and treated almost on a par with experiences. But since it is always debatable to what extent our beliefs are affected by theories (and anyway, why theory-independence should strengthen the claim to truth?), it seems to me preferable to speak about deep beliefs. Now, what makes a belief deep? There are three lines along which one can develop the idea of the depth of beliefs. Deep beliefs

[39] Cf. ch. 1, 8-11.

would be such that:

- we attach very high subjective certainty to them and we would not want to give them up
- they are presupposed by our common everyday practices: linguistic, moral, judicial etc. as well as by our emotional and evaluative attitudes.
- they fit best our overall view of reality.

The first idea seems to capture best what is ordinarily meant when we say that someone deeply believes something. But it is not relevant for our purposes. People can feel certain about most idiosyncratic and crazy beliefs. Different people believe different things. Even if the majority of mankind was strongly convinced of the truth of some beliefs, this would provide no reason for demanding of a theory of personal identity (or any other theory) to conform to these beliefs.

Similar problem afflicts the third proposal. People have different overall views of reality. For some, it will be dictated mostly by science; for some, by religion; for some, by ideology or by spiritual experiences and systems and so on. Of course, if the theory fitted the *correct* overall world-view, this would enhance its plausibility. But this idea, aside of being impractical (except in discussions between persons in agreement about the overall world-view), misrepresents the order of confirmation. We can hardly arrive at the correct overall view of reality if we do not have the correct theories of particular areas of reality. Of course, we want to have a consistent world-view in which all theories fit; but the theory of the self should be made to fit other theories no more than they should be made to fit it. This point applies in particular to the issue of the relation between theories of personal identity and ethics, religion and philosophical *interpretation* of science.[40] There is no reason to accord any primacy to these theories over the theory of personal identity.[41]

We are thus left with the idea that deep beliefs in the relevant sense are those which underlie our common everyday practices and attitudes. These beliefs will also be likely to meet condition a). We are not willing to change our attitudes and practices without a strong reason. As for consistency with our considered overall view of reality, we would certainly prefer the beliefs presupposed by our practices to be consistent with it. But if one discovers a tension, there are two ways of achieving consistency, and the choice may depend on one's philosophical temperament. While some of us are reformists, many would rather take a Moorean common-sense line and reject a theoretical view rather than beliefs we take for granted in our everyday life.

What it is for a belief to be presupposed by a practice? Sometimes the situation is pretty straightforward. Practices are necessarily regulated by some rules. Suppose we take a rule "Only the person who perpetrated the crime is to be punished for that crime". If this rule is to be *applicable*, the world must contain persisting persons.[42] And since we do apply the rule, and we would not do so unless we believed it is applicable, it turns out that we believe that there are persisting persons. But only rarely can we seize upon such clear-cut rules. For example, is blaming regulated by a similar rule? Or by a rule that only a free agent can be blamed? There is room for many a debate here. But it is not controversial that for such practices to exist, we need to use certain *concepts*, like that of agency, and *distinctions*, like that between things done by the agent and not done by the agent. Secondly, practices usually

[40] Interpretation as distinguished from science as such. Theories should be consistent with data provided by science. But virtually every philosophical theory is.

[41] This statement concerns *theories*. The role of the concept in our pre-theoretical thinking and practices may give rise to demands on a theory of the self, as we said above.

[42] The sentence 'There are persisting persons' is not a logical consequence of the rule, but the relation between them is a close analogue of semantic entailment. In all models (worlds) where the rule is applicable, the sentence is satisfied. If we treated "*x* is to be punished for *y*'s crime' as a genuine relation in the world and considered the descriptive statement that such relations hold (which we may want to distinguish from the rule as prescriptive), then we would have a straightforward semantic entailment.

have a *point*. These observations give rise to two philosophical strategies. Our familiar concepts and distinctions can be analyzed by less familiar and avowedly more precise philosophical concepts. If only one analysis proves to be tenable, then we can be said to implicitly believe that the self is something like what the analysis of all these concepts and distinctions says (or, if there is more than one tenable analysis, that we believe in one of the proposed views of the self). Such beliefs would have to be taken into account by the theory because they are entailed by the use of concepts which have an important role (since their use is necessary for an important practice).[43] The way in which roles of our concepts place constraints on theories of personal identity has been analyzed above. We should note, however, how much room for argument is left after the analysis which purports to uncover deep beliefs involved in key concepts and distinctions relevant for the practice has been presented. One can say that *since* this is the analysis of the key concept for the practice, then *of course* this practice, e.g. blaming, is applicable to the objects satisfying the concept as it is analyzed. But one can also argue that we could not really blame people if they were such as the analyzis suggests – for example, on the grounds that the practice would lose its point if targeted at such beings.[44] One can then go on to resolve the incompatibility between analyzis and the practice in one of three ways. First is to say that the conceptual analyzis is not correct. Second, to admit that it captures our actual concepts, but revise the concepts and reject the theory contained in the analysis for the sake of preserving the practice. Third is to change or jettison the practice. Whether this is possible or not and rational or not is a matter for further argument.[45] To end this line of thought on a Wittgensteinian note. Disagreement over the

[43] Deep beliefs do not simply follow from the definition of ‘person’. But they may be uncovered by a deeper conceptual analysis. Secondly, they may be required not by the analysis of a single concept, but by the analysis of a web of concepts, distinctions and their roles in practices.

[44] The differing reactions may be due to differences in approach to concepts and instantiation. See Nozick (1981), 47-58 for a masterful analysis of this issue.

[45] These moves are similar to those we considered when discussing the relation of a metaphysical account to the concept playing a role in important practice. There, the content of the concept has been taken as given. Here, it is the analysis of the concept which is put into question. Is the analyzis correct? When the concept is analysed along a given line, is it fit to play the role we want it to play? If the analyzis is correct, and if the

relation between a practice and the analyzis of the concept may indicate that we implicitly follow different rules which in the actual practice happen to have the same results (or, as Wittgenstein would say, maybe we just do not follow rules at all). This possibility puts into question the assumption that, for example, what we call 'blaming' is a single shared practice underlied by shared beliefs and attitudes. In the light of these considerations it seems unlikely that the appeal to actual practices could yield well-supported and robust constraints on theories of personal identity.[46] Still, it is a striking fact that the acceptance of different rules, beliefs and attitudes gives rise to a more or less uniform practice on behavioural level. It can be suggested that what we all care about is precisely this, the tangible behaviour and its effects in the actual world. And this is the only thing we all care about. *This* is the point of "the" practice. If we follow this approach, we may abandon the futile question of what is rationally required by "the" practice, since different theories with associated attitudes and rules may have the same practical results. And only failure to yield these results would provide a neutral grounds for criticism, since these results are the only thing that we *all* care about.[47] The question which remains is whether the acceptance of the theory, together with appropriate attitudes and rules, would allow for the existence of a practice with the desirable results, given the facts about our psychology.[48] The discussion shifts from ethics to

concept is fit for the role, and if we want to preserve the practice, then the content of the concept *and* the associated deep beliefs constrain metaphysical theories of personal identity in the pragmatic manner.

[46] Does it mean that practices and metaphysics are altogether independent? Not quite. The appeal to *actual* instances of shared practices is inconclusive. But when we consider the extension of our practices to less familiar or openly science-fiction situations, disagreements emerge over what extensions or changes would be feasible and rational. And these disagreements are easily traceable to disagreement over metaphysics. This shows that it is not the case that actual practices do not presuppose metaphysical beliefs. Rather, for different parties, different beliefs underlie the practice. Or, we could say equally well, there are many different practices which converge in actual cases. One could of course debate to what extent the science-fiction cases are really possible. But this seems to me beside the point. Disagreement does not come into being when the hard cases actually emerge. The fact that responses to imaginary cases differ shows that disagreement is already here, albeit dormant.

[47] It can be argued of course that we all *should* have some particular and less straightforwardly practical concerns and not others. Thus the debate over personal identity will be affected by properly ethical debates. The move in the other direction is possible as well: concerns and ethical theories may be criticized on the grounds of presupposing false views about persons.

[48] We should keep in mind that even if a metaphysical theory fails to do so, it is still open for its proponent to claim that while remaining on the level of everyday practice we have to accept some practically necessary

psychology. The crucial consideration in assessing the practical and psychological viability of the theory is whether its acceptance can be reconciled with the having of relevant *attitudes*.

2.1.4. Attitudes

How then beliefs about the self may affect our attitudes? The answer which first suggests itself is that attitudes relevant for the problems of identity are usually *propositional* attitudes, i.e. their object is a state of affairs expressible by a proposition. The attitude itself will be expressible by a proposition of the form "I am A that P" where "A" stands for a verb expressing the kind of attitude or emotion and "P" for the proposition describing the object. Now, it may be the case that either the belief that P may be true or the very ability to frame P presuppose having certain beliefs, concepts or experiences. What could happen if we came to believe in a theory which would not square with these? First, the attitude could turn out to be *unjustified* or even *irrational*. Most interestingly, it could come out as *unintelligible*. This result could be a serious reason to reject a theory: if the attitude is familiar and for all intents and purposes makes sense to us, then the theory which makes it unintelligible simply does not square with facts and so is *untrue*. Less strongly, we would not know how to fit it in our system of concepts and beliefs and so it would not be *believable*. In view of this, it is incumbent on a proponent of such theory to explain how we can be so extensively confused and to show how we could come to believe the theory.

This idea is enough to get us going, but it is notoriously difficult to spell out clearly in what way our attitudes presuppose beliefs and conceptions about the self and how one should go about establishing the claims that they do. To get a bit further, we may do well to draw on the valuable work of Ronald de Sousa on emotions. On his account, emotions have an *object* (de Sousa uses the term "target") and involve focusing the attention on a

illusions, on the level of theoretical thinking we can believe in a different, true account. Thus the appeal to practices cannot even show that the theory is not *believable*. Things may be different with the appeal to attitudes.

property of the object (*focal property*), which, if it is a genuine property of the object and causes the occurrence of emotion, is called the *motivating aspect* of the object. The most interesting claim is that the character of feelings is determined by *paradigmatic scenarios*.[49] This feature of emotions is closely related to perception of *salience* of the properties of objects and situations and perceiving them as relevantly similar or different from other objects and situations (notably, those featured in paradigm scenarios). Now we may see that a change of views about the self may have an impact on our attitudes at many different points. The target may turn out not to exist, or not to have the properties which were supposed to constitute its motivating aspect. By analyzing the nature of the target and the motivating aspect ("that what matters") one can uncover new similarities and dissimilarities. Most subtly, a theory can change our perception of salience of certain features of our lives and of the differences between selves, forms of survival and life and death as such. In consequence, what was apprehended under one analogy or paradigm scenario may come to be considered under a different scenario. All these changes can affect the nature and intensity of our emotions and attitudes.

Having explained how views about the self may undermine our attitudes, we are in a position to complete our account of *deep* beliefs. If a view would undermine a set of attitudes, than an implicit presupposition of this attitude is that we are *not* something like what the view says we are. This is what we deeply believe. If a view were the only view consistent with a set of attitudes, we could be said to deeply believe in that view. Of course, it is reasonable to consider only the set of *important* attitudes.[50] A person could be irrationally afraid of black cats because of an implicit belief in a superstition, but if that was an isolated and unimportant attitude we would not say that she deeply believes in the superstition.

[49] de Sousa (1987), 201: "paradigm scenarios determine characteristic feelings, and an emotion can be assessed for its intrinsic rationality - a kind of correctness or incorrectness - in terms of the resemblance between a present situation and a paradigm scenario."

[50] Subjectively important. The road for arguing which attitudes *should* be important to us is of course open.

We can now consider how staying in tune (or not) with our deep beliefs and attitudes which presuppose them reflects on its plausibility. What do we say when a theory undermines our old attitudes? The answer depends on how high the attitudes ranks on several dimensions (independently of the special considerations brought in by the theory). First, the dimension of *rationality*. Do we normally regard this attitude as rational, better or worse justified, defensible or already suspect? Secondly: is this attitude *necessary* or is it possible to get rid of it (and if possible, how difficult and costly would it be)? Thirdly: is having it *good* for us? Fourthly: is this attitude *central* or marginal to our lives?[51] In general, the higher the position of an attitude in these rankings, the stronger the pressure on theories to preserve it. If a theory undermines a very high-ranking attitude it may be rejected on pragmatic grounds discussed in the section devoted to concepts. Another risk is that the theory may not be *believable*. If an attitude is too deeply entrenched, the only psychologically viable way of resolving the tension with the theory may be to give up the theory. Let us note, however, that even if a theory undermines an attitude in some way, it does not necessarily follow that this would oblige us, as rational beings, to abandon the attitude. First, the change of beliefs may just lower the ranking of the attitude: instead of being rational, it comes out as merely defensible; no longer central, it is just one of the options etc.[52] Secondly, there are other ways of accommodating the tension. Perhaps we are creatures doomed to inconsistency.[53] Or we may need practically necessary illusions, of which we can be free only in our more philosophical moments. Yet another possibility is to develop a two-level account of rationality. In general, strategies already familiar from ethical discussions will be applicable in this context. One kind of arguments deserves a separate remark The way it is often framed

[51] It should be noted that various kinds of “lives” may be in question: human lives; lives of persons; lives of sentient beings. Nature and force of arguments will depend on which of these levels is chosen.

[52] This can happen when the theory changes our perception of salience of properties and differences. This possibility could not be accounted for if we operated only with consistency with attitude’s presupposition, which is all-or-nothing.

[53] A line taken by Nagel (1986), 87ff.

is this: since a theory undermines important attitudes, believing it would make leading a human life impossible, therefore it should be rejected. I think that actually the best way to develop an argument of this sort would be to claim that believing the theory would make meaningful life of any kind impossible. In any case, the focus of the central claim does not rest on the relation between a theory and any particular attitude, but rather on the impact on the system of our beliefs and attitudes as a whole. If such claim could be established, it would give strongest possible reasons to reject a theory on pragmatic grounds.

2.2. Further demands on theories of identity

In the preceding sections I have discussed how *coherence* with our concepts, experiences, beliefs and attitudes affect plausibility of theories. I have emphasized the way in which the *roles* of the concept "person" in our systems of concepts, beliefs, practices and attitudes make it possible to reject a theory if it does not fit well into such system ("rejection on pragmatic grounds' as I have called it). But we may expect of a theory to go beyond mere coherence. We may want it to provide *explanation* and *justification* of certain phenomena and practices.

As I noted in the section *Concepts* such demands may be understood as postulates for preserving a role of the concept. I will call it the *conceptual* way of viewing explanation. Viewed in this way, these demands are still demands for coherence. But there is another way of seeing things. Theories of identity may be regarded in the same light as scientific theories. Call it the *quasi-scientific* view of philosophical explanation. On this view, theories should not be rejected just because they require revision of our concepts. The measure of their success is how well they help us in understanding and explaining a wide range of phenomena. The extent to which we regard a theory as plausible will partly depend on how much we expect of it, and, of course, what we take the *facts* needing explanation to be. If we take this line, demands of justifying practices or attitudes will be construed as demands for a specific kind of explanation. Namely, our belief that a practice is justified - and the truth of this belief - will be the data to be explained. The explanation will have the form of supplying reasons which would justify the belief and which are such that our belief may be traceable to implicit or incomplete grasp of such reasons. It would not do if we had true beliefs only by chance or for wrong reasons. For the demand for justification makes sense only if we make the assumption that our beliefs in question are to a large extent justified or rational.[54] If this assumption is wrong, then the demand is spurious, and consequently the fact that the theory provides a justification cannot be presented as an argument for its plausibility, insofar as we remain on the line of quasi-scientific explanation.

Having explained the nature of the links between providing explanation and justification and the plausibility of theories, I now turn to consider in more detail which demands made on theories are in fact reasonable. There are two demands which should be taken seriously. The first is to explain why our ordinary methods of re-identifying persons are reasonable and justified at least to some minimal extent. The second is the demand for providing means to explain and to justify the link between identity and concerns for which questions of identity seem particularly relevant.

Since a lot has already been said about the second issue, a brief note will be enough here. It seems reasonable to expect from personal identity theorist to answer questions such as: “how it possible to have and come to have concerns like ours if the theory is right?” “why these concerns *seem* rational/justified/important to us?” “what are our bottom-line concerns when we say we have concerns about selves, identity, life and death?” These are all questions about providing explanation or elucidation of mostly non-controversial facts. The case for demanding justification of our important concerns is much weaker if only because there is little agreement on how we should understand our concerns

[54] Could we perhaps rationally think that our beliefs are just *justifiable* while not thinking we have some implicit grasp on the actual justification? I don’t think so.

and which of them actually are important and justified. Therefore I would not regard failure to provide justification for our concerns as a self-standing argument against any theory.[55] In practice, however, there may be very little difference between an attitude coming out unjustified and this attitude - or a whole system of attitudes - coming out as unintelligible or lacking in value and meaning; especially if the attitude in question is a central one to our lives. Since such consequences could provide grounds for rejection of theories, arguments concerning justification of attitudes cannot be totally ignored.

Considering the explanation of the link between identity and re-identification practices, it seems more congenial to think in terms of the conceptual rather than quasi-scientific approach.[56] As I stated it, we may demand that the theory enables us to explain why our ordinary methods of re-identifying persons are reasonable and justified at least to some minimal extent. But much stronger demands are often made: that the theory should assure that our ordinary methods of reidentification are justified and usually reliable, or that they usually enable us to know the truth of identity-claims. Now, *from the point of view of the link between the concept and the practice* such demand is too strong. Consider an imaginary tribe whose members believe in the existence of ghosts. Shamans of this tribe are believed to be able to identify, communicate with and re-identify particular ghosts. Shamanistic practices are as much a part of the ordinary life of the tribe as any other practice. Imagine now that shamans start to dispute what is the nature of ghosts. Can they justifiably demand that any theory of ghosts assures that their common practice is reliable? No, since their practice is in fact *not* reliable and that is because actually ghosts do not exist at all (let us assume we are

[55] To put the idea in another idiom: I do not take the fact that a theory allows for divergence between identity and "what matters" to automatically lower the theory's plausibility.

[56] It used to be popular to make a direct conceptual link between identity and reidentification. But such strong link is not necessary. The concept "person" has a role in epistemic practices which seems not more controversial than epistemic practices related to other objects; and indeed all these practices are interdependent (since we rely on reports etc.). If the theory of identity gave us reasons to be sceptical about the practices regarding persons, it would generate a sceptical danger for the whole system of our beliefs. And we may either say that such scepticism is inadmissible, or that we have a very strong interest in preventing the rise of such worries. Therefore, the concept of persons we want to have is such that it should not embroil us in grave sceptical problems.

right on this point). Their demand obviously begs some very central questions: do ghosts exists? are our practices reliable? The moral I want to draw is that we have no right to assume that we are in a better situation than the shamans. Given concept C and practice of re-identification P, what can be demanded at the start of inquiry into the nature of C-instances is at most this. First, the analysis of the concept C should allow for an explanation of why it ever seemed to us that re-identifying C-s by P is a good idea. From this we can go on to a second demand that C-s should be such that, if they existed, practice P would be the *best practice available,* where the measure is the function of epistemic reliability and practical advantages. To my taste even this weak demand assumes too strongly that our ordinary practices are rational (rather than being mostly conditioned and reflexive without particular regard for rationality). But if one is friendly to such assumptions this condition is reasonable. Thinking along lines suggested by Nozick, we could even say that if P is the best satisfier of concept "knowledge-yielding method" *within a domain* to which C-s belong, then P does yield knowledge (even if in comparison with methods available in other domains it is a bad instance of the concept). Thirdly, P should come out as not significantly less reliable than practices of identifying other common objects, *given* assumptions about the ordinary run of things, *unless* good reasons for thinking that C-s differ significantly from other common objects are adduced. The last clause captures the moral from the Shaman's Dispute, but it makes the demand so weak as to almost seem vacuous.[57] But this is all right. If the theory denying the very existence of C-s should not be rejected just because it makes P unreliable, so *a fortiori* no theory about the nature of C-s should be rejected on these grounds. Secondly, it seems right that inquiries into reliability of methods and nature of objects are inter-dependent, none enjoying priority. Finally, while not performing as a pass/fail test, my

[57] To see that the condition is not vacuous consider a theory which would be excluded: cats are cleverly constructed robots which avoid dissection or any other scientific test, substituting bodies of cat-like organic creatures in such situations. This seems possible, but we have no good reason to think that *cats* as opposed to all other creatures are such clever robots.

condition may be used for argument if we assume that the less reliable, given the ordinary run of things, the practice turns out, the stronger reasons must be adduced for thinking C-s are in some way special for the theory to be plausible. It is at this point that the clause "given ordinary run of things" comes into play. The demands for justifying our methods of reidentification have been mostly used against Psychological Theory and Dualism. I think that these theories pass my reasonable modest conditions just fine. According to Psychological Theory, our practices of identifying persons on the basis of the identity of the body are reliable in the normal run of things. If there is a divergence, it is only in cases when things run very differently from the way to which our methods are adapted. Issues with Dualism are more complex, since it is sometimes denied that a Dualist is entitled to make the assumption that the ordinary run of things is that one self inhabits one body for the whole life. The gist of the problem is this: how the Dualist could justifiably rule out that bodies do not usually come to be inhabited by series of selves, say, a new one every day?[58] But consider the following situation. Our planet is supervised by Extremely Powerful Aliens. They do not tamper with us in any Matrix-like way. They just swap all the time ordinary macroscopic physical objects around us for their replicas. There is no widespread deception of senses, but our judgments about identity are systematically wrong. How could we rule that out? I think the cases are parallel and both are answered by inference to the simplest explanation (simplicity has an importantly *pragmatic* dimension: theories which allow us to live life in the way we want with minimum effort are simplest). Now, as Johnston notes, this presupposes that the behaviour of objects is law-like and that we have some empirical theory about it.[59] And of course every serious Dualist holds that embodiment is governed by some laws. Even if the link between a particular soul and body is contingent (and not every Dualist is committed to that) it does not mean it is chancy. Now, for the purposes of this discussion,

[58] Perry (1978), 9-17; Johnston (1987), 73f.
[59] Johnston (1987), 73.

my memory of my continuous embodiment is as reliable as anything (since it is not less reliable than my memories about physical objects, and these are necessary for reliable re-identification of physical objects). And since I have no particular reason to doubt the existence of other minds I can rely on similar reports of others.[60] Given this, assuming that I am an exception to a law would be implausible.[61]

Johnston actually rests his case against Dualism on an argument about the special case of reidentifying persons during the periods of unconsciousness. He claims that we would naturally say "Fred is here (and sleeps deeply)" without even thinking of any theory of law-like behaviour of simple selves. Now, I am rather surprised that Johnston does not take into account the fact that the belief that the soul does wander off during the sleep was pretty popular in many different cultures. It seems to me no less spontaneous and natural a response than Johnston's. Secondly, if one looks into open eyes of an unconscious man, one may experience a very strong unpleasant impression that "no one is there". This impressions seems spontaneous and natural, and it does not seem to rely on any particular theory. Finally, I do not see why the Dualist should be embarrassed by this example any more than by statements like "Tomorrow we're going to bury poor Fred at St. John's" uttered by a faithful Catholic (we may imagine the man saying in the same breath "We pray that Fred enters the house of our Father in heaven"). It is just a habitual carry-over from normal cases.

2.3. Thought-experiments

In this section I am going to state briefly what I think we can expect imaginary cases to show. I also have a few remarks about some common mistakes in framing thought-experiments.

[60] It is enough to slightly adapt the Extremely Powerful Aliens Case to show that Dualists and Reductionists are in the same position with regard to the other minds problem. This is nothing more than rehearsing Descartes' arguments preceding the Dream Argument and the all-encompassing doubt.

[61] To say in the present context "But you are relying only on your own isolated experience" is as irrelevant as saying the same in the context of knowing the identity and nature of ordinary physical objects. If one is willing to entertain sceptical scenarios like Extremely Powerful Aliens, then one is in just as bad situation in relation to the world and to the other minds. In this sense there is no *special* problem about other minds.

While methodology of thought experiments has grown in the recent literature into a topic of its own, I think that such general remarks as I make - hoping they will be mostly uncontroversial - will be sufficient for the purposes of my work.

My approach is based on three observations. First, there is some *minimal* agreement about understanding the word 'person'; but people have great many different detailed *conceptions* of persons, none of which is obviously absurd. Second, people respond differently to thought-experiments and even responses of a single person may not be consistent.[62] Third, while it is customary to say that thought-experiments are supposed to show what our "intuitions" are, the meaning of 'intuition" in this context is far from being clear.

These observations are jointly sufficient for establishing what we can expect from thought experiments. I will start with expanding on the third, as it will shed light on the remaining two. What can 'intuition" mean in the present context? First, we may think of linguistic intuition or intuition about the use of a concept. If that was the meaning, thought-experiments would be helpful in matters of conceptual analysis. And indeed they are; but as follows from the first two observations, there is no hope for establishing any particular view of personal identity as the correct analysis of a shared concept "person". The existence of different conception of persons and differing responses to cases indicate that, as Johnston rightly says, the common concept of person is quite unspecific. Moreover, it would be too naive to think that responses to thought-experiments are independent from our particular theories and conceptions. It is my experience that the deeper and more carefully one thinks of any case, the more likely it is that one's considered judgment will be shaped by one's previously held philosophical views.[63] On the other hand, considering imaginary cases does play a role in coming to embrace a theory. To conclude, thought-experiments play an

[62] These points are emphasized in the seminal work of Mark Johston (1987).

[63] The views need not about selves in particular. For example, whether one is a physicalist or non-physicalist will make a difference.

important role in articulating our *conceptions* of persons and in drawing out their consequences, but they cannot serve to establish one conception as following from the concept.

But perhaps we could have access to spontaneous pre-theoretical responses to questions posed in thought-experiments? *Such* responses could be called “intuitions”. Well, first of all, it is not clear just what would be the advantage of such responses. Why would spontaneous and philosophically un-informed judgments be superior or more likely *right* in any sense than considered and philosophically informed ones? In any case, this is an idle question since there is no reason to think such responses are available. First, to understand the very problems and thought experiments, one needs a grasp on many sophisticated technical concepts.[64] Secondly, philosophy is not insulated from general culture as it permeates religion, spirituality, arts and politics. Since anyone who attained the level of intellectual development allowing for the command of the concept “person” will be exposed to general culture, no respondent can be philosophically innocent.

Nonetheless, there is a right hunch (“hunch” will be the third analysis of “intuition”) in the thought about pre-theoretical responses. The thought is that our responses to puzzle-cases may be rooted in experiences of the self, attitudes and practices (such as practices of reidentifying persons) *rather than* in particular detailed theories. If our hunches were explicable in this way, they would indicate what we *deeply* believe, and this is something relevant for theories of identity. It should be noted, however, that experiences, attitudes and practices should not be assumed to be data which absolutely precede and are independent of our conceptions. Within any single culture conceptions of the self co-exist and

[64] I think that any philosopher who tried to get responses to standard puzzle-cases from laics will agree. To overcome unwillingness to entertain the required imaginings and to get over confusions and misunderstandings, one has to explain so much that the responder effectively ceases to be a laic. Until that point *no* real response is forthcoming. Of course I may be wrong, and these observations are due to my ineptitude at explaining things to laics.

interact with experiences, attitudes and practices.[65] Even if some of these are universal, this does not show that they are independent of any conception of the self. For it is possible that some such conception (which would have to be a rather thin one) is universally implicitly accepted.[66]

What follows from the above discussion is that our responses to cases are nothing like ready conclusions which could be used for or against any theory. They are, in fact, a *starting point* for a serious inquiry. The first task of the personal identity theorist is to provide a credible explanation of our responses, especially of cases when very similar cases tend to elicit contradictory responses. But if we have abandoned the idea that these responses are primarily due to features of a shared concept, we will not expect the theorists to provide explanation in these terms. Plausible explanations of our responses will point to a wide variety of factors, some of which will be judged irrelevant to inquiry into personal identity.[67] The primary effect of discussing and explaining responses will be simply a better understanding of the richness and complexity of our beliefs and attitudes. Cases and responses may also throw into relief particular features of our experiences, attitudes and practices which, it can be argued, should be accommodated by theories. It is to this heuristic use that I will most often put imaginary cases. Following such exploration, clarification and explanation, it is finally possible to argue that a select set of our responses represents our deep beliefs. Such conclusions will serve to criticize and support theories. I believe that this account pictures quite faithfully the actual procedures used in contemporary literature on personal identity.

[65] Reflection on the self is a universal human pursuit. Even if a culture does not develop philosophy, reflection on the self will be occasioned by reflection on phenomena of death, sleep and dreams or by shamanistic, meditative or drug-induced experiences. Such experiences are known in virtually every culture.

[66] Actually it would not matter much whether a theory could be somehow established on the basis of our universal pre-theoretical responses or turned out to underlie our experiences, attitudes and practices. What matters is whether the theory is *consistent* with them and makes them intelligible. Still, it is better to be clear that the relation is that of two-way and not one-way dependence and influence.

[67] For example, responses may depend on the manner of presentation of the case or on the habit of generous interpretation of fiction. Williams (1970) 52-55; 61f; Unger (1990), 11-13.

2.3.1. Unjustified assumptions in thought-experiments

Many central cases used in discussing personal identity involve some strange happenings to our brains. Happenings which so far have not been in fact observed and some of which are admittedly not practically, or even physically, possible. It is thus not surprising, that there are radically sceptical views about the claims of possibility involved in these thought-experiments and about their usefulness, even if the described happenings were in some sense possible.[68] I do not endorse these views. I have little to say about the scepticism about possibilities. I rely on the popular intuition that metaphysical possibility is much wider than physical possibility and is closely related to conceivability. And then, without the use of metaphysical possibility and thought-experiments one simply cannot do much interesting philosophy. This is my pragmatic reason for rejecting the view. As for doubts about the usefulness of thought-experiments, they seem mostly to rest on the following idea. We acquire and use our concepts in familiar circumstances. The concepts are made for application in these circumstances. So if we turn to consider wildly different situations, we should not expect that our concepts will be determinately applicable, that we will be able to confidently apply them or that they will be applicable at all. We should not expect our responses to imaginary cases to be very illuminating about our use of the concept "person". Now, all this seems to me largely right (except the radical claim that the concepts would not be applicable at all). But this would be worrisome only if we thought that useful responses to cases should follow from our grasp and use of the concepts "person", "self", "the same self" etc. And I have already argued that there are good reasons to abandon this idea. The very non-specifity of our shared concept of "persons", while making thought-experiment not terribly useful for conceptual analysis, makes it possible to frame and discuss far-flung imaginary cases. For the less specific the concept, the harder to find a situation where it could

[68] Johnston (1987); Wilkes (1988), 1-48; Gendler (1999).

not apply. To suppose that we would not be able to apply the concept at all in unfamiliar circumstances is to assume that it has a fairly substantial content, and this, as we said, seems wrong. Since framing imaginary cases seems justified, we are free to try to make use of them. Of course, it is still possible to argue that our responses are so indeterminate, ridden with contradictions and dependent on irrelevant factors as to make them useless. But this cannot be established *a priori*. We will have to examine thought-experiments one by one.

I do not find fault with thought-experiments on a wholly general level. I find fault with particular assumptions made in particular cases. Let us first consider without prejudice what science shows about the relation of mental and physical states. Thus far, science tell us something only about the relation of mental states to states of organic brains which are parts of living organisms.[69] And what science shows is that in normal circumstances, exercise of mental capacities depends on the presence of particular physical states. It will be handy to say that the brain sustains mentality, where "sustaining" is intended not to determine the exact nature of the relation between the two. The nature of these relations is *not* shown by science. If it were, doing philosophy of mind would be pointless. I will assume that it is not pointless.

Now, sometimes the way thought-experiments are framed involves assumptions which go well beyond what science shows and which are for this reason more or less question-begging. An obvious example is the assumption that the relation between the physical and the mental is one of strong metaphysical dependency. For instance, that mental states are *realized* in brain-states. That obviously begs the question against Dualism. But since the bulk of contemporary discussions is actually an internal dispute between Naturalists who can share the assumption, this may not matter much. More importantly, this assumption seems methodologically unsound. For it is assumed that we can rely on conclusions achieved

[69] Science tells us also about *behaviour* of non-living systems such as computers, neural nets, robots etc. To say, however, that it shows us something about the relation between the physical and *mental* states of such objects, is to take a stance which is controversial in the way that saying that biology and psychology show something about the relation of physical states of organisms and mental states of persons is not. Whatever cannot be agreed on by everybody cannot be included in an unprejudiced account.

in Philosophy of Mind in discussion of Personal Identity. But the latter is not posterior in this way. Considerations of problems in personal identity have immediate bearing on problems in philosophy of mind. Moreover, there is considerable overlap as consciousness is a central topic for both. The two disciplines are on the same level. Of course, one needs to start *somewhere*. But then, in making assumptions about the nature of the relation between physical and mental states one should be clear that what one primarily achieves by argument is to show the structure of logical consequences and consistency of one's overall position. Whereas many arguments in personal identity discussions seem to go along some such lines: "Assume (implicitly) that brain-states realize mental states. Therefore when such-and-such imaginary events happen to our brains, our life and consciousness will have to have particular features. *This* shows that such-and-such views about our lives and consciousness and persons are mistaken." Such arguments seem to me pretty useless, unless we have good independent reasons to think our life or consciousness would have the required features. And since the circumstances are usually unlikely to be ever realized and observed, it is not likely that any such reasons would be forthcoming.[70] To the extent that we have independent reasons to hold beliefs about consciousness and persons which the argument shows to be incompatible with assumption that brain-states realize mental states, such argument would constitute a reductio of the Naturalist assumption.

A similar danger is posed by assumptions concerning what can sustain mental states. Consider Bionic Replacement. In this case, cells of my brain and body are step-by-step replaced by a functionally equivalent non-living robotic parts. The conclusion one may draw is that the result is that I survive. The case is used primarily against Animalism. But of course

[70] It is good to realize how strong convictions many philosophers have e.g about Fission, and realize that these convictions rest almost entirely on assumptions about the nature of mind-brain relation. Commisurotomy does not take us anywhere near enough fission to justify such strong intuitions. Nagel's seminal discussion of commissurotomy (Nagel 1979, 154-164) based on Sperry's studies (see Sperry 1968) is still one of the most nuanced.

it works only if one makes the assumption that artificial systems can sustain mentality.[71] Since an Animalist may deny that assumption, the force of the case is weaker than it initially looked. This case is pretty straightforward, since it does not involve appeal to the current findings of science. It is much harder to notice the danger which lies in claiming the power to sustain mentality to *parts* of brains. I will shortly discuss two cases. Giles presents a case where a disconnected part of the brain - too small to sustain a mentality of a person - is in a state normally sustaining pain. He claims that this shows that there can be ownerless pains. Unger devotes much attention to the Spectrum of Decomposition, where tiny parts of my brain are cut out, until we get down to a single cell. The operations are conducted in a way which leaves a functional system as long as possible. Unger assumes that systems around the middle of the spectrum could sustain some mentality and consciousness just as well as animal brains of comparable size could.[72] The moral Unger draws from the consideration of the spectrum is that it is indeterminate at which point personhood and consciousness would be destroyed by the process. There are, however, good reasons to resist the assumption that *parts of brains* or *radically mutilated* brains could sustain mentality. First of all, science so far shows us only something about mentality of *whole* brains under *normal* circumstances i.e. when the brains are safely within original bodies. And it is very doubtful whether science *could* show us anything about mentality of parts or radically mutilated brains, because the processes occurring within such objects would not be normally connected with behaviour. Claiming that these processes do sustain mentality not only goes beyond possible scientific evidence but also ignores all views in philosophy of mind which claim that the link with behaviour is essential to mentality. Analogy with animal brains is also unsound, since we are comparing normally functioning brains with radically mutilated brains at best and with mere parts at worst. Moreover, however we cut up a human brain it does not seem possible to cut it

[71] We may note that in contemplating the case this implicit assumption may be granted due to the habit of generous interpretation. It is not so easily granted in discussions in philosophy of mind.

[72] Unger (1990), 62.

down in such a way as to obtain an exact functional equivalent of, say, frog's brain.[73]

Brain-damage and its adverse effects on mental functions is by now a familiar and intensely studied phenomenon. The problems posed by the Spectrum of Decomposition: "how much damage must happen for the brain to be unable to sustain consciousness?"; "is there a determinate threshold of damage required to destroy capacity for sustaining consciousness?" are certainly valid ones. But we must not formulate the cases in such naive way as if the size or the mere quantity of cells in the brain were the only factors relevant for supporting consciousness! The brain is anything but homogenous. Whether there is a single region in the brain on which consciousness depends and which region it is, is a matter of debate among neuroscientists. It would be wisest to wait for some kind of scientific consensus to emerge on this issue before proposing cases relying on unjustified assumptions. For all we know at this stage, the presence of consciousness may depend on simultaneous activity of many parts of the brain, on the balance of neurotransmitters, and on the way neural connections are configured. I see thus at least two general hypotheses which could yield the result that within Spectrum of Decomposition there is a determinate threshold of damage. First, a particular kind of functional structure may be necessary for consciousness. While it may be possible to maintain *some* kind of functional structure very far into spectrum, the right kind of structure could collapse at a precise point.[74] Secondly, presence of consciousness may demand crossing a certain definite threshold of density and activation of neural connection (and possibly of many other quantifiable factors). Brains of very different sizes and structures could meet such condition and thus sustain consciousness. But since in effect we would have a *law* expressible in quantifiable terms, there would be a definite

[73] Of course we could cut *and* stitch to obtain something like that. But that would be an altogether different case.

[74] Consider an analogy: circulation of traffic in a city (make it a closed system). Let us stipulate that the traffic is working iff it is possible to get from any point near any road to any other point near a road. If we start to close off roads, the traffic may be working for some time. But it is entirely possible that if we close *just one more* road, the traffic will not work. This is just one example of a model not susceptible to the spectrum problem.

answer to the question about the threshold of damage done to any particular brain. In such model it is possible that as one approaches the threshold, the brain increasingly *fluctuates* from consciousness to unconsciousness. Unger stipulates that all changes would be made without interrupting the processes which normally generate consciousness. This inclines the reader to think that consciousness must “fade out” rather than be abruptly “switched off”. But we shouldn’t let the “fading-out” picture be foisted on us. By now it is clear that it cannot be taken for granted that it would be possible not to interrupt consciousness. That is an unjustified and questionable assumption. It is perfectly plausible to think that as brain-damage progresses, consciousness becomes more and more disconnected and episodic until it ceases to occur at all. Here we have graduality perfectly consistent with absoluteness of consciousness.

Chapter 3. Psychological Theory - Metaphysics

3.1. Reductionism

Psychological Theory is the first Complex Views of persons that I want to discuss. Complex Views attempt to provide an informative and non-circular analysis of necessary and sufficient conditions of personal identity. This commits them to ontological Reductionism about persons. The link is clearly stated by Noonan:

> a defender of this thesis [that a diachronic criterion of personhood can be informatively specified] *is* committed to reductionism about persons as this notion is characterized by Dummett, since he is committed to holding that the truth-conditions of statements in which the term 'person' occurs can be given by statements in which the word 'person' does not occur.

It should be noted that Ontological Reductionism may take weaker forms. Following Robinson, I take the minimal requirement for ontological reductionism to be the "a priori sufficiency of the base" thesis:

> (R) The bottom level down is conceptually and a priori sufficient for the higher level states.[75]

Reductionism connects the identity-questions and what-is-x-questions. Reductionism about persons would minimally claim that existence of certain non-personal objects and relations between them is conceptually sufficient for the existence of persons. Yet particular reductionist theories will differ in their characterization of the categorial status, fundamental nature and ontological structure of persons.

The essence of Psychological Theory lies in the thesis that *psychological* relations between objects account for the *diachronic* identity of persons. This, by itself, tells us little about what persons are. Yet in acordance with my postulate (P) any theory of personal identity

[75] Robinson (2009), 532.

must be able to provide some plausible answer to this question. And actual accounts of persons developed in the PT paradigm do provide some answers, full or partial. I will classify them according to their characterization of the ontological structure of persons. In general, three reductionist approaches to the ontological structure of higher-level objects seem possible:

(1) Higher-level objects are *identical* to some bottom-level objects.

(2) Higher-level objects are *constituted* by bottom-level objects.

(3) Higher-level objects are *logical constructs* out of bottom-level objects.

If approach (2) does not collapse into one of the others, then constitution must not be identity and higher-level objects like persons must be fully real objects not identical to any bottom-level object. Therefore the inventory of what there ultimately is must mention persons. This view certainly is not reductionist in the strong Parfitian sense, since there is an important *further fact* of the existence of the constituted thing. But constitution views may meet the condition (R) of reductionism.[76]

3.2. Identification Approach

Persons could be either type-identified or token-identified with bottom-level objects. It is more plausible to think in terms of token-identification, but my arguments would apply to either option.

Since PT is a Complex View, persons are not to be identified with any simple thing. Nor can they be identified with any ordinary material object. In that case their persistence conditions would be physical and not psychological. So what are persons? The classical answer is that persons are minds. Minds may be conceived as *collections* of mental

[76] But note that a position like Baker's which admits of real emergent properties and objects is certainly not reductionist; Baker (2000), 12-17; 25; Baker (2007), 263. Thus it is no surprise that Baker's theory is a version of the Simple View; Baker (2000), 146.

items: experiences, thoughts, memories, desires etc. linked by psychological relations.Nowadays, the most popular identificatory position is probably the 4D view: persons are mereological *sums* of person-stages linked by psychological relations.[77] What these proposals have in common is that they identify persons with composite *set-like* objects. By set-likeness I mean the following structural property: the identity of the set-like object is wholly determined by the identity of its components.[78] Why persons are identified with set-like objects? Because there is nothing else to identify them with, given that they are neither simples nor ordinary material objects *nor sui*-generis objects constituted by or constructed from bottom-level objects.

The fundamental problem with this approach is the danger of excessive essentialism/modal inflexibility. Given that the identity of a set-like object is wholly determined by identity of its components, the object could not have different components than it actually has. This problem has been widely discussed in connection with 4D ontology. Here I want to draw attention to the way this problem appears from the first-person perspective.

Suppose persons are sums of person-stages. I am a person. At some point, I face a choice: I can either do *F* or *G*. This choice will affect the course of my life from that point. There are thus two possible sums of person-stages extending beyond the point of choice. Call them *F-sum* and *G-sum*. When I think of the future before the choice, I have to think: “either *I* will live like this, or *I* will live like that”. I can have experiences which belong to person-stages in *F-sum* or in *G-sum*. Whatever I choose, the future person will be *me*. From this point of view I am *identical* to two possible persons: *F-sum* and *G-sum*. Now suppose I did F. So I am *F-sum*. When I think now about *G-sum*, I have to say this is a *different* person. *G-*

[77] Lewis (1976*a*), 59.

[78] I do not mean to claim that any theory which could be labeled as “bundle” theory is committed to set-likeness of persons. Bundle theories may embrace logical construction approach and perhaps constitution approach (though a constituted thing seems more than a mere bundle).

sum has different members than *F-sum*. So we have a striking discrepancy between the way I regard possible persons *ex ante* and *ex post*. Moreover, if G-sum existed, F-sum would not exist. However, I am F-sum. So if G-sum existed, I would not exist. But this means that it was not possible for *me* to do G; for if I did G, then I (*F-sum*) would not have existed! Modal inflexibility entails necessitarian fatalism. These consequences are unpalatable. If I believed I am a set-like object, then apparently I could not coherently think "I could do this, but I could also do that". But such thoughts are necessary for thinking of oneself as agent and for any kind of planning. If I believed in this theory of personal identity, I could not regard myself as person and live like one. Theory with such features is a non-starter.[79]

The standard response to worries about modal inflexibility is Lewisian counterpart account of modal semantic.[80] This account allows us to say that I could do otherwise, for all that it means is that there is *another* (possible) person who is similar but not identical to me. The main problem is that when we think about *our future* then we simply cannot think in terms of counterparts. Perhaps we can do it when we imagine possible *parallel* worlds. But suppose I wonder whether to go to a cinema tomorrow. Weighing the advantages *for me* of alternative courses of actions, I try to imagine from inside what it will be like *for me* if I decide to choose one. Then I decide "I will go to the cinema". Now, I am presented to myself in exactly the same way when I still consider my options and when I have decided. "I" is used in the same way when I say "I can" and "I will". I have to *identify* the subject of imagined experiences and actions as *me*. Otherwise I could not say "I can do this and I can do that"; and that is essential to my deliberation.[81] Consider now how I would think about going to the cinema if I thought in counterpart theory terms. Then I am confronted with two possible worlds. In one, cinema-going person exists. In the other, there is

[79] See ch. 2, 21.

[80] Applied to persons in Lewis (1971).

[81] When considering our future options we do not seem to think in terms of possible worlds at all. Rather, we think in terms of abilities, capacities or potentialities of actual objects. I would say that I ascribe such modal properties to the me who is here-and-now.

a home-staying person. My relation to these worlds is the following: *I do not know* which world is actual, and which person I am. But I am one of these persons and not the other. This is just how we do not think about our future. Counterpart semantics may seem to offer a passable account of modal predicates. But it gives a completely unbelievable account of semantics of personal pronouns, especially the “I”, in counterfactual and temporal contexts.

The idea that if something else happened than what is actually going to happen in one’s future, one would not have been there from the start, probably strikes most of us as terribly strange. It is, I contend, as strange as the thought that your identity to someone can depend on there being a third person related to both of you. There is an intuition in cases of Fission that whether you are the post-fission person cannot depend on facts other than facts about the two (or one) of you and your relations to one another. This intuitive principle was called “Only *x* and *y*” principle. I think that there is an analogous *temporal* principle. This is: “Who I am now depends only on what happened until now, and not on anything which may happen in the future”. My identity now is a given. Various things can be done in the future by *me* and this does not affect my identity now. This principle seems just as intuitive as the “only x and y” principle. We may call it “Only past and now” principle of identity of persons. Lewisian accounts of persons necessarily violate it.[82]

Another problematic feature of temporal wholes is that they do not exist at any single time. When applied to persons, this seems rather absurd. For we would have to say that, strictly speaking, at most a part of any person exists now. I think it makes little sense to say: „A part of me is present now” or „I am partly present now”. But this is what we would have to say. Moreover, these facts seem to undermine the agency of persons. If I am not around at the time an action is taken - only a part of me is - then it is hard to see how *I* am the agent doing the action. Now, it may be protested that there is a clear sense in which I am

[82] Cf. Johnston (1987), 68.

doing the action: I am doing it because I have a part which does. Perhaps it makes the ascription of action to me *derivative,* but not in any harmful way. Take the case of ordinary physical parts for comparison. I can say “I am writing”. Even though many parts of me are involved in this action, not all of my organism is. My jaw and teeth are not involved. I could lose them in some gruesome accident and still be capable of writing. So, strictly speaking, only a part of me is doing the writing. But this does not stand in the way of ascribing the action to me. However, the cases are *not* analogous. The reason why it is proper to ascribe the action to the whole organism is that its parts would not be capable of performing the action and presumably would not even exist, if there was no organism that they were parts of. The existence of the whole and its relation to its parts is thus relevant to the occurrence of actions. Not so with persons extended in time and their parts. Proponents of 4D ontology seem committed to holding that temporal parts are capable of existing independently of the wholes and that they have all sorts of properties that the wholes can have (except, of course, properties entailing having a temporal extension exceeding that of the part).[83] This is required for making sense of the notion of temporal parts and for temporal parts to perform their job in 4D account of change.[84] To conclude, person-stages seem to be real agents in 4D world, while persons are explanatorily idle. Persons are likewise idle when it comes to accounting for our attitudes. Consider Fission from first-person perspective. We are told that we have two (or more) persons co-existing before fission.[85] Since after fission the two go their separate ways, it would seem that I could look forward to only one of the post-fission lives. This is wrong, however, for it is indeterminate which person *I* am. So even if, from third-person perspective, we can say that one person should anticipate only her future life, from

[83] Note that claiming that some properties, especially mental ones, cannot be ascribed to *momentary* objects would not defeat temporal parts. It would only show that temporal parts necessarily have some minimal duration. A move to strong diachronic holism of the mental could, however, be fatal to the 4D account. If the minimal duration of temporal parts would have to be counted in, say, years, then little sense is left in talking of temporal parts.

[84] Sider (2001), 56; 64-65.

[85] Three on Noonan’s account; Noonan (2003), 216, 227f

first-person perspective I cannot say which of the persons I am, and so I cannot say what I should anticipate. Whether I will anticipate post-fission lives will have to depend on the type of considerations explored by Parfitian Reductionist, which leave personal identity aside. So personal identity is irrelevant to anticipation on 4D accounts. Analogous argument can be given for Fusion and responsibility. In the end, 4D theorists are bound to agree with Parfit: personal identity is not what matters.[86]

If there are persons worthy of this name in 4D world, then these are momentary person-stages.[87] In their case none of the problems with agency and identity we discussed above needs to arise. Lewisian metaphysics can lend itself to Transience View just as well or better than to Reductionism. Moreover, Transience View can deal with the arguments I constructed without relying on Lewisian metaphysics. Strawson distingushes two ways of thinking of oneself.[88] I can think of myself as subject; that is, on his theory, as momentary subject.To express this way of thinking about oneself, Strawson uses "I*". But I can also think of myself more loosely, identifying myself with a human being. Consider now how I* can think of possible persons after the choice. If I* think of my continuants in terms of "I-human being", then there are no obstacles to saying that I-human being am identical with them. For both now and then, there will be one and the same human being, and we need no deep metaphysics to be able to say that. On the other hand, if I* think in terms of momentary selves, then I* can think only in terms of survival: what can happen to my *successive selves*. So I* do not think I* am identical to A* or B*, and A* is right in thinking he is not identical to B*. So Momentary Selves theory has a simple and metaphysically neutral answer to at least some of the problems I posed in this section.

Why then a Lewisian would prefer to be Reductionist rather than Transience theorist? The only reason, it seems, is the wish to preserve our concept "person" and our

[86] Cf. Sider (2001), 202ff.
[87] Lewis himself talks about person-stages as if they were short-lived persons or subjects. Lewis (1983*b*), 76.
[88] Strawson (2005), 68f.

intuitions and interests concerning persons. However, the Lewisian account is in itself strongly revisionist in that persons do not have many features which we (relatively) pre-theoretically take them to have, but turn out to have many very surprising features instead. And we just do not think of persons in Lewisian terms at all. Given that, I do not see why should one still bother to provide a model in certain respects *isomorphic* to common-sense thinking. Noonan, for example, devotes a few pages of intricate argumentation to show that his answers to questions about Fission are in agreement with the deliverances of our intuition (which is supposed to give more or less clear verdict).[89] But the right answers - if they are right - are given for reasons completely unrelated to the reasons we actually have for our intuitions. There is just no way that we could implicitly have a grasp on something like Noonan's reasoning. The fact that there is a de facto agreement between the theory and intuitive answers in and of itself does not enhance the plausibility of the theory if this theory is irrelevant to the explanation of why we have such intuitions (though it protects the theory from charges of obvious implausibility). It is also impossible to claim that Lewisian theory follows from the analysis of what is implicitly contained in our concept. The way which we have to think about the relation between our concept and Lewisian theory is rather this. Given that the world is what it is, i.e. given the truth of Lewisian ontology, sums of person-stages are *the best realization* of the concept "person" available. And so, they *are* persons. Like many others, I say that they are just not good enough. If this is the best we can get in a Lewisian world, than in a Lewisian world there are either no persons or only momentary ones. The reason is that Lewisian accounts do not manage to save the features of persons which are relevant for our persons-related interests. I have argued that making the person a set-like object undermines agency. Others argued at length that such theories cannot account for responsibility, compensation, concern with survival and self-interested concern in

[89] Noonan (2003), 228-231.

general.[90] I conclude that PT cannot be plausibly stated in terms of identification of persons with bottom-level set-like object. Transience View seems superior to this version of PT.

3.3. Constitution Approach

The constitution view differs from the other two Reductionist views by introducing a genuine new substance in addition to bottom-level objects. This new substance is a whole *constituted* or *composed* of bottom-level objects. Now, "constitution" is used in a variety of ways. It used to be quite common to say that a thing is constituted by a material, or a portion of stuff, or an aggregate of molecules or the like. On this usage, "constitution" is equivalent to "composition". But the most common usage these days is to use it for a relation between *two coincident objects* which are, so to say, on the same ontological level. So one ordinary object (an object with a *form* we may say) constitutes another object of the same kind.[91] Say, an anvil constitutes a doorstop. Thus, given stuff S and two apparently coincident things made of that stuff: A and D, we can have two models of constitution. Model 1 says that S constitutes A and S constitutes D, but A and D are not related by constitution. Model 2 says that S constitutes A and A constitutes D. To add to confusion, the idiom of constitution is also being used to state views where constituted objects are "nothing over and above" bottom-level objects; views which in my scheme would be classified as Identification or Logical Construction Views.[92] Now, I want "constitution" to signal the special substantial status of the constituted thing (as against "nothing over and above" reading), but I will stick to the old usage, treating "constitution" as equivalent to "composition" (if only because we do not always have the analogues of lumps of clay or anvils).[93]

[90] Haksar (1991) 158-180, 188-228; Schechtman (1996), 51-66.

[91] Cf. Lizza (2006), 63 on two types of constitution.

[92] So, for example, Parfit adopted the idiom of constitution without changing the substance of his logical construction view, Parfit (1995), 295-297.

[93] The reason I do not stick only with "composition" is that Identificatory and Logical Construction views have equal right to use the term. A term like "constitution" is needed to signal the special status of the higher-level substance in Constitution Views.

What are person composed of? First, they may be constituted by a series of things which sustain our mentality; things like bodies, animals, brains or souls. In Locke's times, one would call such things "thinking substances"; in our times they are often called "realizers". The series of realizers constituting a person may contain just one element. The person is nonetheless not identical to this sole realizer since the person could transfer to another realizer and the realizer could cease to constitute the person. Regarding persons as composed of "realizers" is a popular option in the PT camp. Locke himself, the father of PT, embraced such view. Among contemporary proponents of PT, Shoemaker provides the best example of this approach.[94] Secondly, it is possible to think that persons are composed of person-stages but are not straightforward sums of them. I know of no proponents of this view. Finally, one can regard persons as composed of mental items: experiences, thoughts, dispositions etc. This view goes back to Hume. It seems to be little favoured now, but it has an important virtue. It expresses clearly the crucial insight of PT: *persons are essentially minds*. If a person continues to exist, this is because her mind continues to exist. This thesis is held also by proponents of the other views; even if persons are not just minds, their identity depends on the identity of their mind. Since this is so, we could switch at this point to discussing identity of the mind. But I hope it will be less confusing if I continue to talk of persons while thinking implicitly along the lines of the Humean approach. If my arguments rely on taking persons to be composed of mental items, they can be translated into mind-talk and applied to other views *via* the thesis that identity of the mind is essential for identity of persons.

From the metaphysical perspective that I adopted in this chapter, it seems that the main motivation for Constitution View PT lies in the desire to avoid problems of identification strategy of 4D ontology: problems of modal inflexibility and counterpart

[94] Shoemaker (1984), 113f. See also Unger (2000), 283. I should note that in recent works Shoemaker embraces the One Realizer view which I discuss in ch. 5; Shoemaker (2004) and Shoemaker (2008). This is still a constitution view, but intermediate between PT and PBA.

theory. Therefore the issues of trans-world identity will remain in the focus of my discussion. This will also lets us stay close to our guiding question "what are persons?" While PT theorists focused on persistence conditions of persons, this question has been somewhat neglected.[95] I will try to remedy this situation by applying the notions of form and matter to the psychological realm.

3.3.1. Form

The main task for the constitution theorists is to specify the relation between the constituted substance (henceforth "the whole") and its constituents. On the one hand, the whole cannot be something *separate* from its constituents, something which is only *caused* to exist by them. On the other hand, if the constitution approach is not to collapse into either identification or logical construction approach, the whole must be a new and robustly real substance. Its existence has to bring in more than the fact that a multitude of constituents can be viewed as a whole. There are two ideas which motivate the talk of constituted substances. First, a substance is a *unit*. To be an individual substance is to be something internally unified and distinguishable from the surroundings. Secondly, some wholes seem to have *peculiar properties* which are not straightforwardly reducible to properties of the components. If the components make up a unified whole with such peculiar properties, there is motivation to postulate the existence of a constituted substance.

The themes of unity and peculiar properties are brought together in the idea of *form*. We can think about form in two ways. Looking bottom-up, the form is primarily the *organization* of components which accounts for their unity and the existence of the whole. Looking top-down, we will take the form to be *a set of essential properties* peculiar to the

[95] Persistence conditions say little about what a thing is; trans-world identity conditions say a lot. Given that problems of trans-world identity are actually the main problems of PT, it is surprising they were given relatively little attention. One may think that these are not pressing, since everybody has problems with trans-world identity. Yet even if it is true that everybody has *some* problems, I think that a PT theorist has *more* of them than others.

whole. In any case, having the (same) form is necessary existence and persistence of the whole.[96] The form also explains how the whole can survive a change of its matter. A thing can survive such change because (and only if) its form is preserved. This condition is met in many cases of material change. The main appeal of constitution views is their ability to account for persistence through changes in a way consonant with our intuitions. Another important virtue of the notion of form is that it can be used in the account of *what* a thing is. We can say that a composite substance is the matter organized by the form.[97] Modern constitution views are rarely couched in terms of form, but the concept is eminently useful and I see no reason to shun this Aristotelian legacy.[98]

How can we apply the notion of the form to persons? Let us start with an analogy. Consider a ship and what it can survive:

<u>*Ship*</u>

- can survive some replacement of planks (*components)*

- can survive some changes of shape (*organization* of components)

- has to retain the function (*capacities* and *external relations)*

The form of the ship is, roughly, a ship-shape-like organization of planks which makes the whole have the capacity to sail. As long as such form exists, a ship exists. The identity of particular planks is largely irrelevant for the identity of a particular ship. But perhaps not altogether irrelevant. For we may also ask what makes the ship *this* particular ship. What makes it identical to a *possible* ship? Identity of *some* planks may be relevant to these

[96] In talking about *the* form we may mean either the *type* form or the *individual* form. This distinction will come into play later on in the argument.

[97] This terminology seems to me clearer than the modern vague talk of constituted substances being something more than components but not separate from and “over and above” them.

[98] See Haldane (1999) for another defense of application of the form-matter distinction to contemporary debates in philosophy of mind.

questions. Let us see if we can think of persons along similar lines. It seems to me that people with inclinations towards Psychological Theory think more or less in the following way. Their identity is defined for them by possession of certain experiences and personality (*components).* This, for some people, should result in a roughly similar *narrative.* Say, „marrying and having children" would be constant across all possible lives of the person, even if it were possible to marry someone else than one in fact married.[99] But becoming a Buddhist monk would be impossible. Less specifically, the life should be characterized by psychological *unity*: having a rather coherent psychological makeup, continuity of memory, desire-intention-action links etc. Narrative and psychological unity relations correspond to *organization and external relations.* Thirdly, there are social relations and roles; most notably those connected with one's origins: being a daughter of such-and-such parents, being a sister, a Russian etc. (*external relations).* The possession of basic mental capacities common to all persons is presupposed *(capacities).*[100] To conclude, there is enough correspondence to familiar examples of form-matter distinction to justify applying it to persons.

3.3.2. Form and trans-world-identity

The form can be considered as a *type* of organization of components. Specifying the type-form and matter gets us close to saying what a thing is. It gives some answers to persistence questions. But it does not give the full answer to questions about individuation and trans-world identity. Obviously, the same type-form can have many instantiation in a world; but one cannot be identical to all of them. Even if we tried to make the psychological form of a given person very specific, it seems impossible to rule out many instantiations (think of Fission or Twin Earth). So perhaps we should turn to the *individual* form. The problem is how to cash out this idea. If we do not want to engage with scholastic ideas about essential

[99] This should be constans at least across lives not ending with death in infancy or adolescence.

[100] This, it seems to me, is the thinking which leads people to believe Psychological Theory. Psychological continuity relations favoured by philosophers are not prominent here.

forms, we may conceive the individual form along set-theoretic lines: as a set of trope properties and relations. But then again, identity of the form will be fixed by identity of its elements (relations); and this in turn by the identity of their *relata* (components). So no single component could be different than it actually is. We land in modal inflexibility again. We have to find some way of individuating persons which is free from the involvement with set-(like)-identity. The doctrine of the necessity of origins seems to be a way out. What fixes the identity of the person is that the type-form is initially realized in particular matter: in these components and not others.

How should we proceed with the idea? If we take components of persons to be person-stages or realizers, we can use the ordinary physical criteria. Things are a bit more complicated when we regard persons as composed of mental items. It would be implausible to think that the person's identity is fixed by her very first experience. The prospects are much better if we allow *dispositions* to constitute persons. Then we can say that the identity of the person is fixed by the initial array of mental dispositions.[101] The identity of particular dispositions may in turn be given by the reference to their biological ground. We can thus propose a rough-and-ready general account:

Person A is identical to person B iff the initial stage of life (the initial dispositions individuated by identity of their ground) of A and B is identical & A and B have the same type form (which consists of PC, basic rational capacities and perhaps having particular type of experiences, personality and narrative).

[101] We might worry about the details: how specific the dispositions should be and whether the collection of dispositions at one time should be understood in set-theoretic terms (which entails modal inflexibility). But we could also simply say that an embryo has an initial disposition for having mental life in the future.

3.3.3. Synchronic and diachronic form

Let me now introduce an important distinction. Organization of components - a system of relations - can be *synchronic* or *diachronic.* It is natural to think of the form in terms of synchronic organization: at each point of the thing's existence, the constituents of the thing have to be related in the relevant way for the thing to exist. If the form consist of synchronic organization and properties of the whole which can exist at a time, then I will call such form a *synchronic form.* Such form is present at every point of a thing's existence. An individual form thus conceived could *endure* in the technical sense of this term.[102] Diachronic organization of components includes *inter-temporal relations* between components. If a form includes diachronic organization, it will be a *diachronic form.* Why would one use this idea? The thought goes like this. It is hard to find anything unchanging in the human mind. Rather, what characterizes persons is the orderly pattern of dependence and succession of the contents of their mental life. This kind of diachronic organization of contents constitutes the form of a person. By saying how the contents of different phases of a person's life must be related, we state both the *persistence conditions* of persons, and what the *form* of a person is. This allows us to say *what* a person is. So the diachronic form should have much appeal for the Psychological Theorist.

3.3.4. Diachronic Form PT

The thing to note about the diachronic form is that it commits one to regarding persons as temporal wholes spread in time. First, the individual form will consists of inter-temporal relations. It is a temporally extended object. If the composite thing *is* the matter and form, then it will likewise be temporally extended. Secondly, the relations: "being organized by the form" and "composing the whole" seem to be one and the same relation. Being organized by

[102] Inter-temporal relations which figure in persistence conditions of a thing with synchronic form will be though of as limitations on sorts of changes the thing can survive, and not as a part of the form.

a diachronic form - being in a network of inter-temporal relations - is not a relation that can be relativized to times. Therefore composition relation is not relativized to time. Therefore components are parts of the whole in an a-temporal way.[103]

Diachronic form accounts face similar problems as 4D ontology on account of making persons temporally extended wholes. Here too we have to say that only a part of any person exists at any given time. When it comes to consequences for agency, however, some DF accounts may do better than 4D accounts. The DF view on which persons are constituted of person-stages has, of course, the same consequences as standard 4D views. But the accounts which take persons to be composed of mental items or realizers have the option, unlike the person-stage view, to embrace a *diachronic holism* of the mental.[104] According to diachronic holism, mental states at a time cannot be individuated and have content quite independently of the earlier (and possibly later) mental life. A short-lived object either cannot have experiences, beliefs, desires etc. and cannot perform actions, or at least it cannot enjoy exactly similar mental states and actions as a long-lived person. If so, the actual part of the person is not in the position of an autonomous agent. The fact that one's current mental states are related to a temporally extended network of mental states is relevant for explaining the features of current mental states and for explaining action. Thus the existence of the whole is relevant for explaining action. Ascription of the action to the whole is justified.

Multiple Occupancy is another unwelcome consequence of the DF view. Persons in B-theoretic views are usually defined as *maximal* aggregates of suitably related components. But there is little justification for such assumption. Consider your Initial Part: all parts of you until the age of ten. The mental items belonging to your Initial Part will be suitably psychologically related, will be a part of an intelligible narative etc. This object will

[103] Person and her life has to be carefully distinguished. It is perfectly possible to say that the person's *life* must be characterized by diachronic form, but not the person herself (in this case the constraints put by the diachronic form on life translate into persistence conditions for the person).

[104] See Slors (2001) 88-110, 114-119, 189-200.

have a perfectly good psychological diachronic form. And, certainly, capacity for thought, reflection and agency will also be present. There is nothing lacking for this object to be a person. There are two further consideration supporting the claim of the Initial Part to personhood. First, there are Strawsonian Episodic themes. It is not true that in self-reflection we always view ourselves as maximal wholes. In some moods, I think that I am a person who exists from the birth to the death of this human organism. But in other moods, my self-identification reaches at most a few years back and ahead from now. Like me, many people claim to identify (sometimes) with a segment rather then a whole. Secondly, consider a possible world where you have died at ten. Certainly, you could have died then. In such world, there is a whole consisting of all and only suitably related components which in the actual world are your parts until the age of ten. But then, this whole is trans-world *identical* to your Initial Part. And this whole *is* a person. It would be extremely implausible to deny existence and personhood to your Initial Part.[105] Now, there probably are infinitely many initial parts of you, each of which is a person which co-exists with you. DF view entails infinite Multiple Occupancy.[106]

The existence of your Initial Part leads to another consequence. Consider again the possible world which is identical to the actual world up to the time when you are ten. At that time you suddenly die. Call this world W.

[105] The relation between the whole person and the initial part would be rather like the relation between the sculpture *David* and *David's* hand. If one is willing to say that David exists, I see no reason to deny that David's hand exists. It is just as legitimate an object. And David's hand does not require David's existence. If Michelangelo started with sculpting the hand and then gave up on finishing the job, then David's hand would exist while David would not. Of course, it would not be "David's hand" then (likewise, if you died at ten, the relevant object would not be "your initial part"), but the very same object would exist.

[106] This argumentation applies to 4D accounts as well.

Initial Parts to Double Identity[107]

(1) You exist in W.

(2) In the actual world, you are a maximal whole composed of psychologically related components. This whole has an initial part IP, which encompasses your parts until the age of ten.

(3) In W, you are a maximal whole composed of all and only psychologically related components which in the actual world are your parts until the age of ten. Call this whole IP*.

(4) You are trans-world identical to IP*.

(5) IP* is trans-world identical to IP.

(6) You are identical to IP. (4,5 by transitivity of identity)

(7) You are not identical to IP. (from 2)

Contradiction ((6),(7))

□

By following trans-world identity from the actual to a possible world and back, we arrive at the conclusion that you are *both* yourself *and* your initial part. This cannot be right. It seems that the only way to block this argument, is to abandon trans-world identity for counterpart theory. Yet the principal advance over 4D views which the Constitution View was supposed to offer was that on this view a person could be identical to a possible person with a history different from the actual. Despite its initial promise, DF view faces much the same problems as 4D ontology. It should be rejected for the same reasons. Admittedly, DF view could embrace diachronic holism and in this way avoid problems with agency. Yet even if diachronic holism were true (which is not obvious), this advantage of DF view is outweighed by the strongest point of 4D ontology: its ability to provide general solutions to a vast array of ontological puzzles.[108] I conclude that DF views lead to highly implausible consequences (infinite Multiple Occupancy, counterpart theory) and are probably inferior to 4D identification views.

[107] This argument (directed against 4D ontology) appeared first in Wiggins (1980), 168.

[108] This feature of 4D ontology is emphasized by Sider (2001), 206ff.

3.3.5. Synchronic Form PT

If making persons wholes extended in time is implausible, then we should turn to accounts which say that persons are composite objects existing *at* particular times.[109] The composition relation will be relativized to time: at any time, the person has some components even as it has different components at other times. What accounts for the unity of such whole over time? As far as I am aware, there are two general approaches to temporal unity in contemporary ontology. The first analyzes temporal unity in terms of spatiotemporal continuity. This approach requires the ontology of temporal parts.[110] 4D ontology has already been discussed in section 3.2. This leaves us with the second approach which explains temporal unity by persistence of components (matter) and their organization (form).

Now, it was Hume's celebrated observation that in our *conscious* mental life we see nothing which persists without change from moment to moment. From this he rightly concluded, given his assumption that there are no non-conscious mental states, that the mind does not possess the "real" or "strict" identity, but only an "identity" in a loose sense. For at any time, my mind is constituted by different components, and differently organized, than at any other time. So there is nothing which makes it the case that we have *one* mind, rather than a series of different minds.

Most of us will agree that there is no particular conscious state which persists as long as a long-lived person lasts. So let us turn to mental capacities and dispositions. Reason and character are good candidates for being stable elements of our psychology.[111] Making the persistence of one of these necessary and sufficient for existence of a person would have marked advantages. Making reason the central element would make the *essential* and

[109] Such view is consistent both with A- and B-theory of time.

[110] The point is argued convincingly by Oderberg (1993), 37-62.

[111] By "reason" I mean the system of our basic rational capacities. Although not a Psychological Theorist, Unger makes the persistence of our reason necessary and sufficient for our persistence; see Unger (1990) . The centrality of character for PT is evident in discussions of Young Socialist or Methuselah cases; see Parfit (1984), 303ff, 327ff; Lewis (1976*a*), 65f.

defining properties of persons completely coincide (only Cartesian Dualism delivers the same nice result). Moreover, considerations about continuity of the subject and anticipation suggest that only our basic mental capacities matter for our survival.[112] On the other hand, intuitions which seem to be the root motivation of the Psychological Theory have much more to do with character than with anything else.[113] Each of the option seems to promise a plausible version of PT.

If identity of dispositions determine the identity of persons then we should ask what determines the identity of dispositions in turn. The most natural answer would be: the identity of dispositions (just like any other property) is determined by the identity of their *bearer*. Whether dispositions C and D of the same type are identical depends on *whose* dispositions they are. So then, who is the bearer of, say, *my* mental dispositions? Again, the most natural answer is: I, the person, am the bearer of my mental dispositions. Who else? But if this is the right answer then we would be engaged in defining the identity of the person by reference to *this person's* dispositions. This is viciously circular. But, there always are ways to break out of the circle. First, we could distinguish primary bearers and secondary bearers. The identity of dispositions would be determined by the identity of their primary bearers. Bodies or animals or what have you would be primary bearers and persons would be secondary bearers. So identities of dispositions would not depend on identities of persons. The price for that solution are Multiple Occupancy problems. If my body was the primary bearer of rational capacities, then just how would it fail to be a person? Nay, being a primary bearer, it would have a *stronger* claim to be a person than I who am supposed to be just a secondary bearer. Let's better look for a different solution. Perhaps I could enjoy a different, but no less close relation to mental dispositions than their bearer. I could be *constituted* by my mental dispositions. Now, this seems to make a person into a complex *mode* and not a

[112] Unger (1990), 32f, 78ff.
[113] See ch. 4, 86f.

substance.[114] This looks like a plain categorial mistake.[115] But, perhaps, our categories are not sacrosanct. Still, it may be observed that the bearer of mental dispositions looks like a *subject* of them and, consequently, of the actual mental states. And then again it would have a perfect claim to be a person. Well, it seems an inherent feature of many versions of Psychological Theory that they do away with mental subjects one way or another. So perhaps we could stick to non-personal bearers providing only the metaphysical support, so to say, and to persons who would be properly said to live and enjoy mental life. The two replies made so far accept the idea that dispositions are identified *via* bearers. But it is possible to deny it. Rather, dispositions have *grounds*, and this is what determines their identity. And such grounds - presumably suitably arranged parts of my brain - would be non-personal. Persons could be constituted by dispositions.[116] Or they could be composed of the grounds of basic rational dispositions.[117]

At this point we need to note that the persistence of a particular person depends on the persistence of *particular* dispositions. A series of different instantiations of the same disposition-*type* would not do. First, there would be no stable and persisting element in the thing. One would have to appeal to intertemporal relations between different components to explain the unity of the object and this is to embrace the Diachronic Form approach. Secondly, accounts in terms of relations between different instantiations of types are always subject to branching/fusion problems. The only known solutions to the branching problems available to Psychological Theory are temporal parts ontology, and closest-continuer theory

[114] This remark applies to many versions of Diachronic Form view as well. However, if one blurred the difference between substances and *events*, and made persons consist of mental events, the present charge could be evaded. Dispositions, however, are not events.

[115] We should not be confused by the possibility of ‘person’ being just a phase sortal, or one body being able to support more than one person (either serially or simultaneously) into thinking that persons could be modes. *Personhood* and *personae*, as Johnston call instantiations of particular person-types, are modes. Persons are not. In dubious cases we should resort to constitution relation rather than say that persons are modes.

[116] Cf. Dainton (2009), 113, 232f.

[117] Unger’s position could be construed in this way; Unger (1990), 108f.

of diachronic identity.[118] We can leave 4D ontology aside. What of the closest-continuer theory? Why should we not say that the person is constituted by such-and-such components at a time and is simply able to persists under such-and-such conditions? Well, if the person were composed only of mental items, then I do not see just *what* a person would be. By assumption it would not have any persisting parts and this would make it very unlike any ordinary material complex objects. And it is by analogy to these that we seek to make the status of complex persons intelligible. Here, a person would seem something ghostily presiding over the mental states, its sole permanent property being "is constituted by". I submit this is a hardly intelligible picture. Suppose, on the other hand, that persons were constituted by physical realizers. Then, first, there are standard problems of constitution, since a person would at any time differ from the material object constituting it only by her persistence-related properties. And this being so, Multiple Occupancy looms large. Secondly, closest-continuer theories violate the intuitive "Only *x* and *y*" principle. While presumably no principle in philosophy is entirely safe, I think that the case for this one is very strong indeed.[119] In effect, the closest-continuer view would face all standard problems of constitution views *and* some serious problems of its own. This is not an attractive option.

So, to repeat, the persistence of a particular person has to depend on the persistence of particular dispositions. But, obviously, the persistence of a particular disposition depends on the persistence of its ground or bearer. The idea that a particular disposition could out-live its bearer makes no sense. A particular disposition can no more survive the demise of its bearer than a particular redness can. If so, then the persistence of the bearer of mental capacities is a necessary condition of the persistence of a person. And this bearer will either be a physical object or an immaterial object. Since in stating necessary conditions of persistence of a person we must refer to persistence of such objects (which in

[118] I take the adding of the non-branching clause to definition of identity to be equivalent to embracing a closest-continuer theory.

[119] See Oderberg (1993) , 156-163;Wiggins (2001), 96ff; Noonan (2003), 129-143, 171f, 214-227.

turn will either have specifiable physical conditions of persistence or their persistence will be primitive) it is impossible to state *purely* psychological necessary conditions of persistence of persons. It also follows that if the grounds of particular mental dispositions were our bodies, Body-Transfer would be impossible. For that would require that numerically the same disposition is grounded serially in two non-identical substances. This is absurd.[120] The person cannot outlive the ground of its mental dispositions. Still, the person could be distinguished from the ground if the ground could outlive the person. So far, we have the result that the persistence of the ground is a necessary condition of the persistence of a person. Perhaps the persistence of mental dispositions should be a necessary *and* sufficient condition. So, for example, if I fall into a persistent vegetative state, then my body continues but *I* do not survive. And if, after obliteration of my mental capacities, the brain was reconfigured to again instantiate *similar* mental capacities, they would be capacities of a different person than me. Therefore the person could not be identified with any ordinary material object. Call this a Mixed Psycho-Physical Approach.[121]

Now, I am inclined to say that by distinguishing the person from the physical bearer of mental dispositions this approach multiplies entities beyond necessity. It seems better to say that 'person' is merely a *phase* sortal. A physical substance which at some point is a person may cease to be a person, but there is nothing about this fact which would demand positing a new object. Subsequently, the same substance may again come to be a person. Now, *in a sense*, there will be two persons in such situation, given that there are two different cases of instantiation of *personhood*. We could say that there are two *personae*. But if "I" and

[120] It could be claimed that Bionic Replacement counts as body-transfer, since the organism ceases to be and I get a bionic body. In my view, however, the best thing to say about this case is that there is *a* physical object which survives throughout, and I would call this 'my body' (see ch. 5, 110f, 118-121). In any case, the *physical* continuity is preserved. If dispositions are *realized* in physical objects, then their persistence requires physical continuity. So purely psychological account of continuity is still impossible.

[121] A Mixed Approach is advocated e.g. by Garrett (1998), 56f. Note that if mental dispositions could be given *purely* physical necessary and sufficient conditions of persistence, such view would count as a version of PBA, albeit a psychology-sensitive one.

"the person" refer to the substance involved, then there is only one: the body, identical to the person, identical to me. This position seems to have good prospects of avoiding problems with constitution, cohabitation, "thinking-animals" etc. The main reason to reject this position is, I think, the belief that we are *essentially* persons. But, like Olson, I do not find this thought compelling, at least as long as personhood is defined in terms of mental *capacities.*[122] First, there is the thought that if an object has a *potential* for mentality, and then develops mentality, then it is very hard to see how it would fail to be the subject of the developed mentality (and how an entirely different subject would arise to deprive the developing thing of the entitlement to the mental capacities). Secondly, to call a vegetative state *persistent* is tendentious in the present context. It is persistent only relative to our limited technical means. Even if there are no capacities present, the body in such state - and many a corpse too - has the potential for mentality. Most of us would be inclined to regard a resuscitated person as *the* old person brought back to life (so *a fortiori* we would regard a person brought back from a vegetative state as the old person). This suggests that it is the *potential* for mentality, rather than *capacity*, which is essential to persons. Under such formulation, it is plausible to think that we are essentially persons. An embryo is a potential subject of mentality. Therefore I could have been an embryo. But a chair, however we look at it, is not a potential subject. So I could not be a chair in any possible world. All this seems right. And it is perfectly consistent with identifying a person with an ordinary physical object. Since standard arguments from vegetative states and corpse-identity against identification of persons with ordinary physical things do not seem compelling, and such identification has marked theoretical advantages (simplicity and prospects of avoiding constitution and cohabitation problems) I conclude that PBA is preferable even to a moderate Mixed Account.

We thus witness a complete collapse of Synchronic Form PT. Arguably, it should

[122] Olson (2007), 45.

collapse into straightforward PBA or substantival Dualism. If we make basic mental capacities and subjecthood central to our account of persons, the drift towards non-psychological theories seems irresistible. For persistence of such features does not depend on anything but on the persistence of the substance which develops, realizes and exercises such features. A character-based account would have better prospects of resisting the collapse. However, it can hardly claim support from the intuition that we are essentially persons since this intuition comes down to the idea that we are essentially subjects or minds; and possession of character does not seem to be relevant here. In the next chapter I will show that no sensible theory can accommodate intuitions about the centrality of character anyway.

3.4. Logical Construction Approach

The logical construction approach to personal identity is inevitably associated with the name of Derek Parfit. Indeed, this approach is often called Parfitian Reductionism. The main idea of this approach can be stated succinctly: facts about persons are nothing "over and above" facts about mental events and relations between them. But what exactly this phrase is supposed to mean is not so easy to explain.[123]

In *Reasons and Persons* Parfit claimed that this means that a complete description of reality could be given by an impersonal descriptions in which there is no reference to or quantification over persons.[124] This should make one wonder in just what sense persons are supposed to exist if a complete description of reality could omit them. And indeed, the Logical Construction view is sometimes regarded as a way of saying that persons *do not really exist.*[125] It would be a version of No-Self theory. This is not how I want to construe this

[123] On Johnston's analysis, Parfitian Reductionism should be cashed out in the idiom of constitution, Johnston (1997), 151-154. Others treat logical construction reductionism as *eliminative* reductionism; Haksar (1991) viii, 184 and Olson (2007), 130n. I want to state it as a distinct position intended to be non-eliminative.

[124] Parfit (1984), 212.

[125] See ft. 123.

view. Moreover, Parfit himself has since retracted his claim.[126] Therefore I will not start with the idea of complete impersonal description of reality.

It will be best to start with the favourite intuitions and analogies offered by the proponents of this approach. The first intuition is that there are no mental subjects unifying and individuating mental states. What there fundamentally is, are bundles of interrelated mental events. We should not read too much into the word ‘bundle’. Its use does not imply the existence of a set-like object or constituted whole or anything like that. It is, rather, a familiar term used to highlight and make vivid the *looseness* of connection between mental events composing our mental life. Not only is there no unifying mental subject. There is no metaphysical *deep unity* of mental life and persons either. So also, as Parfit put it, there is no “deep further fact” about existence and identity of persons. All these ideas are aptly illustrated by the analogy between persons and clubs (or nations).[127] Clubs are very loosely unified objects. There is nothing at the center, or underlying, or over and above individual members which unifies the club. Clubs are ontologically and conceptually dependent on their members and relations among them, but not *vice versa*. And it is rather compelling to think that there are no facts about clubs which do not come down to facts about members and their relation. If there were facts about clubs as clubs which would add something of substance to facts about members and their relations, then indeed a club would have to be some metaphysical entity hovering above the members. Well, most of us do not think it’s the case. Facts about clubs are “wholly constituted” by facts about members and relations (“wholly constituted” is an expression equivalent to “nothing over and above”). ‘Constitution’ does not denote here any peculiar interesting metaphysical relation. My analysis of what it means to say that facts about clubs are “wholly constituted” by other facts would be as follows. By

[126] Parfit (1999), 218, 221f.

[127] Many people, I among them, would say that clubs and nations do not really exists. If persons are on a par with clubs then persons do not exist. So Logical Construction is a No-Self theory. Even if this is right, it is not obviously or trivially so. It does not stand in the way of regarding Logical Construction as *intended* to provide a realist account.

referring to a given fact about a club we thereby either plurally refer to many facts about members, or else we refer to a very complex fact about members and such facts are ontologically (and perhaps conceptually) necessary and sufficient for the given fact about a club. Of course, this does not imply that the meaning of statements about clubs could be adequately translated into statements featuring members and their relations and not clubs. The second thing to note is that whether a given fact about a club is a *further* fact, or whether it is in some way *identical* to some complex fact about members is not determined by this analysis. But in any case it cannot be a deep fact. Note also a very important feature of this analysis: even if every particular *fact* about a club could be identified with some fact about members and their relations it does not follow that the *club* could be identified with any particular whole composed of members. Even if a club has these members at a time, it could have different members. To adopt a familiar concept, this analysis is consistent with *multiple realizability* of a *particular* club. Finally, questions about existence and identity of clubs may have no determinate answers. Does a club still exist when only two members are left? If the club goes dead and after few years some members reconvene, is it still the same club? Such question do not have determinate answers and it does not even seem a sensible idea to think that there should be determinate answers to such questions. Quite obviously, the answer which we want to give will depend on what *convention* we want to adopt. At this point the analogy brings us to a new idea: the existence and identity of some objects depend on our conventions. Clubs often quite overtly exist due to and by convention: their founding acts or statutes. But even if there is no such formal act, it is obvious that clubs exists only if some people regard themselves as members. To conclude, clubs are conventional and thought-dependent entities. As for clubs and their members so for persons and mental events.

All this is well and good. We should now have a fair idea about *facts* about persons. But *what* are persons? Even if we added some detailed account of persistence conditions (like the one actually developed by Parfit) to all I already said, we would still get no clue to the answer. So what are persons? This question troubled me greatly when I was reading Parfit.[128] I think I have an idea how to answer this question now. The answer is: we simply need not say. If this means that a person has no genuine intrinsic nature and has no fully specifiable individuating properties then so be it.[129] Even if demands for providing these were reasonable in cases of ordinary objects (and this is open to question), this relative lack of essence is precisely what distinguishes logical constructions from objects at the bottom level of reality. If this is right, we can make good sense of the idea that logical constructions are less real than fundamental objects, but not altogether unreal either. We can say when there is a person. We can say when a person persists. We can distinguish and individuate persons at a time. We can talk about mental events being parts of persons' lives. And that is all we really need. Certainly this is all we need for practical and everyday purposes. Now, someone could press: "But what does it *mean* to say that a person exists? Even if a particular fact about the existence of a particular person is wholly constituted or plainly identical to the fact that such-and-such suitably related mental events exist, the statement "Person A exists" cannot *mean* the same as "Such-and-such interrelated mental events exist". For if the meaning of such statements were the same, then we could adequately translate all statements about the existence of persons into statements referring to or quantifying over non-personal objects only. But then we could give a complete impersonal description of reality. And this, for all intents and purposes, means that persons do *not* exist". Here the ideas of conventionality and thought-dependence are to the rescue. An analogy to secondary

[128] It seems I was in a good company. Olson expresses similar concerns, Olson (2007), 131.

[129] To be sure, I do not claim that Logical Construction theorists do not give any essential properties and individuating properties to persons. Of course they do. But this approach neither posits a new primitive substance with a specific nature, nor identifies a person with a set or a constituted whole or anything else. There is then a clear sense in which it just does not say what a person is. To say that a person is a logical construction is not to say what a person is but, as we saw, it is to say something about personal *facts*. Furthermore, matters of trans-world identity (or deciding the closeness of counterparts) are largely undetermined. Of course, it is open to the proponents of Logical Construction to accept my idea of individuation of persons by the initial stage of life. But my suggestion is that they do not *need* to accept this or any other comparable principle.

properties may be helpful. A Lockean would say that a thing *is* red only because and in virtue of the fact that it can be *perceived as* red. To put the idea of thought-dependent existence in a nutshell: the possibility of *seeing* things as such-and such *makes* them such-and-such. And so, I suggest, it is with persons on the Logical Construction approach. The statement "There is a person" should be analysed as having some such statements as its logical consequence: "There are some suitably related mental events which can be *appropriately* regarded and conceptualized as a unit for certain (predominantly moral and practical) purposes". And I would offer the following rough analysis of what the statement "There is a person" *means:* "There are things which can be appropriately regarded as making up a rational, self-aware moral agent etc. (add any content of your favourite nominal definition of "person")". The inclusion of appropriateness into the analysis allows for oblique reference to conventions. And, certainly, conventions governing the use of concepts would have to be taken into account. So it may be that reference to the concept "person" is ineliminable from our language.[130] Now, what about persons themselves? Are they ineliminable from ontology? Well, on the present analysis the existence of a person is a *further* fact beyond the existence of mental events and their relations. But, to be sure, this is no deep metaphysical fact. Rather, it is a fact about applicability and appropriateness of some mental acts, descriptions, conventions, concepts etc. This is very much in Parfitian spirit. If there is such genuine further fact, then of course it could not be omitted in a complete description of reality. But this fact is nothing else but the fact of the existence of a person. Therefore facts mentioned in a complete description of reality would entail the existence of persons. There is no elimination of persons from our ontology.

There are two questions about the logic of my analysis. First, is it circular? The analysis is not formally circular, since we do not *use* the word "person" in analysans. We

[130] Note that I make no demand for persons themselves to have any comparable concept or actually apply it to themselves.

used nominal definition of “person” to get rid of it. But did we thereby eliminate the use of the concept person? I think so. We do not say about anything that it *is* a rational agent etc., only that some things may be regarded as such. So at most we mention the concept. Now, to understand the analysis, we should understand what it is to regard mental events as making up a person. So, if you will, the grasp of the concept “person” is pragmatically presupposed by the analysis. If this means that the analysis is in a sense circular, it is not viciously so. On the contrary, this sort of circularity ensures that there is no possibility of elimination of the concept “person” (or at least of the necessarily related mental acts and attitudes) from our thinking. The second question is this. Given that the analysis is not straightforwardly circular, does it allow us to reformulate all statements about persons into statements which do not refer to or quantify over persons? Suppose it does. How then does it help with the worry about complete impersonal description about reality? Well, I would just say that if facts about existence of persons logically follow from the facts mentioned in the complete description, this is just as good as if the persons were referred to or quantified over in the description. To protest that in spite of such logical entailment persons do not *really* exist, is to fail to grasp the idea of logical construction. Secondly, there is a way to say that complete impersonal description involves quantification over things which are persons, albeit on a nonstandard reading of “are persons”. To see how this is possible we need to consider the biggest logical challenge this approach faces.

The apparent logical form of the statement “There is a person” involves the ascription of a monadic predicate to a single thing. But our *analysans*, which is very roughly “There are things which can be regarded as persons”, ascribes a many-place predicate to a multitude of things. How is such a wild divergence of logical form between the *analysans* and *analysandum* possible? Standard examples of reduction give us one-to-one correspondence between types or tokens of reducible objects and properties and reducers.[131] Consider then the case of ‘polywater’. As Railton observes:

> … the reduction of ‘polywater’—a peculiar form of water thought to have been observed in laboratories in the 1960’s—to ordinary-water-containing-some impurities-from-improperly-washed-glassware contributed to the conclusion that there really is no such substance as polywater.[132]

This reduction was *eliminative*, as opposed to the vindicative reduction of water to H_2O. What accounts for the difference? One suggestion would be this. If we think on Kripkean lines, then some such analytical equivalence was involved: “Polywater is *the* substance which in fact caused the occurrence of these phenomena”. However, it turned out that there were may different substances: water with such-and-such impurities, water with a bit different impurities and so on. There was no *single* substance *additional* to water. This is the reason why the reduction was eliminative. The case of persons seems just the same: there is no new single object in addition to multitude of mental events. So we should regard this as the case of eliminative reduction.

My answer starts with the observation that the reason why in the case of polywater the lack of a single substance led to elimination was that the analytical equivalence included the condition that there be a single substance. Since nothing satisfied this condition, the right conclusion was the elimination of polywater. But our analysans has a different form which allows for there being no single thing with which to identify the person. We have to see how is this possible. Consider the case of reduction of redness to something physical. First, “There are red things” could be analyzed as “There are things each of which has the property F, in virtue of which things cause sensations as of red”. This would lead to type identity: R = F. What if there is no such single property? Then we would say that “There are read things” means “There are things each of which has a property in virtue of which

[131] Think of water and H_2O, secondary properties and causal dispositions/grounds thereof, mental states and functional states.

[132] Railton (1989), 161.

something is causing sensations as of red". This would lead to multiple realizability and token-identity of instances of red with instances of physical properties (or to identification of red with a second-order property of having suitable physical property). So even if there was no first-order monadic property to which redness could be reduced, this would not lead us to say that redness should be eliminated. Still, the logical form of ascribing a monadic property to an object is preserved by analysantes. Consider now a non-standard approach. This would say that "There are red things" means the same as "There are sensations as of red and there are things in the world causing them". The logical form of this analysis does not entail that the objects causing sensations have any property in virtue of which they do this. For all this sentence says, they could have no non-causal properties. This analysis does not produce the demand for identifying redness with anything in the world. So a reduction-conducive analysans can have a different logical form from the analysandum, and this makes reduction without identification possible. But how plausible is this approach when applied to the statement "There is a person"? How can the serious departure from the apparent logical form be justified? I see three possible positions. First, we could say that the apparent logical form of this sentence is misleading. What we really mean by saying that there is a person is that there is a bundle of mental states. And the logical form of a statement to the effect that there is a bundle does not posit a single substance. The deep logical form is different from the "superficial" one. Let me just say that I do not find this very convincing. Secondly, we could retract the claim that our analysis fully preserves the meaning of "There is a person". To be sure, this statement logically entails the analysans. But it also says something extra. What is this extra element? Well, probably some Cartesian intuition: that there is something really deeply unified, and perhaps existing over and above the flux of mental events. The logical form of our sentence about persons is the product of our regarding the bundles as units for certain purposes. And the extra Cartesian element is perhaps engendered by our use of the form suggesting that there is some real underlying bearer of properties. But, whatever this extra element is, it is wrong. Still, enough of what we say by the sentence is true to make us regard it as generally true. For what is true - i.e. our analysans - is sufficient for saying that there is a logical construction, an object existing by convention. This position seems close to Parfit's intentions in Reasons and Persons. The third account starts with an innocent-looking paraphrase of "There are persons". Render it as "There are things which are persons". This paraphrase is ambiguous. It can mean "There are things, each of which, individually, is a person". But it could also be read as "There are things, which, plurally, are persons". The gist of the proposal is to read "being a person" as a predicate plurally applying to many things, and not as a monadic predicate applying individually. Some examples to make this idea more familiar. First, we can say "These things were the cause of the accident". On the face of it, this sentence implies the existence of an object called "the cause". But none of the things, by itself, can be identified with the cause. And we do not think that many things together constitute a new mysterious object "the cause". Rather, jointly, they are (in plural sense) the cause. Consider now our favourite example. Imagine you attend a fair at which various clubs recruit members. You enter one tent and someone greets you by saying "Welcome! We are the local basket club". This statement is immediately intelligible. The members can truly and rightly say "We are the club". They are, jointly and plurally, the club. And if people make France, then Louis XIV was nearly literally right in saying "I am France" (assuming he was the only necessary and sufficient member of France). Because statements like "We are the club" are so easy to understand, it is also immediately understandable how "There is a club" can be equivalent to "There are some people so-and-so related". If people are analogous to clubs, we can profit again from the analogy and say that some mental events are, plurally, a person. The consequence of the plural reading is that even an impersonal description of reality would quantify over things which are persons - in the plural sense of "are".

I conclude that the Logical Construction approach is a defensible position distinct from Identification and Constitution accounts on the one hand and from Eliminativism on the other.

3.4.1. Plausibility of Logical Construction

You may wonder why I insisted on there being an analytical paraphrase of statements like "There is a person". The reason is that unless such paraphrase is provided, Logical Construction Reductionism is a non-starter. First, what we mean by 'person' would be left a mystery. Saying that a person is a logical construction is not to say what it is. Furthermore, if my understanding of logical constructions is right, one cannot grasp what 'logical construction' means without understanding that there can be an analytical equivalence of the relevant kind.133 Finally, an analytical paraphrase is required for any reduction to get off the ground.134 Consider a property F to be reduced. If F is conceptualized as a primitive property, then there is no space for reduction. Butler's dictum "Everything is what it is, and not another thing" has its rightful application in such case. So for reduction to be possible, F must be conceptualized in such a way that its intrinsic nature is open to question. It must be conceptually possible for there to be many possible stand-ins for the nature of F. But this means that there is an analytical or definitional phrase of the following sort: "F is the property s.t. φ(F)" where condition φ can be met by candidate reducers. φ will typically be a causal role, functional role, relation to our perception etc.

[133] Perry (1975), 86: "The logical constructor attempts to analyze sentences about objects of some category into sentences about objects of some other category"; Cf. Olson (2007), 130n. According to Howard Robinson (personal communication), logical construction theory properly understood claims that there is meaning equivalence between statements about persons and statements quantifying only over lower-level entities. This is consistent with Parfitian Reductionism as I construe it, but, I think, not required. The *concept* 'person' may not be wholly reducible; it is enough that there are *some* meaning equivalences required for reduction to get off the ground; and that truth-makers of any statement about persons could be fully described in impersonal terms (perhaps by an infinite disjunctive clause). This, I think, is the idea behind Parfit's claim that we could give a complete description of reality in impersonal terms.

[134] It need not be an analytical paraphrase in the reductive language. Reduction-conducive paraphrases can be stated in a meta- or higher-order-language quantifying over predicates/properties of the reductive language. Recall the example of redness we discussed, or think of the functionalist analysis of mental states.

I have little to say about the credibility of a radically reductionist proposal for reading statements like “There is a person”. The best renderings I can think of sound incredible to me. But, perhaps, one should go about it in a roundabout way. One may first come to think that personal identity consists in nothing more than in the holding of some psychological relations and that there can be no metaphysical subject over and above the stream of mental events. Then reductionist analyses of straightforward existential statements about persons may sound more credible. And if they do not sound quite convincing, this can be blamed on an extra Cartesian residue. Such approach may have some plausibility.

How does Logical Construction approach deal with the challenges besetting other forms of Psychological Theory? The problems of the “thinking-animal” type do not seem very worrisome simply because logical construction theory has no use for the idea of the underlying mental subject at all. Multiple Occupancy, on the other hand, cannot be avoided. True, one could argue that there is only one right way of defining persistence conditions for persons, and that the right account entails that persons are constructed out of a maximal set of R-related events. However, there are other consistent conventions. For some purposes we may regard mental events occurring in one day as making a “one-day person” etc. Even if such logical constructions were not persons simpliciter, they would be just as real as persons, and would have just as good a claim to being rational agents. Does Multiple Occupancy undermine the idea of agency of persons in this case? Not necessarily. It is necessary for the occurrence of many actions that a self-conception entailing long persistence be operative. But that there is such conception is more than sufficient, on our account, for the existence of a person having the self-conception. Therefore this person’s existence is, at least, a necessary condition for the occurrence of many of her actions. This observation suggests a way to distinguish between actions of long-lived and short-lived persons. When a long-life self-conception is presupposed by an action, this action should be ascribed to a long-lived person, and not to any short-lived person; and vice versa. The actions that do not presuppose any particular self-conception would be shared by all persons present. This proposal accords well with the fact that the presence or lack of a particular self-conception on a given occasion is often a reason to disown actions. Saying “I was not myself then” is a way to disown actions on such grounds and to express the feeling that the agent acting on that occasion is alien despite being closely related to one. If the proposed rule about ascriptions of actions is accepted, then we can seriously regard co-existing persons as different agents. Somewhat surprisingly, Logical Construction approach can make better sense of the agency of persons than the other two psychological approaches.

As regards the issue of trans-world identity, Logical Construction does not require counterpart theory. The argument from Initial Parts to Double Identity will not work, since the identity of logical constructions does not depend on the identity of components in the way set-like objects or constituted wholes do. Let us to try to go through the argument. I could have died at ten. Then I am identical to a possible person, call it Milosz*, the existence of which consists in the occurrence of all my mental events that occurred before I was ten. Is Milosz* identical to an actual sub-person S the existence of which consists in the occurrence of the same events? No. The convention required for regarding the events in the actual world as a unit is a different convention than that which is involved in the case of Milosz*. Milosz* is a person, and S is a sub-person or a short-lived person. They are not identical. What can be regarded as units under the same convention are events making up my whole life. Therefore Milosz* is identical to Milosz. The fact that identity of logical constructions depends crucially on conventions blocks the argument to Double Identity and allows for multiple realizability of persons. There are no obstacles to identifying oneself with possible or future persons. So far, out of all PT theories, the Logical Construction fares best. Its refutation will have to wait until ch. 8 where I discuss challenges posed by the concept of anticipation.

3.5. Concluding remarks

Given my focus on the ontological structure of persons, some of the challenges posed to PT views may not be specific to them. Some of them - notably Multiple Occupancy problems - will indeed reappear in my discussion of PBA theories. Likewise, a PBA 4D theory would face exactly the same challenges I posed against Lewisian PT accounts (and for this reason I will not discuss this option). Nevertheless, the views about the nature of personal identity do have impact on the accounts of persons' structure. The idea of Diachronic Form naturally belongs in the PT camp. Similarly, if one is willing to go for Logical Construction theory there is little reason, I think, to give much weight to physical criteria of persistence. Giving these some or even much weight is by no means precluded by my arguments; and taking Parfit's PT account as the best representative of Parfitian Reductionism is mostly a matter of convenience. But not only that. The plausibility of an overall view of persons depends also on how the accounts of the nature and structure are conjoined. Thus, I believe, Logical Construction PBA view may be one of the least plausible views overall; but Logical Construction PT is a quite formidable theory. The same goes for Constitution views: I argued that the Synchronic Form PT is less plausible than a Mixed Account, and this, in turn, than PBA views.135 But I am getting a bit ahead of myself: the overall plausibility of theories cannot be assessed just by considering how they deal with metaphysical puzzles. Their intuitive appeal has to be appraised as well, and to this task I now turn.

[135] In ch. 5 I will qualify this claim: I will argue that the Serial Realization PBA Constitution view is even less plausible than the Synchronic Form Constitution PT. But another Constitution PBA view: the One Realizer account will emerge as probably the most plausible Reductionist view overall.

Chapter 4. Psychological Theory - Intuitions

In the previous chapter I have focused on the ontological structure of persons within PT. That discussion could abstract to some extent from how particular PT accounts characterize the psychological relations binding a person together. When it comes to intuitions, however, differences between the accounts are much harder to ignore.[136] I must therefore make a disclaimer. I am going to focus on the *classical* neo-Lockean PT theories which define personal identity in terms of psychological continuity - a complex relation built on (quasi-)memory-, anticipation- and intention-links and on the persistence of beliefs, desires and character.[137] The reason is that the classical neo-Lockean approach is a direct competitor of the Consciousness-Based approach: both are fundamentally motivated by the intuition that personal identity is closely tied to the *stream of mental life* but analyze this idea differently. PT accounts which go beyond the classical approach sometimes are meant to complement it rather than supplant it.[138] If the classical approach is misguided, as I hope to show, this will affect such accounts as well. But other accounts - notably the Narrative Approach and agency-centered views[139] - may be motivated by quite different intuitions than the classical neo-Lockean approach. I think such views address a *different* sense of “identity”, “survival” and perhaps “person” than the strongly metaphysical sense that is at stake in the disputes between the classical PT, PBA and SV. I will argue that intuitions on which the classical PT is based are also largely irrelevant to the metaphysical dispute. The strategy of investigation and stratification of intuitions developed in this chapter can be applied to non-classical approaches as well. Carrying out this task is however, hardly possible, within the limits of the

[136] “Intuitions” are understood as responses indicating our *deep beliefs*, see ch. 2, 39f.

[137] Parfit (1984), 204f.

[138] See e.g. Schechtman (2001).

[139] Koorsgard (1989), Rovane (1998), Frankfurt (1999), Velleman (2001) and Velleman (2005) are good examples of the agency-centered approach.

present work.

4.1. Cases of radical psychological difference.

What makes the classical PT appealing? In my view, the distinctive attraction of this theory is due to our responses to a family of cases. These cases are: Amnesia, Conversion, Methuselah's Case, Reincarnation and Alternative Lives.[140] The cases describe situations when there is a significant psychological difference between a given person and a later or possible person that, *but for this difference,* we would not hesitate to identify with the former persons. But, given this difference, many people strongly feel that this is a different person. When the cases are considered from the first-person perspective, the response often elicited is: "It would no longer be me".

It is noteworthy that none of these cases is of science-fiction kind. With the exception of Reincarnation, they are minimally speculative, involving little or no embellishment of actual situations. This is a major difference from the other family of cases supportive of PT: the Body (and Soul) Transfer family. Partly for this reason, Body Transfer cases generate weaker and more questionable responses. In view of this, the Radical Psychological Difference cases provide the clearest and most direct link between the theory and our everyday experience and thinking. The aim of this chapter is to cut this link.

4.2. Factors of psychological differences

The psychological differences described in our cases may involve several factors. I will divide them into four groups. First, there is personal episodic memory: memory that such-and-such particular events happened in one's life. Episodic memory is to be contrasted with factual memory and how-to memory. These kinds of memories can be grouped along with

[140] The first four cases are familiar; Alternative Lives case less so. This is the case of life one would have if one were brought up very differently from the way one was brought up in fact; or the life one would have if one took a different decision on an important occasion. The question is: is the possible person who has a life so different from one the same person as one?

skills and abilities. We may call it a know-how group. Thirdly, there is the world-view. World-view is meant to encompass not only general theoretical beliefs about the nature of the world and humanity, but also beliefs about values. Finally, character. Broadly construed character includes desires, aims, tastes, quirks, mannerisms, habits and character traits as such. There are also factors which, while not internal to the mind, are important for these cases. These are one's relations with other people and the world. Call it the environment factor.

4.3. Amnesia

Classical Neo-Lockean approaches tended to focus on episodic memory as the most important factor in survival. But in our cases it is a minor factor. It is not the perception of failure of episodic memory that generates the responses seemingly along PT lines. Character mostly does the job. First, compare Amnesia with Conversion. The latter case does not involve any failure of memory. What changes is the world-view and character. And yet this case elicits responses at least as strong as any other. Secondly, consider Amnesia closely. Standardly, Amnesia cases include irreversible loss of memory.[141] Let us now stipulate that the amnesiac person exhibits exactly the tastes, habits and character traits of the pre-amnesia person, and also all her desires and aims insofar as these do not conceptually depend on specific episodic memory. So for example, the person would retain the desire to have a happy family, while she could not have the desire to care about her forgotten husband. How would we respond to this case? My response - and I believe it is a majority's response - is that without a doubt this is the same person as the one from before the amnesia. If others tell her who she is, she can in principle reassume her old way of life (although in actual cases, people may find it difficult to take up their old life). The plot of many a book and movie revolves around an amnesiac hero striving to recover his memory and identity in an alien or deceptive

[141] Standard *philosophical* cases of amnesia, not the standard *medical* cases.

environment. We take the hero's striving to be something eminently natural and understandable. We have no doubt that he is searching for *his* memory and identity. For in his place we would do just the same.[142] These considerations show that the role of episodic memory is at best secondary. Amnesia cases seem to give much importance to episodic memory. This is a factor which features prominently in the case and the lack of which seems to generate the response "It would not be me". But this is because, I suggest, one spontaneously tends to assume that one's life would have to change very much if one lost one's memory. And *this* leads to the reaction. But when we consider cases when many other factors are retained, so that the old life may be resumed, we are convinced that the person survived amnesia. Let us then turn to the case where the change of character and way of life is most prominent.

4.4. Methuselah and change

Consider Methuselah.[143] We imagine him to live for 1000 years. We should imagine Methuselah with a body of an ordinary adult for most of his life, say, from when he is 20 to 970 when he begins to age at a normal rate. His psychology, however, is not likewise immune to change. It works in much the same way as our psychology. Throughout this exceedingly long life, there will be many psychological changes: old memories, desires and interests will gradually give way to new ones; character, views and relationships of Methuselah will change accordingly. But in order to make it a distinct case, we should assume that Methuselah does not suffer from any abrupt psychological change like amnesia and conversion. Still, the 80-year old person called Methuselah, and 910-year old person called the same will differ enormously in their memories, character, views and relationships. An intuition has it that these two are different persons. Gradual changes over long time

[142] The fact that we empathize with the hero and put ourselves in his boots shows that we are not simply taken in by the way the story is framed. On the other hand, when we look *forward* to undergoing amnesia, we may doubt whether we would survive. There is a temporal asymmetry here. More on these issues below.

[143] See Lewis (1976*a*), 65f; Parfit (1984), 303ff; Noonan (2003), 107f, 119f.

achieve the same result as Amnesia and Conversion over short time. To generalize: if there is too much difference in psychology between an earlier and a later person, then these two persons are not identical. This, indeed, seems to be the foundational intuition of Psychological Theory. This is how people's psychological make-up is treated in non-philosophical contexts: as that which makes for people's "identity" and what makes them distinctive individuals. Psychological Theory derives most of its appeal from apparently taking up this insight.

Let us now examine the consequences of the idea that Methuselah's life contains several psychologically - and hence numerically - different persons. The first consequence is that psychological continuity is not sufficient for identity. The non-identical "Methuselahs" are psychologically continuous as there is no abrupt psychological change like amnesia or conversion in Methuselah's life. The lack of such landmark changes has a further effect. It makes it impossible to neatly divide Methuselah's life between individual persons. Suppose that 200-year old Methuselah (call him "M200") thinks about his past and future. He thinks "I arose some 100 years ago. I don't remember much from before that time, and these earlier Methuselah's were very different from me. I have no plans reaching further than 100 years from now. After that time so many changes will occur, that whoever exists then will no longer be me". Now, 250-year old Methuselah ("M250") will think just the same. If both are right, then they are different persons, for their lives span different times (M250 outlives M200 by 50 years). However, their lives overlap. Both of them live in the period between 150 and 300 of Methuselah's life. To make the point more general: since the changes are always gradual, we should say that for every couple of years *n* and *n+1* there will be a single person living both in *n* and *n+1*. Therefore if we make "Methuselahs" have a lifespan of at least 2 years, we will have overlap between persons. Multiple Occupancy follows. This does not commit us to saying that persons share temporal parts; they can just share ordinary parts of their mental lives like experiences and dispositions. But now the following problem arises. Do these persons share *all* their parts or just *some* of them at any given time they co-exist? Take the latter option. First, we should say that the old person has only the old dispositions. Secondly, she has only these experiences which fit her personality or arise from her character. This means that one person would have only some of the experiences which occur in this body – in spite of their being co-conscious with other experiences. Moreover, it would mean that as the old person gradually wears out - she becomes aware of fewer and fewer things; while the new person becomes aware of more and more things. This is simply ridiculous. So we should say that the persons share all their parts, all experiences and dispositions. But then, why should the old person vanish shortly afterwards? The new parts are just as much the parts of that person as the old parts. Why should the earlier parts be privileged? If the person is in an equally good shape as in the beginning, not lacking parts at any moment, then it can go on persisting just the same in the future. Since the person could persist until the given time and gain new parts, it can to persist even further in the same manner, because there is no significant difference between her persisting from t_1 to t_2 and from t_2 to t_3. The psychological theorist can finally try a mixed account. The persons should share all experiences, since otherwise we land in absurdities. But their dispositions can differ: the old person has only the old dispositions, the new person, the new ones. When enough of old dispositions vanish, the old self perishes. This account escapes absurdities of the previous accounts. But the fact that persons do not share dispositions has important consequences for the account of their actions. To count as an action of a person, the act should be suitably related to her dispositions and desires.[144] Since the two persons would differ in this respect, the new self would rightly disown the actions arising from the dispositions and desires of the old self and vice versa.

[144] Suppose that instead we define action by reference to conscious intentions. Conscious intentions would be shared. If so, the old person would come to want the same things as the new person. But this constitutes an excellent reason to say that she acquires a new character and dispositions. So she assimilates the new dispositions, and we are back to the second account.

There would be no unified agency in this case and there should be a sense of this disunity. This sometimes happen. Radical and abrupt psychological change can produce something similar to cases multiple personality. A fresh convert can have moments of relapse into the old habits, and he can disown them as belonging to his old self, not yet wholly mortified. The "old man" and the "new man" can be in conflict. But this does not always happen when psychological changes occur. Indeed, usually nothing like that happens. Our dispositions, desires and beliefs form a fairly unified functional system. This system is, to use Haksar's term, our mind in the weak sense. And such disunities as there are in this system are not (except for the aforementioned special cases) equivalent to disunities between older and newer selves. For example, certain inconsistencies and tensions are very much central to my current personality. Their removal would constitute a radical psychological change.[145] But smooth and graduate psychological changes certainly preserve the unity of our mind in the weak sense. There is no sense of disunity of agency. It is therefore altogether implausible to describe the situation in this case as involving the presence of two persons with different dispositions and desires. When changes are smooth, we should say that the old person gradually *assimilates* new traits while losing the old ones. The same person changes. The idea that there are non-identical persons within Methuselah's life must be wrong. It may help here to consider the analogy to the change of an ordinary material thing, say, a tree. We do not think that tree-change involves overlapping arboreal objects e.g. a sapling and a young tree. Rather, we think that as long as an arboreal object is not abruptly deprived of most of its parts, it can survive the loss of some amount of its components; and that it is able to assimilate new parts in their stead. And so the sapling through a gradual loss and assimilation of matter survives and becomes a young tree. Similarly for persons. If the change is gradual,

[145] The change from the "old man" to the "new man" in conversion (the paradigm case of abrupt psychological change) happens mostly along these lines. The "old man" has inconsistent dispositions and desires. The "new man" has only the "right" motivation - it is simpler than the "old man". But then such selves cannot *coexist.*

the person's psychology is not radically mutilated. What happens, is that it gradually loses some components and assimilates others. Given that the changes in Methuselah's life are of this *assimilative* nature, we should conclude that the old person manages to survive them. The analogy to the life of the tree brings forth yet another thing. We do not think that preservation of some material parts is necessary for the survival of the tree. We know that living organism undergo complete changeover of their microscopic components every few years. What counts is the continuity of life-processes. Why then should we think that sameness of some parts of psychology is necessary for survival of the person? Here too, only the continuity of mental processes should count for survival.

This conclusion has important consequences for Psychological Theory. On the face of it, the adherent of the theory could agree that Methuselah remains a single person.[146] After all, there is no breach of psychological continuity in his case. But Lewis knew better when he wrote "It is incumbent on us to make it literally true that he will be a different person after one and one-half centuries or so".[147] Gradual changes occurring for a long time produce as significant psychological differences as abrupt radical changes. And for our

[146] The third possible way of dealing with the case is to deny that our concept "person" applies to Methuselah. This is implausible. Not only does he satisfy the nominal definition. We can imagine that he would not change. He could forever retain a godly, pious and patriarchal disposition. In this case, we would not hesitate to say he is one person. A person who could change radically, but happens not to change. Finally, if we believe optimistic prognoses about human life-span, it may be possible for us to live some 200 years. Methusaleic problems may become prominent, but we would hardly cease to regard ourselves as persons for this reason.

[147] Lewis (1976*a*), 66. Lewis' treatment of the case shows how little appreciation of persons' ability to change is there in his theory. "Methuselahs" are viewed four-dimensionally as possessors of a section of Methuselah's life centered on a particular moment. This reflects the perspective of this moment - the thought that changes going beyond some point will undermine one's identity. But this ignores the sense that, as one moves on, one remains the same person who thought that thought, while assimilating new experiences and traits. And from this new moment's perspective, one's borders have already moved forward as well. If taken literally, our intuitions are plainly inconsistent. The first intuition, on which Lewis solely relies in this case, involves the lack of attention to our ability to change and to assimilate new things. It also gives no clear picture of how we would come to perish through change. As my argument shows, there is no way to plausibly fill out the picture. Lewis' account actually obscures the need to provide it - "Methuselahs"'s borders are rigidly given by *definition*. This is totally at odds with how we view ourselves. The intuition that we are able to change while persisting is incomparably more fundamental and common-sensical than the vague intuition about the amount of psychological similarity necessary for identity. So *if* we want to construe both intuitions as being about strict numerical identity, we should jettison the latter and keep the former.

intuition it is the difference which matters, not the way in which it arises.[148] So if intuitions about Amnesia and Conversion are right, the intuition about Methuselah's case is also right. They stand or fall together. As I shown, Methuselah-intuition falls. Therefore Amnesia and Conversion, the corner-stone cases of Psychological Theory, also fall.

4.5. Asymmetries. Explaining the intuitions.

Methuselah's case reveals inconsistency between the view from one point in time and the view from later points. At some point one's life seems not to stretch beyond some limit in the future. But at some later time, one identifies both with oneself at the earlier moment and with oneself at times which lie beyond the previously posited borders. Methuselah's case is not the sole case in which asymmetry can be discovered. Amnesia and Conversion involve straightforward asymmetries. The fear of far-reaching psychological change is very real. In *prospect,* such change feels a lot like death. But, interestingly, it is rarely seen as such in *retrospect.* After conversion one may feel "reborn", or simply healed, or feel that one's true self has finally emerged into light. And if the feeling of being disconnected from the previous *way of life* may be common, it is not so common to feel disconnected from the past self.[149] On the contrary, one may discover that for the first time one is in a position to tell a consistent narrative of one's life - Augustine's *Confessions* provide a famous example. When we think of Amnesia, we tend to share the backward-looking perspective of the post-amnesiac person who tries to retrieve her past "identity". Amnesia and Conversion thus involve the following asymmetry: when one views the change prospectively, it seems one will not survive it; but when it is viewed retrospectively, one thinks one is the person who

[148] Cf. Schechtman (2003), 243f.

[149] Some people may feel disconnected. My point is that it does not follow from the nature of the process. Other people undergo equivalent processes and do not feel disconnected - even though in prospect they thought of the change as equivalent to death. What Schechtman calls the 'empathic access' undoubtedly plays a major role in the feeling of disconnectedness - indeed this may be what the feeling mostly *consists* in. But then, whether the situation is construed as disconnection from *one's* old *way of life* or as disconnection from an old *self* seems to depend on the way the people *conceptualize* their situation.

survived through the change. Interestingly, an exactly inverse asymmetry surfaces in another case: the Alternative Lives. When we look back on an important past choice, we may feel strongly that "this choice made me who I am". And when we turn to consider what could happen if we did otherwise, we may feel that the "other me" who chosen otherwise is not me. But looking forward to a choice, we regard both courses of life as possible for us. One can be either of the future persons.[150]

How do we explain and resolve these asymmetries? I will take Alternative Lives as my example. Explanations will apply, mutatis mutandis, to other cases. My proposal is that the non-identificatory backward-looking intuition is not about strict numerical identity at all. It is, by and large, a correct intuition about the sort of person one would be. Past choices made me the sort of person I am. Different choices would make me into a different sort of person. The "other me" is alien to me, but this does not prevent him from being me. But alienness makes it hard for me to emotionally identify with him. Saying "this would not be me" is a short way of expressing the lack of emotional identification. On the other hand, Alternative Lives pose no obstacles to forward-looking identification. We take our possible future selves to be strictly identical to us by default.[151] This explanation makes both intuitions correct, as they concern different senses of "identity".

This may not be the only reason for asymmetry. An interpretation which makes the non-identificatory intuition more connected with numerical identity goes as follows. Perhaps in considering the counterfactual situation I see *nothing in common* between the "other me" and me. That is, I fail to see a factor which would make him me. This seems to indicate that we believe that there should be something stable and unchanging throughout the life of a person, and we fail to see that it is present in the counterfactual situation. But since the identity of the thinking substance is not in doubt here, this something stable would have

[150] Cf. Belshaw (1992), 110f.
[151] Cf. ch. 3, 49f.

to be something mental. Moreover, it would have to be something distinctive and unique: perhaps my character or perhaps a general “feel” of being me. Unfortunately, a Humean point seems to be valid here: there is no such peculiar mental feature which remains stable and unchanged throughout a human life. Since the non-identificatory intuition rests on an illusion, it is wrong. If the forward-looking identificatory intuition relies on the same illusion, it is likewise worthless. On this interpretation, while both intuitions concern numerical identity, none of them deserves to be taken seriously. Still, the idea that there should be something unchanging throughout the human life is not shown wrong. In the absence of any unchanging mental feature, this intuition supports Substantialist accounts: either Dualism or PBA. The forward-looking intuition may be therefore quite right, but for reasons inimical to Psychological Theory.

Finally, the non-identificatory intuition may implicitly entail Momentary Selves Theory. It is true to say “If I did otherwise, I would not be here (in this overall situation)”. Now, *this* fact can be felt to have a bearing on the identity of the thinker. If so, the identity of the thinker must directly depend on the present situation. And this seems possible only if one is a momentary self who, of necessity, cannot exist beyond the present situation. The forward-looking intuition, on the other hand, would involve identification with a persisting human being. This is the same as the Substantialist explanation of this forward-looking intuition. The three explanations I offered are not opposed to each other. They can apply to different people. In some cases, incompatible beliefs and tendencies may be jumbled together.

We can bring together the elements of these three tentative explanations by using the concept of emotional identification. This kind of identification - and not purely theoretical identification - is prominent in our cases. Emotional identification is based on the perception that one has something in common with the self with which one identifies. And as this something warrants emotional identification, it is not anything as abstract as numerical identity of the substance. It is something which has to do with the sort of person one is. Finally, emotional identification in our cases is present-centered. That is, one's present state is the terminus of comparison for the states of the self to-identify-with. But now consider cases when the comparison related to emotional identification is not present-centred. My identity can be determined by a certain ideal - which can be realized at some moment in the past, or be just an abstract idea. If, at present, I do not realize this ideal, then I may feel that I am somehow falling short of being fully myself. Or more mundanely: I feel that I'm not acting like myself; I feel alien to myself. In such cases, it is the present which is measured against the past (or the abstract or the eternal) and not vice-versa. If there is a difference, I dis-identify with the present me, while identifying with the past/ideal me. But then, it would be hard to find anything more absurd than the thought that I am not strictly identical to myself at present. Evidently, what we have here is emotional dis-identification which does not involve a theoretical judgment of numerical identity. I submit that the identification at play in the Psychological Difference cases is of the same kind. So we have another strong argument for the thought that our intuitions about these cases are not, at the bottom, intuitions about numerical identity.

What remains to be explained is why deep psychological changes feel like death in anticipation. For they do. And it is because they feel like death that it is so natural to think that they bring about the demise of a person. However, a change does not have to share this feature with death to feel similar. There are other features to share.

First and foremost, death means the loss of one's life. Deep psychological change means the loss of one's present life (or, if you like, the loss of one's present way of life; I am not using "life" in any technical sense here). The loss suffered at death is irreversible and takes away all possible life. The loss brought about by a deep psychological change is less radical: the present life is lost, but it is supplanted by a different life; sometimes the change may be reversible. The notion of life is at once more vague and more comprehensive than any particular psychological concept. To live a particular life is to have particular ambitions, goals and concerns; have certain memories and treat some of them as central, important and fond memories; have certain dispositions and tastes; have one's own world-view; be a master of particular skills and knowledge. But all these internal factors are less than half the story: to live a particular life is also to live in a certain kind of environment; to share one's day with these people and not others; to have one's day structured by particular tasks, activities and habits. You can, no doubt, add to this list. The vagueness and comprehensiveness of this notion makes it fit for the explanation of our reaction to cases. For if one is asked: "What is lost through an extremely radical psychological change?", the natural answer is "Everything". It is the seeming totality of the loss which makes the change seem so much like death. It is not the loss of any single factor which does. As I have shown in the case of Amnesia, if we take out one factor, but leave enough of other factors intact, then the change no longer seems to undermine identity. This is also the reason why all explanations attempted by proponents of Psychological Theory are unsatisfactory. At the bottom, we are not concerned about preserving intact any single mental factor. Who of those for whom deep psychological changes feel like death would be able to tell what it is?[152] Nor are we especially concerned about the manner of change. What we care about is to continue living a familiar life. A life which may involve innumerable departures from the present state; but which will still bear at

[152] Some people may be able to tell - those whose life is absolutely dominated by a clear ideal or desire. Their "identity" depends on it so much, that the loss of it feels like death. This is Parfit's *Nineteenth Century Russian* (Young Socialist) case described in Parfit (1984), 327. But even here, matters are less than clear-cut. Suppose that our hero loses his absolute faith and ardent desires; but instead of becoming the landowner, he still works for the socialist cause, remains in the socialist milieu and spends his days in much the same way as he used to. But this is mostly out of habit; he is half-hearted about everything. Does it now seem like he ceased to exist? My feeling is no. When we think how he feels, we will guess that he feels empty; lifeless. It will be appropriate to say that he has lost his soul; that he became zombie-like. But even if his existence is miserable or worthless, he clearly survives.

least a strong family resemblance to the present life.[153] This, I think, represents our feelings adequately.

The totality of the loss is one factor making deep psychological change similar to death. The fact that one loses the familiar life gives rise to two further similarities. First, in both cases the fear of the Unknown may be at play. Secondly, there is the feel that something *unimaginable* happens. The sense in which one's death is unimaginable is highly debatable. I suggest that it is because we cannot *make sense* of some situations, that we feel that they are unimaginable. We cannot fully accept them and integrate them into a meaningful narrative. The pictures of such situations remain sketchy and strangely detached from our overall mental picture of the world. For this reason it seems like we do not fully imagine them. Let us note, in parenthesis, that we may have much the same feeling with regard to relatively trivial changes: overcoming minor but deeply entrenched phobias or habits. Speaking in public may be unimaginable for some. One can in no way picture *oneself* coming to do these frightful things. "I will just die if I come out in public". So we say. In retrospect, such fears are laughable; but at the time, they are very real. Fear of amnesia or conversion is the same. It is generated by exaggerated attachment to one's present habits and way of life. We should not take it too seriously.[154]

Thirdly, a radical psychological change undermines the *meaning* of one's life, just like death. This is clearly visible in the Reincarnation case.[155] The lives of any two

[153] You will note that while the phrase "present life" makes perfect sense, given my notion of life, the phrase "momentary life" makes none. The "present" on which our emotional identification is centered is, usually, a very extended present. The present moment is not importantly different from yesterday or tomorrow - while the period surrounding the present is importantly different from the times lying 40 years ahead.

[154] There are good practical reasons not to take fear of radical psychological change seriously. Imagine an incarcerated mafia boss. If he believes in modern Psychological Theory, he would be crazy if he tried to earnestly undergo re-education. Nor should we attempt to re-educate him - if we are opposed to capital punishment, that is (the boss himself need not be totally opposed to death sentence, if he cares in Parfitian manner not about identity, but about there being a close successor to abide by his ideals and rule the Family). Belief that radical psychological change brings about the literal demise of the person *does* make one too conservative and inflexible. This is not merely a fancy objection. The fear of psychological changes is a strong and real feeling. But often it should be overcome.

[155] There are many different visions of reincarnation. Here I consider the most barren vision which is maximally

successive reincarnations are radically disjoint. It is hard to tell a narrative encompassing the two. One story abruptly ends and another, altogether different story ensues. Moreover, the whole series seems to be getting nowhere. There is no single overarching goal; just an endless repetition of the same schemes over and over again. If one contemplates the prospect of such meaningless series drawing on eternally, then to feel the desire for liberation - even if it meant cessation of life - is a natural response indeed. The whole series of disjoint lives seems meaningless because it lacks both a narrative unity and the unity of a goal.[156] The meaning of individual lives is also undermined by a radical psychological change, especially by the loss of memory. If the past life is altogether forgotten, past hopes, ambitions and loyalties abandoned, then all what happened turns out not to make a lasting difference. From the perspective of a later moment, it is as if the past life never existed. And this is also the threat which death poses to our lives.

We are now able to give a fair explanation of what is going on when we consider and react to cases of radical psychological changes. The feelings about one's life are basic, and they are largely appropriate and reasonable. Due to these feelings, the prospect of a radical change feels much like the prospect of death. Moreover, the alienness of the person emerging from the change makes it hard to emotionally identify with him. It seems that there is no more *meaningful* link between the old and the new person, than between two non-identical persons.[157] Such intuitions and feelings - involving no judgment of numerical identity - may be expressed by saying "It would no longer be me", where the statement is recognized as being figurative and rhetorical, expressing, more poignantly, the same thought

conducive to negative assessment.

[156] The quest for liberation could, of course, give a narrative unity to a series of lives. But the idea of and desire for liberation seem consequent upon the prior negative assessment of the cycle of reincarnation. Still, I have no intention of denying that a series of lives *can* be regarded as meaningful - for example, in the light of *boddhisattva* ideal. What I try to do is to analyze more closely the negative assessment of reincarnation, when it seems very much like death and certainly not better.

[157] The mere survival of the thinking substance - even if it is enough to ground concern about future pain - is not something meaningful in itself. If I become a living "vegetable", the substance survives, I survive, but all meaning of my life is lost.

as “It is just *as if* I died and someone else took my place”. Such feelings make it natural to think that the emerging person would literally be a numerically different person; that the change really *is* death. The thoughts we encountered in the three tentative explanations of asymmetries may also contribute. There is the thought that every person should have some distinctive mental features. A thought perfectly valid for what Johnston calls *personae* - persons *qua* possessors of a particular life in my sense. Not valid, however, for persons as such. There is the thought that there should be something stable and enduring to guarantee a person’s identity. Again, a valid thought about the substance; not valid about the contents of the mind. Finally, there is the thought, deriving no doubt from the usual present-centredness of emotional identification, that I really exist *now*, and selves existing at other times are, for this sole reason, *in some sense* different selves. Probably there are yet other thoughts on the comparable level; the level of hunches and beliefs already philosophical in nature, but not systematized and thought-through. And there is plenty of good old confusion.

4.6. Deep belief in Psychological Theory?

Psychological Theory provides a simple explanation of our responses to cases of radical psychological difference. We implicitly accept Psychological Theory - so we respond with appropriate judgments about numerical identity. This explanation is too simple to be generally true.[158] *If* we have such implicit beliefs, then they are inconsistent with deeper and more common-sensical beliefs: as I have shown with Methuselah’s case, they are inconsistent with the belief that people can survive through assimilative psychological changes. Moreover, Psychological Theory cannot provide a good explanation of the asymmetries of our intuitions about cases - that is, an explanation which would support this theory. If such intuitions rested on implicit acceptance of something like Psychological Theory then they

[158] Some people *do* believe in something like Psychological Theory prior to encountering the fully-fledged philosophical theory.

would be inconsistent, and so would bring no support to any theory. And if my explanations are right, intuitions involved in these cases do not support Psychological Theory at all. There are also good reasons to doubt whether any identifiable psychological factor could provide explanation of our feelings about the cases. On the other hand, the explanation focused on the idea of the loss of the familiar life seems to capture these feelings quite well. It turns out that our intuitions are, at the bottom, not about numerical identity at all. But these bottom-line correct intuitions may be infelicitously expressed. They may also quite naturally give rise to thoughts which are about numerical identity all right, but which are not well thought-through, vague, and inconsistent with what we think at more lucid moments. I have no simple story to tell about our intuitions. They are complex; they are messy; they may be confounded. But this picture is, I submit, much more believable than any simple explanation.

The cases of radical psychological difference are very important for Psychological Theory. They provide the initial motivation for seriously considering this theory. Secondly, these are cases with which we are most familiar in the non-philosophical world. So these are cases which have the highest chance of uncovering beliefs which underlie our attitudes - in other words, our deep beliefs. It turns out, however, that our reactions to these cases are primarily due to concern about our present life and to emotional identification, and not to beliefs about identity. The belief in Psychological Theory, if present, is an unfortunate by-product. Therefore, as far as these cases go, belief in Psychological Theory is shown not to be a deep belief of ours.

Two other families of cases remain to support Psychological Theory. There are Body-Transfer cases. These cases, are, however, most science-fiction-like of all, and notoriously inconclusive. What William’s famous discussion of these cases succeeds in showing is, I think, that our reactions to these cases may be interestingly asymmetrical and are anything but firm. This may not be true for everybody. Some people may have a firm intuition one way, and others the other way. And this, I suppose, does show that some people do implicitly believe Psychological Theory, while others believe something else. In any case, these thought-experiments do not uncover any shared deep belief. So they fail to provide a serious support to the theory. Perhaps most promise for Psychological Theory lies in Margins of Life cases: early human life and senility. These are actual cases. And many people seem to react to them in line with Psychological Theory. Many, but not all; not even a majority. In fact, the vast majority reacts to senility in a way inconsistent with Psychological Theory, firmly regarding senile people as familiar old mothers, fathers, friends; or themselves, in fearful anticipation. Moreover, the reactions favourable to Psychological Theory seem to be due to the very same factors which operate in Radical Psychological Difference cases. In short, if we treat fetuses and senile seniors specially and fail to identify with them in retrospect/prospect, it is because they bear no meaningful link to us who live a full personal life. My life as a person starts later and may finish earlier than my biological life. These thoughts are correct; and they are sufficient to justify our attitudes. Belief in PT is an idle addition. Hence, again, it is not a deep belief.

4.7. Reincarnation: an asymmetry. Sparsity of intuitions about identity.

I have argued that our responses to Radical Psychological Difference cases show little about our beliefs about numerical identity. There is, however, one asymmetry which is highly relevant. There is a powerful intuition about Reincarnation and Alternative Lives in parallel worlds. This is: I cannot say what would make the imagined man *me rather than* somebody else. It seems that we would have to appeal to a primitive simple subject to give an answer to Reinarnation: what makes it me is that the same subject persists.[159] But such subject has no distinguishing features. So it does not help at all with the question about parallel worlds. For

[159] Buddhists find Reincarnation conceivable while disposing of the idea of persistent self. But then they do not hold that anything else accounts for personal identity over time. For Buddhists, there is no such thing. So the future person is no more me than anybody else.

what would make this subject *me* rather than somebody else?[160] However, if I consider the actual world and start from my present state of consciousness, then provided that there is still a being with first-person perspective, there is no limit to changes I can imagine to happen to me, if only I make some story of how this change unfolds. Moreover, there are no limits to what we can imagine our past to be.[161]

There is thus a strange contrast between our stance with regard to parallel worlds and to the actual world. The explanation lies in the perspective: whether we adopt the first-person or the third-person perspective. From the first-person perspective, the person remains me as long as we can track potential continuity of our consciousness from a given point to the desired point. This is the only intuitive limitation. But when we consider a merely parallel world, we have no starting point for tracking consciousness, for the question is whether I exist in this world at all. So for any given point in a possible life I can ask: but is it me? Therefore we have to rely only on the third-person perspective. And then, we are altogether at a loss as to what to say and how to proceed. So, given no assumed connection with the actual me, it is a conceptually open question whether a person is me.

First-person perspective gives us a very meager intuition about ourselves: we are the locus of possibly continuous consciousness. But this is the only intuition we have which is uncontroversially about the identity of a person. Putative intuitions about the role of psychology in identity either concern preservation of familiar life or personality or express emotional dis-identification. Intuitions about re-identification of person by physical criteria can plausibly be interpreted as concerning only the usual evidence of identity. This leaves only the first-person perspective as a source of intuitions going beyond wholly general metaphysical considerations.

[160] The same goes for the body, see Maddell (1981), 23-26.
[161] Cf. ch. 6, 139ff.

4.8. Identity and self-concern

I want to finish this chapter with some comments on the link between identity and self-concern. One of the reasons that Radical Psychological Difference cases are thought to reveal our intuitions about identity is that the prospect of radical psychological change seems to affect self-concern.[162] If I think that such change will come upon me, I may feel that whatever happens afterwards is of little concern to me. The person after the change is, as far as self-concern goes, in much the same position as any persons who is not me. But how could this be if I believed he is me? There are various reasons. First of all, he will be a different *sort* of person than me; his goals and values will be different. Why should I care about such person - a person I may dislike or despise? Moreover, if this person is not meaningfully linked to me any closer than any stranger, then why should I care about him more than about a stranger? It may be said: well, he will be *you*. But regarding *that* as a "reason" for concern is really nothing else than to bring the person into the scope of self-concern.[163] And this, by assumption, is lacking and we should be giving *reasons* for feeling it. In my opinion, self-concern is a basic attitude and any thought of justifying it is ridiculous. I do not believe that we treat identity to a future self as a *reason* for being concerned. The impression that we do is created by the following procedure.[164] First, it is described how various, usually unpleasant, things will happen to *a* person. As a generally benevolent person, you feel somewhat concerned. Then you are asked to imagine that this person is *you*. Now, you feel greatly, and differently concerned. So it seems like the fact that the person is identical to you gives you a reason for being specially concerned. But this appearance is misleading. We are, simply,

[162] As well as other attitudes commonly thought to presuppose identity. I focus on self-concern because it is most straightforwardly connected with identity.

[163] By self-concern I mean concern about oneself coupled with the sense that one is, in some way, special to oneself. I do not define it as being specially concerned about oneself as opposed to others, *because* one is oneself as opposed to others. I doubt many people have *this* attitude, which I find hardly intelligible anyway. In any case, if one finds identity with one *sufficient* for being specially concerned for a future self, then *eo ipso* one has self-concern with regard to this future self.

[164] Perry (1976), 74f.

specially concerned about our futures. And this attitude is a basic one. What happens when you are told that the person in the example is you, is that you *recognize* this as a case where your special concern for your future applies. That is all. To be sure, there may be reasons for failing to be concerned about oneself: if one finds oneself despicable, hateful, worthless or one's life meaningless. Lost concern for oneself may also be restored, if one realizes that one's life and personality have value. But then, these are reasons to be concerned about *anyone*. There are no reasons for being *specially* concerned about oneself. This comes naturally, or it doesn't.

While saying all that, we are very far from the ideas encapsulated by Parfit in the phrase "identity is not what matters". For Parfit, identity with a future self is, in ordinary cases, *sufficient* for special concern about this self, because in such cases the fact of identity consist in the holding of psychological relations which really matter. Identity is not, however, necessary for special concern about a future self.[165] What I suggest is precisely the opposite. So far, there is no reason to doubt that identity is *necessary* for self-concern. But quite often it is not sufficient.

What does it mean to say that identity matters? It can mean three things. First, that the fact that a future self is identical to us gives us a reason to be specially concerned about this self. This idea is, as I suggested, mistaken. Secondly, that preserving our identity is something we are concerned about. Preserving identity means nothing else than prolonging one's existence. Obviously, many of us care about it. But even if existence as such has some intrinsic value for us, we mostly value it as a precondition of the goods of life. In this sense, identity certainly is not what matters most. Thirdly, that "identity matters" may mean that we are specially concerned about *our* future, as opposed to other people's future. We are, most of us and most often, so specially concerned about our futures. We may fail to be specially

[165] Parfit (1984), 261-264, 308-312.

concerned about some period within our future for a great number of reasons - ranging from stupidity and lack of imagination to adoption of universalized benevolence. But even if the knowledge that this is *our* future is not always sufficient to make us concerned about it, it does not follow that normally we are not concerned about ourselves. It simply shows that we must pass further conditions to become objects of our own concern. This is just common sense.[166]

The hard and interesting question which remains is this: given that the lives of radically psychologically different persons are not meaningfully linked, and that we emotionally dis-identify with the radically different persons, what meaning is left in saying "It would be *me*"? If it is just some philosophical abstraction like "the substance remains the same" that would make no practical difference, then we could doubt whether this does us any good. We may doubt whether this answer has much to do with our actual concept "person" or, in any case, with the concept we would like to have. And if it does, then perhaps, after all, personal identity is not what really matters. Now, if my arguments so far were right, then on the level of concepts and deep beliefs the link between identity and our practical and emotional attitudes is not as close as we sometimes think. We should not expect personal identity to carry all the burden of our attitudes. The fact that one would be numerically identical to a person after amnesia or conversion may turn out not to be the most interesting, important and practically relevant fact. However, the fact *may* make a difference. We should bear in mind that our intuitions about cases are not firm and univocal. On the contrary, they are full of asymmetries. If we notice these asymmetries, if we consider the perspective of the persons after change, our attitudes may change. We may become concerned. This is especially true when we consider the pain that these persons may undergo. As Williams and others noted, there is a tendency to be concerned about the suffering of the radically changed

[166] Cf. Johnston (1989), 386.

person.[167] The Young Socialist may not care about happiness of the future wealthy landowner. He may even wish him a good deal of misfortune. All this he can do with assurance and grim satisfaction. But can he remain equally adamant when he learns that the future landowner will be cruelly tortured, far beyond what he may conceivably deserve? It seems that when it comes to this, we are very far from certainty. What is so special about future pain? It is not simply, I believe, that we become much more serious and focused when we consider the prospect of pain. This may have a minor role. More importantly, fear of pain is a very simple attitude. For it to be an *appropriate* reaction to an imagined situation it is sufficient (and, I think, necessary) that something painful would happen to me in this situation. This contrasts sharply with, say, concern about happiness. A good deal more than the possibility of happiness and identity is involved in our concern about happiness: things like liking the person, thinking her worthy of happiness, her being emotionally close to us and so on. There are too many factors which may distract and mislead us. So the question is: do we experience fear of the pain inflicted on the person? Do we find it a natural and appropriate reaction? And if yes, how do we reach these reactions? It seems that we do it by trying to *anticipate* the pain. But then, it seems that anticipation and identity go hand in hand: that I can anticipate having an experience if and only if it is *my* future experience. This common-sense position will be defended in ch 8. If it is right, then the fact that a future person is identical to us makes an obvious and important difference. We can anticipate her experiences. And this, at the very least, makes fear of her suffering appropriate. It does not mean that we will automatically feel fear; nor that there is any good reason to feel fear.[168] Still, we may feel special concern about this future pain of ours and we will not be irrational

[167] So the least the Pain Avoidance Test used by Williams, Unger and others show is that there is no *universal* tendency to be unconcerned about the radically changed persons. Williams (1970) 51-57; Unger (1990), 32f, 78ff.

[168] To say that fear is appropriate is just to say that the situation is fearful; i.e. meets conditions of the concept (evil may happen to us) or that it is relevantly similar to a paradigm scenario for fear; see de Sousa (1987), 181-184, 201-203. Still, one may think that actually feeling fear may do no good, be dishonorable, not *macho,* and what have you. Or one may simply fail to feel fear.

in doing so.

Possibility of anticipation seems to be a sufficient condition of appropriateness of self-concern in general. And identity is, I will argue, necessary and sufficient condition of anticipation. Theoretical identification with a future person does not guarantee that one will anticipate her experiences nor that one will be specially concerned about them. But it guarantees the appropriateness of such acts. If one knows one survives, one may anticipate one's future experiences and become concerned about them. And this, in turn, may affect one's present attitudes, plans and decisions. Meaningful links between lives of selves are not just given; they may be chosen and created. Personal identity may be just a precondition of meaningful relations one can have to oneself. Not more, not less.

There is no place here to systematically discuss the whole issue of "what matters in survival"; to dispute Parfit's claim that "identity is not what matters in survival" or claims that self-concern can be extended to people other than oneself. But I think I have sketched how someone who rejects PT can give a reasonable answer to questions about the importance of survival of the self.

Chapter 5. Physically Based Approach - Metaphysics

It seems fair to say that, for some time now, the thinking of mainstream Reductionists has tended to converge on the idea of *embodied mind*. This is what we are: a mind realized in a physical structure. Though the spirit is anti-Dualist, there is a duality here, and a tension. What is more important for our essence and identity: what is realized or the realizer? This question needs to be asked when we are confronted with the examples showing that it is not necessary for the realizer to realize a mind and, conversely, that it is not necessary for the mind to be realized the way it is. Psychological theorists regard mental continuity as more important. Those who take the Physically Based Approach believe that it is the continuity of physical realization which is fundamental for personal identity.

5.1. Bionic Replacement and the metaphysics of subjects

The structural differences in ontology of different Physically Based theories are best highlighted by the Bionic Replacement case.[169]

Bionic Replacement. All cells of your body are slowly and gradually replaced by artificial, bionic cells. The bionic cells are functionally equivalent to organic cells with regard to a range of functions: transferring and processing sensory inputs, realizing thinking and other mental functions, transferring behaviour-directing impulses and executing behaviour. The bionic cells are not, however, alive. With regard to the brain, the replacement is done in a way which preserves both the structure and the pattern of functioning of your brain. At the end of the process, there will be a fully functional non-organic body. Will you survive the process?

[169] Unger (1990), 121f.

It is very plausible for a Reductionist to say that you survive Bionic Replacement. The train of thought leading to this conclusion may go like this.

Step 1. We could slowly and gradually replace old organic cells by new organic cells. This process actually happens. There is no doubt that a person survives the process.

Step 2. The bionic cells are functionally equivalent to organic cells in all respects which seem to matter for mentality.

Step 3. Therefore, all along the process, there occur kinds of physical events which normally sustain mental events.

Step 4. Therefore mental events occur all along the process.

Step 5. The bionic replacement could even occur without interruption of consciousness.

Step 6. Given that, and given that the process preserves physical and psychological continuity, we should conclude that mental events which occur throughout this process belong to one and the same person.

Ergo, the person survives bionic replacement.

This reasoning is open to criticism. But it does sound very reasonable. The reason it does is the Reductionist habit of focusing on mental events and causal relations among them while paying less attention to the *subject* of mentality. It is when we ask about the subject that problems surface.

The subject is the thing which performs thinking and all other mental activities. It is the substance to which mental states are ascribed. There is a strong pre-theoretical presumption in favour of saying that persons are subjects.[170] Few philosophers embracing a

[170] Shoemaker (1963), 43; Strawson (1999), 113.

Physically Based Approach would dispute this. Yet they may prefer to talk about “realizers” rather than subjects. The realizer of a mental state (or property or capacity) is the substance whose structure (and possibly other properties) realizes this mental state. ‘Realization’ is a handy word, intended to cover whatever intimate metaphysical relations is believed to hold between physical and mental states: it may be identity, it may be supervenience, perhaps it may even be some tight causal dependence. Now, if mental states are physically realized, then it seems they are states of the substance which realizes them. So realizers are subjects. Conversely, it is hard to see from the materialist perspective what else is there to being a subject of a mental state than being its realizer. Now, the problem which Bionic Replacement poses is this. The brain, the human body, the animal - these ordinary physical things do not survive the process of bionic replacement. What then, if anything, survives the process? Does the subject survive?

Serial Realization.

In response to Bionic Replacement, one can say that the person survives the process of replacement of one realizer by another. This is a Lockean answer. The difference from PT lies in the importance of physical continuity between successive realizers. With regard to the question “what is a person?” such Physically Based theories will have the same options to choose from (identification, constitution, logical construction) as PT theories. One option is to regard the person as serially constituted by organic entities (brains, bodies or animals) and their bionic replacements. We can call it “two low-level substances - one higher level substance survival” model of Bionic Replacement.

Survival of One Realizer.

According to this view, there is a physical substance which is, so to say, on the same level as brains, bodies and animals, and which survives Bionic Replacement. This substance first has biological properties and then bionic properties. The sortals under which it could be correctly subsumed would be "person" or "realizer" or "a person's body", but not any familiar natural kind term.[171] The advantage of this approach is that, intuitively, Bionic Replacement is unlike annihilation of one thing and substitution of a new thing. Rather, it seems like some straightforwardly physical thing survives the process. Call it "one substance survival" model of Bionic Replacement.[172]

No Survival. Persons as Animals or Brains.

It is possible to deny that any respectable physical thing survives Bionic Replacement. After all, brains, human bodies and animals, which are prime candidates for respectability, do not survive. If we are identical to animals or brains or human bodies, then we do not survive Bionic Replacement. The metaphysical view which distinguishes this group from the other two is this: every person can be identified with a substance which falls under a different substance sortal than 'person'. Human persons will be biological substances. Nothing bars the proponents of this approach from admitting that there may be non-biological persons: immaterial persons, AI persons etc. But no biological person can become a non-biological person and, presumably, vice versa. The denial of survival of the person in Bionic Replacement follows from this metaphysical assumption. Call it "one substance non-survival" model.

[171] Cf. Persson (2005), 37-44. Persson actually argues that we are not identical to things of *any* kind, but that we can still be said to exist. This is scarcely intelligible. Persson lands in this confusion because he identifies the body with the organism; but there are other options to consider.

[172] Depending on how one construes constitution, one could also say that persons are constituted by bodies etc.; meaning by that no more than that the person spatially coincides with her bodies. In terms introduced in ch. 3, One Realizer accounts should operate with Model 1 of constitution. Accounts using Baker-like Model 2 fall into the Serial Realization camp. See ch. 3, 55.

5.2. Serial Realization

The first dilemma facing Serial Realization is this: is the person the same subject as an ordinary physical substance which is the realizer? If the answer is yes, then the person is simply identical to the realizer. So the person does not survive Bionic Replacement, contrary to the assumption. If the answer is no, then we face the Multiple Occupancy problem. As we said, mental states seem to be states of realizers. So the realizer is a subject, and since it realizes rational thinking, it is a person. So if the person which survives replacement is different from the realizer, we have two co-existing persons. But it is hard to see how there could be a person in addition to the realizer. Where does a new subject come from? This is the problem extensively discussed by Olson under the heading “Thinking-Animal Problem”. The Thinking-Animal problem seems worse even worse for Serial Realization PBA than for PT. Psychological theorist can appeal to a dualistic intuition. Persons are, on this approach, essentially mental substances. They have different persistence conditions from ordinary physical substances. Realizers are ordinary physical substances. It seems easier to swallow that two substances of different kinds coexist and share mental states. Furthermore, by being essentially mental substances persons may have some kind of leverage over realizers. Perhaps they, and not the realizers, should be regarded as primary subjects. Such considerations seem hardly available for a physically based approach.[173] Moreover, the main motivation for PBA seems to derive from two ideas: (i) persons are ordinary physical things; (ii) I am my body. These are potent materialist intuitions. To the extent that one stresses the differences between persons and things like animals and bodies, one cuts oneself off from these intuitions. But then, what is the motivation for holding that physical continuity matters for identity? If the motivation is the wish to make a person a respectable physical thing, then this motivation is badly served by the Serial Realization account. Identifying a person with a familiar and

[173] See, however, Unger (1990), 131-134.

respectable physical thing like the animal (or body or brain) is a much better way to satisfy this motivation.

One may wish to avoid talking about subjects altogether. One may choose to talk only in terms of basic mental dispositions and capacities.[174] These can fill in for the subject, given that one often thinks that such capacities make for the essence of a mental subject. An account making mental capacities central is offered by Unger. The fundamental problem with this idea has already been exposed in ch. 3. One cannot say that *the same* disposition persists if the ground of the disposition is replaced. True, it is possible to hold that a person is constituted by a series of physically continuous realizers each of which realizes the same *type* of capacities and that it is the fact that they are suitably causally related which makes them all belong to one person.[175] One could support this view by saying: "If you look at the persistence of ordinary physical objects, this is what happens. First, you have one mass of particles (or some other stand-in for Aristotle's 'matter'). This mass is organized in a certain way - it realizes a certain form. Then some particles go away and some are taken in. You have a new mass of particles. It realizes a form of the same type as before. But the form is not numerically the same. Only the sameness of type is preserved. And that is all there is to persistence of an ordinary physical thing". This is good as it goes. But this reply leads us away from Serial Realization and straight to One Realizer account. First, we should distinguish between the organization of matter (a set of relations among particles or parts) and the properties of the composite substance (say, "being 5 feet tall"). And we should say that we have numerically one property of a composite substance persisting through the change of matter and its organization. Secondly, let us note that in the proposed answer we talked about masses of matter and not about substances in their own right, like brain or

[174] See Unger (1990), 108f.
[175] I take Unger to offer such theory, *loc. cit.*

bodies.[176] So we should say: of course, this is exactly what happens when you undergo bionic replacement. One mass of matter is replaced by another. The same type of organization is preserved. And due to this your mental trope-properties persist. But then, this means that this is a clear case of an ordinary substance persisting through change and not a case of substance-changeover. So, if there is a good metaphysical reason for thinking that physical continuity is necessary for identity, it is a good reason to think that we do not have a substance changeover in Bionic Replacement.[177]

Might we have a basic intuition that physical continuity is necessary for personal identity, regardless of its role in ensuring the persistence of the person's substance? Not at all. The belief in importance of physical continuity is derivative. It is because we believe that persons are ordinary material things that we think that physical continuity matters for personal identity. If we believed that persons are immaterial substances then, obviously, we would not think that physical continuity matters. Secondly, compare Bionic Replacement and Brain Transplant with Teletransportation. Probably, there is a difference in majority responses to these cases.[178] Very probably, there is a difference in confidence about the correctness of one's judgment. The majority confident response to first two cases is that the person would survive bionic replacement or brain-transplantation. The majority moderately confident response to Teletransportation is that a person would not survive it. Responses asserting preservation of identity in this case seem often less confident and in minority. The contrast between these cases supports PBA against Psychological Theory.[179] It makes the appeal to physical continuity criteria persuasive. But how is the contrast to be explained? A

[176] Compare Johnston (1992), 101f, on the importance of the categorial distinction between masses of particles (and similar objects) and constituted objects of common-sense type.

[177] This argumentation is also consistent with No Survival approach. From the Animalist point of view, not enough or not the right kind of organization is preserved in the process for me and my dispositions to survive.

[178] Counting only serious responses, not affected by distorting influence of the way of presentation of cases, tendency for generous interpretation of science-fiction and the like. One should also count responses to Teletransportation only of people who thought through the cases of Duplication by "Teletransportation" and Branch-Line Teletransportation.

[179] Cf. Unger (1990), 87ff.

simple and efficient explanation is this: we think that a person survives as long as we think that something is literally preserved or transported in the process. This substance is taken to be our substance; the realizer, the subject. If cases of Body-Transfer, like Teletranportation, are described in a way which presupposes, or suggests, or easily allows for thinking that something really moves from one body to another, we tend to think that the person goes along with this something (i.e. with his substance, that is, with himself). When the cases are described as a process whereby the initial substance perishes and a similar psychology is immediately instantiated in a different substance, when there is no suggestion of anything literally moving from body to body, then we have a strong tendency to think that the person does not survive. Then it seems that what happens is a mere copying. Now, on Psychological Theory the cases of Bionic Replacement, Brain Transplant and Teletransportation do not fundamentally differ. Psychological Theory does not seem to have a convincing explanation of why our responses to cases differ; and why do they differ depending on the manner of presentation.[180] This is one of the implausibilities of Psychological Theory. Now, Serial Realization PBA does not differ from Psychological Theory in this respect. According to this theory, there is no fundamental *metaphysical* difference between Bionic Replacement and Teletransportation. Both cases can be correctly described as cases where the old substance perishes, while dispositions exactly *similar* to its dispositions get to be instantiated in a new substance. But then we should say the same thing about these cases: the person does not survive. Could the difference be due to our belief that physical continuity is relevant for identity, a belief independent of beliefs about what happens with substances? No. For this

[180] There is a version of PT which says that identity is given by psychological continuity secured by *normal* causal relations. It is in fact one of the Mixed Account and could just as well be counted as one of Serial Realization views in PBA. This theory could posit a difference between Brain Transplant (and perhaps Bionic Replacement) and Teletransportation. But the appeal to normal causation is extremely implausible. It does not explain our reactions: why we can think, and with such ease, that body-swapping takes place in Prince-Cobbler case and in Star-Trek Teletransportation. If we had any thoughts that normal causation matters for identity, we would have thought that these cases are deeply odd (in the way we find Fission and Fusion cases odd and disturbing, no matter how they are described). As often happens in discussion of personal identity, the theory gives an answer congruent with our intuitions, but for entirely wrong reasons.

would not explain why we do not immediately find extraordinary cases of Body-Exchange (like Teletranportation) odd and disturbing.

Unger presents a motivation for importance of physical continuity which is different from mine. Unger also thinks that the belief that physical continuity matters is dependent on the overall materialistic world-view and that it is a derivative belief. But, according to Unger, it is not derived simply from the belief that we are ordinary material substances. Given the assumption of materialism, this belief is derived from several beliefs. First, there should be no gaps in the existence of an individual. Secondly, any process that involves an individual's surviving cannot allow for it to be accidental that there is no time-gap in the individual's history. Thirdly, for the kind of things we are (mental and conscious), it is necessary that whenever we exist, our very own particular mental capacities exist.[181] I think all this is basically right. However, these three beliefs simply derive from - or partially articulate - the belief that persons are substances, and substances of a particular sort. There should be no gaps in existence, and not accidentally so, for only then is the thing *one* substance. To allow for a changeover of the underlying substance is to compromise the metaphysical rationale for rejecting intermittent existence. In fact, Unger expressly allows for *metaphysical* dis-continuity, while holding onto physical continuity. Thus, he allows that something may be physically continuous relative to a context or a standard; while on a lower level and according to a more stringent standard, it is not continuous. Considering Zemach's example of all particles of my brain "flashing" in and out of existence at a rate of million times a second, Unger denies that it should count as a case of physical discontinuity. Yet, surely, it is a case of metaphysical discontinuity. But if the ban on intermittent existence is not derived from metaphysical considerations, then where does its force come from? Some could say that only empirical, physical facts and not any "deep" metaphysical facts matter to

[181] Unger (1990), 24f; 113-117.

us and for application of our concepts. Not Unger, however. For the beliefs from which he derives the importance of physical continuity are anything but every-day, ordinary-life, ordinary-speech beliefs. They are deep and complex metaphysical beliefs. By sliding from metaphysical continuity to mere physical continuity, Unger undermines his account. If the ban on gaps in existence concerns only gross gaps, then it is not any better supported - indeed seems equivalent to - the belief that physical continuity is necessary for persistence. If, on the other hand, it is a metaphysical principle, then first, it is not consistent with Zemachian discontinuities. Secondly, unless it is seen as deriving from beliefs about the unity of substance, it lacks justification.[182] Finally, Unger's account leaves the three beliefs disparate and independent. If instead we consider them as different facets of metaphysics of substance, we can see them to have a common ground. This increases their plausibility. To conclude, it is most plausible to think that the importance of physical continuity is ultimately rooted in the belief that the unity of substance is involved in personal identity.

With regard to the ontological structure of persons, Serial Realization PBA account does not differ from Constitution PT. On both views, successive stages or realizers of persons are primarily bound by causal relations. The difference lies in the postulated nature of causal relations. On the most influential versions of Psychological Theory, causal relations need not preserve physical continuity, while they have to preserve psychological continuity.[183] On Serial Realization PBA, physical continuity must be preserved. How does this theory fare in comparison with Psychological Theory? To my mind, it fares significantly worse. First, the fundamental ontological problems pertinent to substance changeover are the same for both theories. Multiple Occupancy problem might be worse, if only slightly, for the

[182] Oderberg's work could be used against Unger on this point. Making continuity context-sensitive could probably allow Unger to avoid Oderberg's main objections to spatiotemporal continuity accounts. But then the unjustified ban on intermittent existence is vulnerable to counterexamples presented in Oderberg (1993), 195-198. While these counterexamples may be disputed, they serve to make a valid point. We do not have a shared basic belief in the impossibility of intermittent existence. So if we uphold the ban, it is because of our metaphysical convictions about substancehood.

[183] Nozick (1981), 39; Parfit (1984), 279f, 284-287; Shoemaker (1984), 108-111.

Physically Based Approaches. Most importantly, Serial Realization undercuts the justification for adopting physical criteria of identity. I have pointed to three intuitions favouring physical criteria: (i) persons are ordinary physical things (ii) persons are not distinct from their bodies (iii) physical continuity is necessary for sameness of substance which is the thinking subject, the sameness of which is necessary for personal identity. Serial Realization goes against all the three intuitions. The only remaining motivation, as far as I can see, is the link between our practices of re-identification and physical continuity. But it is entirely plausible to treat physical continuity as nothing more than *evidence* of personal identity.[184] In the absence of independent reasons, the fact that physical continuity is used in re-identification practices does not provide any reason to think that it matters for identity. So Serial Realization PBA seems to lack justification. Moreover, if intuitions (i)-(iii) are rejected and we get clear about the limits of what follows from re-identification practices, then it starts to seem that the attachment to physical criteria is just an unthinking carry-over from familiar cases. If one wants to go for serial realization of persons, then Constitution PT is a more plausible option, given its strong intuitive appeal. But this theory is also ultimately implausible. There is no plausible Reductionist constitution view.

5.3. One Realizer

In comparison with serial realization, the thought that there is one physical realizer surviving Bionic Replacement seems a rather good idea. First, Disposition Problem does not arise. There is no denial of the idea that sameness of subject is necessary for identity. This means that at least one motivating idea about the importance of physical continuity for identity may be saved: physical continuity matters because it ensures the continuity of a physical thing which is a person. And this, in my opinion, is the most respectable metaphysical rationale for Physically Based Approach. What about the other two intuitions: that persons are ordinary

[184] See ch. 2, 36f.

physical things and that they are not distinct from their bodies? Here, things depend on what we want to say about the likes of human body and animal. Do these exist alongside the person? Suppose they do. Then we have coincident entities. One can say that some of them constitute others. Such locutions are harmless as long as it is clear that these substances are not in any way more ontologically *basic* than the person.[185] On the present account, they are all on the same level, sharing all particles of which they are composed. So back to intuitions now. As to the idea that persons are *ordinary* physical things, well, their situation is not any worse than that of statues - certainly rather ordinary things. But as for non-distinctness from the body, this intuition may seem violated. Yet, this is not really the case. First, expressions like "my body" may refer just to the matter of which we are composed.[186] In relation to this matter, I am just as close as the substance called "human body". I am not a thing separate from this matter. It is certainly true, on this view, that I am a complex physical/bodily being. And this is all that our materialist intuition really says. We should observe that we (even many of those who are committed materialists) have another important intuition: that I *have* a body; and that a body conceived just as an object among others is somehow distinct from me. Now, on the present account we have a nicely corresponding difference: a difference between a substance with personal identity conditions - me; and a substance on a par with rocks and chairs, with purely physical condition - the body which I have. So, far from violating central materialist intuitions, One Realizer theory accounts for them rather nicely. Secondly, it is possible to hold that there is just one substance - the person. Consider this piece of reasoning. What you see when you look at me is just one substance, me. If you like, you may say that I am my body and that I am an animal and so on. But, I could survive Bionic Replacement. Therefore, my body and the animal could survive Bionic Replacement. This reasoning is not, I think, any worse than the opposed Animalist reasoning. The difference lies in taking

[185] At least as long as it is clear that is the person who is the subject and the realizer of her mentality and *not* these other things.

[186] Unger (2000), 286-288.

different properties to be essential. Do we have to think that biological properties are essential for animals? Unger argues briefly, but convincingly, against this thought.[187] What about "my body" then? Well, if one can say that an animal can survive bionic replacement, one should not find it problematic that my body can survive it.

Of the metaphysical problems, Multiple Occupancy is the most threatening to this account. It can be avoided if one denied existence to such sub-personal entities as animals, human bodies and brains or equated them with the person. This seems rather plausible to me. But suppose one prefers to hold that persons co-exist with animals etc. How bad is the problem of Multiple Occupancy then? Not worse than for other theories - and at this point it seems that *every* Reductionist theory will have this problem. It is open to a One Realizer theorist to adapt the standard response of PT to the Thinking Animal problem, i.e. to distinguish between *essential primary* bearers of mentality (persons) and *non-essential derivative* bearers of mentality (animals); perhaps adding that the latter do not have psychological properties or that they have them in a merely derivative way. This requires holding that we are *essentially* mental beings.[188] This strategy may even work better for One Realizer PBA than for Psychological Theory. D. Hershenov argues that even if a Psychological Theorist can claim that the animal does not think, he cannot deny the existence of an essentially *sentient* being which, however, does not have higher mental capacities (like the capacity for self-consciousness) essentially.[189] Such sentient being is not identical to the person, but it is able to think. In contrast, the One Realizer theorist may hold that only some minimal mental capacities are essential for persons. So it will be possible to claim that the sentient being is identical to the person and so Hershenov's version of Too Many Thinkers problem does not arise. MacMahan presents another interesting solution to the problem: we

[187] Unger (2000), 288-290.
[188] Unger (1990), 109 and (2000), 282.
[189] Hershenov (2006), 233f.

are separable *parts* of an organism.[190] The organism therefore can be said to think in a derivative sense: in virtue of having a part which thinks. It is clear that on this account an animal is not really a thinker in its own right. This response makes a good sense to me.[191] To conclude, One Realizer PBA has ways to confront the Multiple Occupancy problems which are not worse or better than those available to other accounts. Given that One Realizer account is free from many other serious problems, that it can recognize that persons are ordinary real substances and does not clash with (majority) intuitions concerning some central cases, it should be recognized as the best Reductionist account so far.[192]

5.4. Animalism

Proponents of the previously discussed approaches take the lesson from Bionic Replacement to be that persons cannot be identified with any thing defined in *biological* terms. But, while it is natural for a Reductionist to think that we could survive Bionic Replacement, it is also possible to deny it. Today, the most notorious No Survival view asserting that our essence is biological is Animalism. According to this theory, human persons are identical to human animals: members of the *Homo sapiens* species.[193]

There are three reasons why Animalism is implausible. First, there is the case of Dicephalic Twins.[194] Fortunately, this is an *actual* case. It shows our actual practices of

[190] If this reply is to be consistent with the One Realizer response to Bionic Replacement, then one should say that this part is a realizer or person (and, presumably, is made of the same matter which is commonly regarded as making up a brain); or, if one says that this part is the brain then hold that the brain is identical to a person and could survive Bionic Replacement. It would not do to say that this part is *constituted* by the brain, for then, first, it is dubious if the person (as opposed to the brain) should be counted as a part of an animal; second, one winds up with the Thinking Brain problem.

[191] In consequence, I think that the Thinking Brain problem is harder for the Animalist, than the Thinking Animal problem is for people like Nagel, Unger and McMahan. It is curious that in his new book Olson largely *ignores* McMahan's anti-Animalist argumentation.

[192] Curiously enough, it is not easy to find a clear actual example of the One Realizer PBA view. Johnston seems to embrace this view, given his inclination to accept personal survival in the Bionic Replacement case (Johnston 1997, 263). Shoemaker in his recent writings proposes a One Realizer view; although of course this is stil a PT, or rather a Mixed Approach account; Shoemaker (2004) and Shoemaker (2008). The borders between PT and PBA may be blurred once there is an agreement about the idea of the *embodied mind*. More than any other views, One Realizer accounts are located in the borderlands.

[193] Wiggings, van Inwagen, and Olson are mong the best-known proponents of this theory.

[194] McMahan (2002), 35-39.

identifying persons and adopting practical attitudes towards them. The vast majority response to this case is that there are two distinct persons inhabiting a single organism. It seems clear that what underlies this verdict is the commitment to a principle of individuating persons - either “one mind - one person” principle or “one consciousness - one person” principle (both principles yield the same verdict in this case). So the case provides a straightforward counterexample to Animalism and supports a principle of individuating persons incompatible with Animalism. This, admittedly, is a strong claim. But let me put it this way: if the actual Dicephalic Twins case does not show the way we individuate people then *what else could show anything*? Of course, we could say that one should consider the whole range of actual cases rather than an isolated case to adequately grasp the way we apply the concept “person”. Fair enough; but I do not know any actual case which would generate *equally* strong and univocal responses inconsistent with the responses to the Dicephalic Twins case. There are cases of Multiple Personality Disorder and Commissurotomy. But, first, these are not yet even fully understood by science. Secondly, even if we knew all the externally observable facts, it may still be open to question whether we should say that there are two minds (or two centres of consciousness) or one but less unified than in standard cases. Thirdly, even if we agreed that there were really many minds or many centres of consciousness involved, then great many people would say that *precisely for this reason* there are many persons involved.[195] Thus the cases of Multiple Personality Disorder and Commissurotomy in no way provide a counterbalance to the Dicephalic Twins case. On the contrary, the non-Animalist principles of individuation give the best explanation of our responses to a wide range of cases. There are two ways in which the Animalist could still defend his view. First, he could point out that other Reductionist views also allow for the person to have at least partially functionally dis-unified mind and wholly dis-unified consciousness (think of Parfit’s

[195] For philosophically informed discussion of MPD and other dissociative states see e.g. Humphrey and Dennett (1989); Wilkes (1988), 100-131; Braude (1991); Hilgard (1991) and Radden (1999).

My Physics Exam case). But if that is acceptable, then why having wholly or almost wholly functionally dis-unified mind should be impossible for a person? If Animalism were for independent reasons much more plausible than other Reductionist alternatives, then although its response to Dicephalic Twins case is discomfiting, this should not count as a decisive reason to reject it. This may be fair as an *ad hominem* argument against some Reductionist; but it would save Animalism only if there were no better alternatives on offer. This, as I argue, is not the case: apart from more plausible Reductonist theories there is even more plausible Consciousness-Based SV. The other way to defend Animalism is to accommodate our response to the Dicephalic Twins case. The Animalist would have to claim that in this case there are two distinct organisms sharing most of their parts. On what grounds would we count here two organisms? The only reason seems to be that we have two *brains* in this case. If so, the Animalist has to make the brain the central and essential part of the organism; sufficient for its individuation and presumably necessary and sufficient for its persistence.[196] Stated this way, Animalism only marginally differs from the Brain Theory.[197]

The second reason to reject Animalism is purely metaphysical. For various reasons (some having to do with the Thinking Brain and Homunculism problems discussed below), the Animalists cannot avoid claiming a special metaphysical status to *living* things. The claim is nothing less than this: the only composite things which *really* exists are living things. Only life generates the kind of unity among parts which makes it possible to say that they compose something.[198] This position seems altogether implausible. First, there are problems of vagueness. The notion of life is, indeed, singularly vague. Now, what Olson calls "biological minimalism" is supposed to contrast favourably with compositional universalism: i.e. the view that *any* collection of objects composes *something*. According to biological minimalism, there is a deep metaphysical difference between the right arrangement of things

[196] See Johnston (1997), 263; van Inwagen (1990), 169-181.
[197] Cf. Olson (1997), 140ff.
[198] Olson (2007), 226.

(in a living being) and any old arrangement. If there were no such deep differences, and it was just a matter of how we can "cut up" the world, then it is hard to see a reason for restricting ourselves to sparse ontology.[199] But, if so, there has to be a sharp divide between the situation where there is a living organism and a situation when we do not have a living organism.[200] But this, of course, is not the case; both because our notion of life is singularly vague and, more importantly, because our world is so terribly gradual.[201] Secondly, biological minimalism entails that there is a deep and sharp difference between living things (really existing) and artifacts (not really existing). This is unbelievable. In fact, the very notion of life seems to be a largely *functional* notion. The difference between artifacts and living things seem to be merely a matter of *degree* of complication. And, here too, there is no sharp divide to be seen.[202]

Thirdly, Animalism faces the Thinking Brain problem. It seems that the brain is the organ of thought. If it thinks, then we have two co-existing thinkers: the brain and the

[199] If it were just the matter of how we want to cut up the world, it would depend on our interests how to do it. And while precision, sparsity and elegance are virtues, other considerations can outweigh them. Of course, Olson can say that unrestricted Multiple Occupancy entailed by compositional universalism is unpalatable. But it really is so only if we take metaphysics of substances and real unities seriously. I see nothing troubling in the idea of thousands of logical constructions coexisting. Parenthetically, this only shows how utterly outlandish is the idea that persons are logical constructions - Multiple Occupancy *is* unpalatable which shows that we *do* deeply believe that we are substances in a robust metaphysical sense. It is a great virtue of Parfit's treatment of these issues that he acknowledges that we have such belief.

[200] Compare what Olson says about some composite non-organisms: "There is something at least a little bit peculiar about undetached brains and heads - not to mention upper halves and left-hand complements. Their boundaries are to some extent arbitrarily drawn", Olson (2007), 219. This seems to imply that the boundaries of organisms are *not* arbitrarily drawn. But this is just incredible. Perhaps in some sense they are *less* arbitrary - but it is hard to see how mere degree of arbitrariness can matter for metaphysics. Perhaps, taking Johnstonian line we could say that the boundaries of organisms are not arbitrary from the point of view of our *concepts*, even though they are arbitrary in relation to reality. But this anti-metaphysical line is surely not going to help such metaphysicians as van Inwagen and Olson.

[201] Hence, even if we could remove the vagueness from our concepts, we could only do it by making a wholly *arbitrary* cut.

[202] Olson thinks that he could adopt the account of composition in terms of functional unity to his Animalist theory (Olson 2007, 227f). However, he thinks so because he believes that only the whole organism would count as a functional unity. Its parts would not, and that would help with the Thinking Brain problem. But Olson comes to believe this on the basis of a very bad argument. He says that since particles composing a part (say a brain) would belong to a larger set of functionally united particles (composing the animal), they would not compose anything. But, obviously, a whole organism can also be functionally united with other organisms: in a team, or a collective Chinese Room etc. By the parity of reasoning, the parts - organisms - would not exist. This is absurd. The functional unity view is of course *consistent* with Animalism; but it leaves Animalism unjustified. For the functional unity theory of composition see Hoffman and Rosenkrantz (1997), 80-90, 128-134.

animal. On what grounds can we say that it is not the brain but the whole animal which thinks? Of these two, the brain seems to have a stronger claim to being the thinker. Now, this claim rests on the possibility of making a sharp distinction between mental events - standardly taken to be located in the brain - and their external causes.[203] The plausibility of this assumption can be questioned. Olson discusses these issues under the heading of "thinking-subject minimalism". He formulates the thesis of thinking-subject minimalism in the following way:

(TSM) If *x* thinks in the strictest sense at time *t*, then *y* is a part, at *t*, of *x* if and only if *y* is directly involved in *x*'s thinking at *t*.[204]

Olson offers three considerations against thinking-subject minimalism and identification of the subject with the brain. First, specific *parts* of the brain are directly involved in different *specific* mental activities. So it seems that we have many narrowly specialized subjects in the brain - many homunculi. Secondly, it is impossible to defend brain-centered views from the danger of Homunculism by defining unity of mind in terms of functional unity and insisting that there is a necessary correlation between the number of minds and the number of thinkers. For, Olson claims, for this "psychological individuation principle" to be true, we would have to be something entirely composed of mental events. We would not be brains. So this principle is not available to brain-centered views. Thirdly, Olson claims that it is impossible to draw a clear boundary between what is directly involved in mental activity and mere external causes.[205]

It should be noted that Olson's claim relating "psychological individuation principle" and mental-bundle theory is false. To give an example, Shoemaker's Constitution

[203] Schechtman (1997), 154, 161.
[204] Olson (2007), 88
[205] In this Olson follows Schechtman (1997) and Clark and Chalmers (1998).

View PT is committed to the principle, but it certainly does not say that persons are mental bundles: they are constituted by realizers of a functional system - the mind.[206] Olson asks how “causal relations among mental states and actions could entail both the existence and the precise number of things that are not made up even partly of mental states or actions”.[207] Now, for the functionalist it is not the *actual* causal relations but *potential* causal relations which matter. The program need not be actually run to exist. It determines how things go *if* relevant inputs are provided. There is also no reason why the functionalist should focus on particular mental states rather than on general mental capacities. So, thinking along such lines, we may offer some such functionalist definition of a thinker:

(i) *x* is a thinker iff *x* is *structured* in such a way that *x*’s structure realizes a functionally unified mind

which means that the thinker is so structured as to have a *capacity* to form beliefs, desires etc. all of which *would* be functionally unified (i.e. they would have a capacity to relevantly interact). But then, of course, a thinker *can* be such a suitably structured material being. There is nothing hard to understand in the idea that a material being persists as long as it maintains the structure necessary for realizing a *capacity* for mental activity. From here it is but a short step to the solution of the Homunculism problem and to something akin to (but importantly different from) Olson’s thinking-subject minimalism. We can say:

(ii) For any *x* and *y*, *x* and *y* are parts of one *thinker K* at *t* if and only if (*if x* were directly involved in a mental act *s* at *t* and *y* were directly involved in a mental act *z* at *t* then *z* and *s* would be functionally unified)

[206] Shoemaker responds to Olson (on different grounds than I give here) in Shoemaker (2008), 320ff.
[207] Olson (2007), 138.

A thinker can be defined as follows:

(iii) *T* is a thinker if and only if (*T* has at least one part directly involved in thinking & all *T's* parts satisfy the right-hand condition of (ii) & there is no *x* s.t. *x* is not a part of *T* and there is a part *y* of *T* s.t. *x* and *y* satisfy the right-hand condition of (ii))

Olson is wrong when he proposes that thinking-subject minimalism defined in terms of actual direct involvement in thinking is our reason for holding brain-centered views. I believe that when we entertain brain-centered views, we think along the functionalist lines of (iii), which is incompatible with thinking-subject minimalism as defined by Olson. But I also think that, quite independently of any thought about brains, we do believe in thinking-subject minimalism. Quite simply: the subject is the thing which does the thinking. Certainly, the subject is not something which only indirectly or derivatively thinks. The subject is not like a committee which can be said to "think" because its members think.[208] But if the specific acts of thinking are performed by different parts of the brain then the brain itself - the putative thinker - is precisely in the position of the committee *vis a vis* its members who really do the job. If we take the link between direct involvement and subjecthood seriously, then we have to conclude that there is no *single* subject of our mental life.[209] Thus, as Dualists claim, standard Materialism is unable to account for the unity of the subject.[210] Nor is functionalism going to be of much help here. It provides a valid *formal* solution to the problem of Homunculism. But it does so by abandoning thinking-subject minimalism. It primarily explains the unity of *mental life*. But this explains only the unity of *what is produced* (or

[208] Cf. Unger (2005), 72.

[209] It is incredible to think that the *whole* brain - every part of it - is directly involved in a specific mental act; Olson (2007), 91.

[210] The argumentation goes back to Plotinus; Emilsson (1988), 101-105.

realized) and not the unity of t*he thing which does it*. And it is obviously fallacious to infer from the unity of the product that there is one thing which produced it.[211] If this is right, then to arrive at a satisfactory metaphysical solution to Homunculism problem, one either has to embrace simplicity of the self, or reject thinking-subject minimalism.[212] Taking the latter option, one can either deny the "subject" part or the "minimalist" part. The former is not really an option for Physically Based Approaches. We assume that there is nothing more to the idea of the subject than the idea of the "realizer". And, certainly, one wants to be able to speak of realization if one is to start debating the Psychological Theory. Secondly, without the appeal to the sameness of realizer-substance, the Physically Based Accounts are inferior to Psychological Theory.[213] So, if one agrees that Homunculism is a problem, then one has to reject *minimalism*. What are the consequences? In order to reject thinking-subject minimalism, one has to argue that there is no way to draw a clear distinction between direct involvement (or realization) and mere causation. The arguments for this idea are actually quite convincing. Now let us ask what then *makes it true* that something has mental properties. Olson offers a passing remark: "The organism has a nonarbitrary boundary, and it would appear to be the largest thing *whose behavior we can explain in terms of its thinking*. Though there may be a real sense in which thinking is something an organism does, there seems to be no real sense in which thinking is something a brain does".[214] If *explicability* in

[211] The same point holds for activity. Four people can together carry a coffin. From the fact that there is one "carrying of the coffin" it does not follow that there is a single thing which does it (non-derivatively). Perhaps it could be claimed that different specific acts are activations of a single *capacity* of the brain. But in fact the reason to postulate a single capacity is that the products would be functionally unified. If only the parts are directly involved in thinking, then only *their* capacities (e.g. the capacity to support thoughts functionally related to thoughts supported by other thoughts) are directly involved. Capacities ascribed to the whole are something derivative. So the appeal to capacities does not solve the problem of thinking parts.

[212] Simplicity of the self may be directly derived from thinking-subject minimalism. First, one assumes that one is a subject of several different mental acts simultaneously. Since it cannot be the case that one's different parts are involved in different acts, one is either simple or all of one's parts are directly involved in each mental act. Then one has to claim that divisibility entails that whatever is a joint activity of several parts has to be analyzed as a *derivative* activity parasitic upon separate activities of each part. But the subject's mental activity is not derivative. Therefore one's mental activity cannot be a joint activity of many parts. Therefore one is a simple substance.

[213] See above: 117f.

[214] Olson (2007), 93. Italics mine. Olson does not say explicitly what is the "real sense" in which the organism

terms of thinking is what makes ascriptions of acts of thinking *true,* then we have here an unmistakably Instrumentalist position with regard to mental states. It is no accident that Olson arrives at such position. If there is no way to draw a clear boundary between what realizes and what causes mental states *within* the body, then there is little reason to draw a line at the boundary of the human body. But then, either it is impossible to draw *any* boundary of what realizes mental states, or the boundary is drawn somewhere in the environment. Consider the latter option. Particular mental states will be realized by the states of constantly changing *configurations* of human bodies and objects in their environment. This is an Extended Self view. It can be stated as a Constitution View. The self could be defined in a way structurally similar to the Animalist definition: as the whole composed of everything that is caught up in a single *mental* life.[215] This view is incompatible with Animalism. Consider then the other option: there is *no* boundary of what realizes mental states. This means that ideas of direct involvement and realization have to be abandoned (unless one wanted to say that the whole Universe realizes all mental states). What account of the relation of the mental and the physical can one give then? I see only two Materialist options. One is Emergentism. However, if mental properties emerge from the complex relations of the animal and the environment, is there a reason to ascribe mental states to the animal *rather than* to configurations of animal bodies and other objects? None that I can see. To save the unity of

thinks and the brain does not. The context seems to indicate that “being the largest thing whose behaviour we can explain in terms of its thinking” is relevant (though, in fact, being the *largest* thing cannot be relevant - recall the example of the collective Chinese room). Perhaps Olson is saying only that if thinking can be ascribed to the brain, then with at least equal justification it can be ascribed to the organism. On this reading, Olson does not address my question about what makes the ascriptions of mental properties true. There is some justification for this reading, for when Olson returns to the problem on pages 215-219, he says that up to that point he has not ruled out the possibility that both brains and organisms think. Two of the solutions he proposes do not look any better than Psychological Theorists’ solutions to Thinking Animal problem. The third proposal - denying the existence of brains - would neatly solve the problem. But this solution relies on the untenable Animalist theory of composition. So on the cautious reading Olson is left without a good solution to Thinking Brain problem. Secondly, the question about truth-makers of ascriptions of mental properties has to be answered anyway. I argue below that Olson is left with the choice is between Extended Self, Emergentism and Instrumentalism.

[215] Extended Self views most probably fall into One Realizer PBA group. They could also belong to Constitution PT.

the subject one can postulate that emergent mental states are had by an Emergent Self.[216] But, obviously, an emergent self is not going to be identical to an animal. Hence, Emergentism is not an option for the Animalist. The only other option is Instrumentalism. This, however, is but a version of Eliminativism about mental states. Now, if one goes Eliminativist about mental states, there seems to be precious little meaning in holding onto the idea of a persisting self. Eliminativism is a high price to pay for Animalism. We should be willing to pay it only if all alternatives to Eliminativism and Animalism are shown to fare much worse than this couple. But this, fortunately, seems far from being true. Finally, the main appeal of Animalism lies in its apparent consonance with our "common-sense" and ordinary ways of thinking and speaking. It is hard to find anything more at odds with these than Eliminativism. Embracing it would undercut the main motivation for entertaining Animalism. Olson avoids these dilemmas because, given his views about composition, organisms are the only candidates for thinking beings that really exist. But this, as we saw, is hopelessly implausible.

To sum up, there are three good reasons to reject Animalism. First, there is a serious counterexample to this theory - the actual Dicephalic Twins case. Secondly, Animalism relies on an implausible theory of composition. Thirdly, it cannot satisfactorily deal with the Thinking Brain problem. Animalism either has no better solution to its version of Multiple Occupancy problem than any other theory; or it provides a solution only at the price of going Eliminativist about mental states.

5.5. The Brain View

The second theory in the No Survival camp identifies human persons with brains or, better, parts of brains.[217] Another theory has it that brains (or specific parts thereof) are the core of the person, while the person also has other bodily part. On this theory, a person is a body

[216] See Lowe (1996), 46-51; Baker (2000), 17.

[217] Nagel (1986), 30, 40-51; McMahan (2002), 68, 92. McMahan is actually not committed to identifying the person with a part of the brain. He only holds that the existence and functioning of some parts of the brain are necessary for personal identity.

controlled by the brain which is her central part, the survival of which is sufficient and necessary for the person's survival.[218] All these theories may be regarded as variants of what we will call the Brain View. The differences between these variants will not concern us here, so for simplicity I will talk of the view that persons are brains.

The Brain View does not encounter any counterexample comparable to Dicephalic Twins; nor does it seem to require any idiosyncratic theory of composition. But I will argue that it cannot explain the unity of the subject. The only reason to identify a person with a brain is that it is the brain which seems to support a *unified* mind. It is not just that a brain is a physically unified object which, some way or other, supports mental life. For this holds good of an organism, even in the case of Dicephalic Twins. Still, we count the twins as two separate persons because we are able to clearly distinguish two separate and independent minds. Therefore, the brain's capacity to support a unified mind is essential to its identification with the person. Now, the brain is a complex object. Its capacity to support unified mentality depends on (and seems derivative from) its parts' capacities to realize mental states unified with states realized by other parts. However, the parts also seem to have the capacity to realize *disunified* mental states. The common assumption is that each hemisphere could support mentality and consciousness independently of the other. This is shown by the fact that people survive massive strokes amounting to destruction of one hemisphere. Now imagine that instead of doing anything as brutal as destruction or extraction of one hemisphere, we are able to insert millions of Nano-Robots into the person's skull. We will have these nano-robots cut off all direct communication between hemispheres, but without impinging in any way on the physical unity of the brain.[219] Suppose the nano-robots

[218] Johnston (1997), 263

[219] The halves of the brain may still communicate indirectly, since they are in the same body (so one half can immediately observe the effects of what the other is doing); they may also be interdependent in virtue of being in the same *chemical* environment. I do not think this is enough to generate a genuinely functionally unified mind (as opposed to two minds which are coordinated). To clear all doubts, we should to imagine Fission with transplants.

just regulate the level of relevant neurotransmitters in the synapses at crucial junctures so as to prevent communication. Call this case *Nanorobotic Commissurotomy*. What would be the upshot of this operation? It seems most plausible to think that the two hemispheres would support two *separate* minds, despite being physically unified (just like the brains of Dicephalic Twins do). Let us now ask what happens with *mine*, i.e. the whole brain's, capacity to think. Well, it seems that in this situation the parts of the brain *cannot* support a unified mind. Therefore they do not have the capacity to do so. Therefore the brain does not have its capacity for supporting a unified mind.[220] But this capacity is identical to *my capacity to think*. Why? Well, if I am the brain, then my capacity to think has to be the brain's capacity to think. But if the parts of the brain come to support independent minds, there is absolutely no reason to ascribe thinking to the whole brain, rather than to the individual parts. If two hemispheres supported two separate minds from the beginning, the situation would be essentially the same as with Dicephalic Twins. We should say that in such case there are two separate minds and two persons, each identical to a hemisphere supporting its unified mind. But, certainly, the fact that separate minds *arose* at some point because of our nano-operation is irrelevant to our conventions for counting minds. So the post-operation situation would be this: there are two minds, each supported by a hemisphere. Thinking and capacity to think are ascribed to the hemispheres and not to the whole brain. Therefore *my* (the brain's) capacity to think no longer exists. But *my capacity to think is my essential property*. Therefore I do not survive the operation. Nor is there any reason to identify *anyone* with the whole brain. For the reasons we used to identify me with the brain now indicate that we should identify the two post-operation persons with the hemispheres. But while I do not survive the operation, and after the operation there is no person identical to the brain, the

[220] It could be said that the brain still has the *potential* for supporting a unified mind. But if this were relevant, we could count *any* two brains as one person, since they have a potential to fuse into something supporting a unified mind. Even if fusion were for some reasons physically impossible, this certainly is not the *reason* why we count persons as separate.

brain clearly survives the operation. Therefore neither I nor any other person is identical to a brain.

Having the capacity to think essentially and being identical to a brain are incompatible properties. Anything which can be defined in terms of having the capacity to think will not be identical to a brain. Olson is thus right in suggesting that if MacMahan has any view as to what we are, he has to think that we are something like a *functioning* brain.[221] The functioning brain will not be identical to the ordinary brain, but will presumably be something constituted by the brain. But if the person is only constituted by the brain, then I see no good reason to reject the survivalist solution to Bionic Replacement. MacMahan's reasons may have been this. In line with what he says about the organism (p. 92), MacMahan would presumably say that the capacity to think is really the brain's capacity. But then we cannot have the brain replaced and have the same person (because of the familiar Disposition Problem). However, if my capacity to think is my brain's capacity to think, then I am my brain. If so, we should ditch the assumption that I have this capacity essentially, for it is not true of the brain. Yet without this assumption MacMahan's position looses much of its appeal. To conclude, one cannot have an account which (i) satisfies the intuition that the capacity to think is essential to persons (ii) avoids the problems of constitution and (iii) identifies persons with ordinary objects which seem to have the capacity to think (animals or brains). MacMahan could only seem to pull it off by leaving the notion of the "functional brain' unanalyzed.

If the Brain View cannot give us all we want, how plausible is it? A Constitution Brain View seems unattractive. If persons are characterized in terms of constitution, they seem perfectly capable of surviving Bionic Replacement. Consider then Identity Brain View.

[221] Olson (2007), 86. MacMahan (2002) is noncommittal with regard to what we are on p. 88; seems to suggest that we could be regarded as identical to the brain or a complex made up of the brain and the mind on pp. 92-94; but on the p. 92 he also talks of us being "functional brains". It is safest to say that MacMahan does not have a settled view on what we are, but that he *should* regard persons as constituted by brains and not as identical brains.

At the price of abandoning the assumption that the capacity to think is essential to persons it may avoid the problems of constitution. It avoids the troubles of Animalism. But it faces a serious problem which derives from our Nanorobotic Commissurotomy. Our Identity Brain View should say that I survive this operation, even though I am then not able to think. And, by application of our conventions for counting persons - which, as we saw from Dicephalic Twins case, rely on counting unified minds - we should say that there are two new persons. Now, these persons will be identical to sub-parts of me. But since these parts existed even before the operation, then so did these persons. The consequence is that there are always at least three persons in my body: me, and two people who we may call by familiar names "Lefty" and "Righty". Now, we should be puzzled as to what to say about the mental life of these three persons before the operation. It seems impossible to say that only the whole brain (me) thinks and the hemispheres do not. For the whole brain's capacity to think derives from its parts' capacity to think thoughts unified with thoughts produced by other parts. So it certainly seems that the parts think. So both Lefty and Righty think. But, just as certainly, it seems that each of them is capable of thinking only the thoughts produced by himself. Yet, these thoughts are unified with thoughts produced elsewhere. And our conscious thoughts (and, in general, all conscious mental items) are usually all co-conscious. Now if Lefty is aware of an item A produced by him, and A is co-conscious with an item B produced by Righty, then it follows that Lefty must be aware of B. This is what co-consciousness means. But since Lefty has only the capacity to think his own thoughts, he cannot be aware of B. So Identity Brain View apparently leads to a contradiction. In sum, the above considerations make it doubtful whether any attractive version of the Brain View can be stated.[222]

[222] Starting from Puccetti's view that there are always two persons within a normal human body, each associated with one hemisphere (see Puccetti (1973) and (1989)), one could yet entertain a Half-Brain View. But then, the relation of this view to the biological approach to identity is intangible. A hemisphere does not seem a robust biological entity on a par with the animal and the whole brain. Reasons to adopt the Half-Brain view have only to do with the realization of mental states. And so the Half-Brain view would better be stated in terms of constitution consistent with the One Realizer view.

To conclude this chapter, the One Realizer view remains the best Reductionist account so far. As we have seen, though, it come under pressure of anti-Reductionist arguments developed primarily by Dualists. I have touched on the Homunculism challenge. The next three chapters will develop further objections.

Chapter 6. Physically Based Approach - Intuitions

PBA has a somewhat narrower intuition pool than PT. There are three main motivations for PBA:

(A) "Common-sense" and phenomenology-based identificatory intuitions captured by the phrase "I am my body".

(B) Metaphysics: identification of persons with subjects of thought coupled with materialism. Our mentality seems to be sustained by a physical thing. Since there is no additional immaterial substance to do the job, this physical thing seems to be the subject of thought. So this physical thing is the person.

(C) Considerations of practices of identifying persons. We identify and re-identify people by using physical criteria. Even the grasp of our own identity relies on the use of some physical criteria.

PBA does not rely heavily on exploiting imaginary cases. With regard to these, it is mostly defensive. The proponents of PBA usually want to show that anti-PBA cases are defective or that our responses actually support PBA.

Central metaphysical issues have been treated in the previous chapter. I have little to say about intuitions from group (A). The common-sense or ordinary-talk intuitions are of little metaphysical significance, as they are not sensitive to the distinction between *being a body* and *being embodied*. Sobering remarks on the topic made by Unger and Strawson seem to me quite sufficient.[223] As for serious phenomenological work, discussion of it would take

[223] Unger (2000), 286ff; Strawson (1997), 363-370.

us too far away from the central topics of personal identity. Let me just say that there is no agreement among people doing serious phenomenology about the relation of the subject to the body on the experiential level.[224] As for problems of identification, I have argued that PBA cannot be supported by charging its adversaries with making our practices of identifying persons unjustified.[225] What is left is to consider intuition-based *objections* to PBA. The cases usually used against PBA and in favour of PT have been discussed in ch. 4 and were found unconvincing. In this chapter, then, I will focus on Body-Transfer and Body-Exchange cases.[226] My aim will be to show that:

(i) one can imagine oneself having a different body, which means that
(ii) we have no basic conceptual or experience-based intuition that we are our bodies and that
(iii) our grasp of our own identity does not depend on the application of any physical criteria.

6.1. Imaginability of Body-Transfer and Body-Exchange

In Body-Transfer cases, we imagine a person to have first one body and then another body. By imagining Body-Exchange I mean imagining a possible world where the person starts out with having a different body from the one she actually has.

The situation would be most favourable to PBA if body-transfer were not really imaginable.[227] It could be claimed that in imagining body-transfer, one either has to imagine some sort of ghostly body moving from one gross body to another (if so, one does not imagine a full body-transfer), or one has to rely on the objectionable, unclear or void idea of immaterial substance. Such claims are wrong. We can clearly imagine a continuous stream of

[224] See e.g. Cassam (1997); Strawson (1999), 104-106; Zahavi (2005).
[225] Ch. 2, 34-37.
[226] Other cases usually used against PBA and in support of PT have been discussed in ch. 4.
[227] PBA theorists usually reject transition from imaginability to possibility. This is a defense of PBA in the ontological debate. I do not discuss Body-Transfer as an argument in ontology, because the discussions on the move from imaginability to possibility seem always to end at loggerheads. Here I only ask whether PBA is supported by intuition. If we can imagine Body-Transfer, then PBA is not supported by any basic conceptual or experience-based intuition.

consciousness surviving an exchange of the body which supports it. If we take ourselves to be the subject of this stream, we thereby imagine ourselves surviving body-transfer.228 We do not need anything more to successfully imagine body-transfer. There is nothing inherently objectionable or unclear in the ideas of stream of consciousness and its subject.

An interesting challenge to the intelligibility of Body-Transfer comes from Bernard Williams.[229] We can imagine having a different psychology than the one we actually have. If, in addition, we can imagine having a different body, the following possibility opens up: I could exchange both psychology and body with another person. If so, the following also seems possible: from the start, I could have somebody else's life (all physical and psychological properties). For instance, I could have the life of Napoleon. Conversely, it is possible for Napoleon to have my life. Finally, we conclude that there is a possible world where I have Napoleon's life and Napoleon has mine. Now, Williams questions the intelligibility of this outcome. If all physical and psychological properties are exchanged, then it seems that all properties we use to *individuate* persons have been exchanged. We seem to be left with two "propertyless" selves. The problem is: how to distinguish imagining the exchange of life between me and Napoleon, from imagining the situation when Napoleon and the actual I exist? Put another way: if I strip myself of all individuating properties and take on all Napoleon's properties, do I still think about myself, or do I think about Napoleon? If we cannot distinguish between these two scenarios, we should conclude that there is no conceivable difference between them. We are confronted with one state of affairs: my self which has Napoleon's life is identical to Napoleon's self. Since we assumed that Napoleon and I are two different selves, we end up with a contradiction.[230] It is impossible to imagine

[228] Cf. Dainton (2008), 18ff.

[229] Williams (1966), 41-43.

[230] It is possible to draw a different conclusion: that there is just one Self common to all people. This kind of Monism has been entertained in classical Indian philosophies. Maddell's Simple View seems to lead to this position. Maddell would have a hard time explaining why Monism is wrong, given that he does not accept the "one consciousness - one person" principle and given that he offers no way of individuating any other

having Napoleon's life. One of the assumptions making life-exchange possible has to be rejected. Intelligibility of psychological exchange seems hard to deny. Therefore we should jettison the idea that we could exchange bodies. It was this assumption that led us to losing our grip on individuation of persons.

This engaging argument is wrong. I will show that it is possible to clearly imagine having Napoleon's life. The clue is to go step by step. Here is the procedure:

1) I can imagine having some experiences different from the ones I had in fact.

2) I can imagine having lots of them. I can imagine alternative courses of my life.

3) I can imagine a world in which I as I am now, actually turn out to be identical to Napoleon. Somehow I have I miraculously survived for the last two centuries. I had a temporary amnesia, but this should not count as a breach of psychological continuity, since I imagine regaining all memories now. Secondly, my body underwent some unusual changes. We can limit them only to outward appearance if we allow ourselves to imagine that most of our childhood memories were false.

4) If I can imagine all that, then I can imagine that I, that is Napoleon, died at St. Helena, and never made it to my current stage.

5) Finally, I can imagine everything as above, and add that some man just like me exists in the imagined world.

What makes this procedure different from imagining being Napoleon in one go? What makes it work? My answer is that by using this procedure we do not lose the grip on the reference of "I" and "Napoleon". I assume that these terms directly refer to persons. But in order to directly refer to persons, we have to have a way of picking them out. Now, when we try to

self than ours (since third-personal descriptions never suffice to individuate a person); Maddell (1981), 131ff.

imagine two “propertyless” selves which are supposed to switch properties, it seems that we have no way of distinguishing them, so we are not able to unambiguously directly refer to them. But in my scenario, we start with one self which we can easily pick out and refer to. Then we move on to consider possible worlds which are increasingly remote from the actual world. But in each step the difference is not such as to put into doubt the identity of the person. The differences are such as we normally allow to define alternative courses of a persons life.

It might seem that in my procedure I actually rely on bodily continuity. It is preserved in the crucial step 3, when I imagine myself to be Napoleon. Does my scenario provide then any argument against PBA? I think it does. For in step 5 a man appears who possesses the body which is mine in the actual world; while I have Napoleon’s body. So I imagine a situation in which I have a different body from the one I actually have. If we remain in the Kripkean framework, and think that “my body” directly refers to my actual body, we will have to conclude that it is not *possible* for my body not to be identical to my actual body. But as far as *imagination* is concerned, we can imagine my body turning out to be non-identical to actual Milosz’s body.

My scenario has two interesting features. First, it works only for identification with persons who live earlier than one. This follows from the nature of step 3. I can imagine that it *turns out* that I am identical to a person from the past who made it to my actual stage. But it makes little sense to say that I could turn out to be a person who has not yet been born. I cannot imagine a person from the future having *these* present experiences.[231] This means that we lose the anchor which enables us to uncontroversially pick out a particular self at the outset. The only safe way to coherently imagine being a different person is by imagining that I had her experiences, forgotten them and managed to live until I have reached my present

[231] I could imagine that it turns out that I actually live in what I taken to be the future. Say, it turns out, that it is 2500 AD and not 2009. But this is not imagining being someone from *my future*. It is imagining that I exist at a later time than I think I do.

stage. The only way to imagine a person from the future having my actual experiences is to entertain a time-travel scenario. In this case, experiences which take place in what is objectively future relative to the present, turn out to be in my past according to my *personal time*.[232] An interesting corollary follows:

(SIPT) Self-identification presupposes the grasp of the flow of personal time.

Secondly, my scenario does not describe an *exchange* of lives between me and the person in the past. We imagine that the self called "Napoleon's self" is me. So there is no second self which can take my life in exchange. I can imagine some self having my body and life, but, within the confines of our imaginative project, it is impossible to say that it is Napoleon's self. But if we take this scenario to support the idea of simple selves, then, using this idea, we can proceed to imagine being born in the future or exchanging lives.

I do not think that these thought-experiments show that it is possible for a person to have a different body and life. They show, however, that (i) our understanding of the reference of "I", and (ii) our implicit understanding of ourselves and (iii) the ability of self-recognition in imagination do not rely on the use of any description involving the body. How then do we recognize ourselves when we imagine that "we are someone else"?[233] I claim that we do not carry out any specific act of recognition of the self in imagination. We have no need to do it. This follows from the way imagination works. To explain this claim I will use the analogy to the theatre.[234] Theatre involves three parties: the audience, the actors and the author. The audience can adopt various modes of experiencing things in the theatrical

[232] Cf. Lewis' distinction between "external" and "personal" time in Lewis (1976*b*), 69-72.

[233] The phrase "imagining to be someone else", if taken literally, would mean imagining being identical to a person which is non-identical to one. That would be imagining a contradiction. That is not what I mean. I mean imagining being identical to a person which in fact, but *not according to the imaginative project*, is non-identical to one.

[234] Cf. Williams (1966), 35ff ; Wollheim (1984), 65-71.

framework. Suppose we are watching *Othello*. We see a man who in fact is an actor. We have no trouble reidentifying him as he appears in successive scenes. But *who* do we see? There are two ways of perceiving this man. We can "suspend disbelief" and *see* the man *as* Othello. The man which we see *is* Othello. "Who is *this*?" one can ask pointing at the man: "*This* is Othello; and the woman over there is Desdemona." Call this the *immersive* way of seeing. But it is also possible to see the play as play. Then the man is seen as someone else than Othello. He and his actions only *represent* Othello and his actions. Call it *representation-conscious* way of seeing. The second distinction concerns perspective. We may watch heroes and action from the outside or, to use Wollheim's term, *acentrally*.[235] But for the audience in imagination –and for an actor in a play – it is possible to watch the action from the perspective of one of the heroes. In this case, we *centrally imagine* somebody's actions. Now, when we use our imagination we perform single-handedly the functions of the author, the actors and the audience. So what happens when we centrally imagine being identical to someone else? How, say, can we imagine being Napoleon at Austerlitz? As the imaginative audience we are in the same position with regard to ourselves as the theatre audience is to actors. The audience which is immersed in the play sees the man as Othello. His words and actions are seen as Othello's words and actions. Now, in my private theatre, I can see not only words and actions, but thoughts and feelings as well. But the principle is the same. The actor which I see is *in fact* the actual me. But I *see* myself *as* Napoleon. My self is seen as Napoleon's self. My thoughts, feelings, words and actions (those that are a part of the imaginative project) are seen as Napoleon's thoughts and actions. "Who is there?" you can ask. It is Napoleon. It will help if we realize that an actor in an ordinary play may be at the same time a spectator immersed in the play. Immersion is not delusion. The actor will not frame an identity statement using descriptions actually true of him; he won't have absurd

[235] Viewing things acentrally does not always mean viewing them from the third-person perspective. Wollheim rightly observes that in imagination things can be viewed from no particular perspective. The right analogy in this case is with painting, which need not involve the use of perspective.

thoughts like “The so-called employee of this theatre *is* Othello”. No. What the actor does is to directly refer to himself, directly experience himself and see-himself-as Othello. Asked “Who are you?” he responds “Othello” and so on.

What is peculiar about centrally imagining someone is that we do not produce in our mind any new item which represents the imagined person. It is simply our self which is so seen.[236] Things could not work otherwise. If the actor is to represent actions and words, he must produce similar actions and words. If I want to imagine thoughts and feelings, I have to produce some thoughts and feelings. But I cannot produce in myself the property of being a self different from myself. This is absurd. The only alternative is that *I* am seen as the person. Now, since I deal with this one directly given self only, the issue of *recognizing* it in the imagined world does not arise at all. I do not see another self and its properties in the imagined world; I see this self as having properties different from the actual. I can be said to *project* properties onto this self, or, equivalently, to project this self into an imagined situation.[237] I do not have to use any description to recognize myself in the imagined world.

Memory works in a similar way. My present self is directly given to me in experience. Now, memory, it seems to me, involves presentation of *this* self as having been *really* involved in a past situation. Again, there is no issue of identifying a “past self” as me; I am not aware of any such additional thing.[238] On the linguistic level, we simply ascribe

[236] This claim is not necessarily opposed to Velleman’s distinction between the “notional” and the “actual” self in imagination. Velleman discusses in fact only the *representation-conscious* way of seeing: “the image in my mind, *regarded as a copy* of NB’s visual impression, is an image of whatever NB is supposed to be seeing”; “[the notional subject] gets into the act by *being thought of* as the subject, the person reflexively presented by the image”; Velleman (1996), 181f (italics mine). In representation-conscious cases indeed we can talk of two subjects. We *see* the actor - the *representing* subject - and we *think* of Hamlet - the *represented* subject. Velleman’s distinction, as far as I understand it, comes down to this simple distinction. My claim is that the representing subject is one and it is my actual self. This is consistent with Velleman’s distinction. However, there is a further issue. What characterizes the *immersive* way of seeing is, precisely, that the *representing* and the *represented* subject are seen as one. The notional and the actual subjects coincide. Velleman does not consider the immersive form of imagination, and this vitiates most of his subsequent argument about identity and anticipation.

[237] Cf. my discussion of Martin’s account of “projection” in anticipation, ch. 8, 196ff.

[238] What about people who *do not* identify with “the past self”? I think they just use a bad idiom. What probably happens is that the past situation is not presented to them as having *really* happened to their present self. Strawson expresses his impressions in almost the same words; Strawson (2005*b*), 67f. Now, remembering in

tensed predicates to the present self directly referred to. This account explains in the simplest manner possible why memory is taken to be (i) veridical by definition and (ii) to entail identity by definition. It is tempting to extend this analysis to anticipation. I will argue for such extension in ch 8.

My account of self-recognition in imagination has several virtues. It is simple. It is general. It can be extended to memory and anticipation. The notion of seeing-as, on which it relies, is a potent tool for explaining phenomena of imagination, perception, art and religion. Finally, my account provides an elegant explanation of such phenomena as the apparent conceptual link between memory or anticipation and identity; and the asymmetry between the ease of imagining one's alternative lives when the imagination is anchored in the present and the difficulty of identifying oneself in possible worlds without such anchor.[239]

The account presented here is quite close to William's account in "Imagination and the Self". The central idea that it is my actual self which is the actor representing the imagined person is the same. And yet the conclusions are diametrically opposite. Why? William's diagnosis of the situation can be put as follows. There is a feasible project of *centrally imagining* Napoleon. In this case my actual self - the self with all its actual properties - represents Napoleon. But this project says nothing about the real me. The real me is not the part of the imagined world, as the project is about Napoleon, and not me. The second, impossible project, is imagining that is really possible for *me* to be Napoleon. This project is about me all right; but it is impossible to carry out because of the loss of the grip on personal identity at some point. The illusion that I could imagine really being Napoleon derives in part from confusing these two project. Now, to assess the merits of William's position we need to ask one question: what is it about the actual me that enables me to represent Napoleon? We need to distinguish between the features of actors *in virtue of which*

this way is just like vividly imagining something which happened to someone else, while knowing that this situation really took place. So this case does not pose a problem for my account.

[239] See ch. 4, 102.

they represent personae, and the features which are irrelevant. Let us take Othello again. The actor which plays Othello is usually dressed up; he is a human being; a male, or at least someone who looks like a male. Many other features will be irrelevant: being short, being 35 years old and having children etc. If such properties are known by the audience, they will be bracketed in the act of seeing-as. But looking like a human male is certainly relevant; this feature enables the actor to be seen as Othello. Imagine, for the contrast, the following situation: a Desdemona-look-alike woman dressed in a white robe enters the scene and says Othello's lines; and then a tall Afro-American actor enters and says Desdemona lines. Unless we get some further clue as to who is who (e.g. when other personae address Desdemona and Othello by name) we would be confused as to what is going on. In the end, anything can represent anything. But it will only do so in virtue of having some distinguishing features which, in extreme cases, will be arbitrarily chosen. Moreover, for there to be *immersion*, the relevant features cannot be arbitrarily selected: there has to be some minimal resemblance. Consider Othello-Desdemona gender exchange. Is it possible to view such play in an immersive way? Not really. With the distinction between relevant and irrelevant features in hand, we can go back to imagining being Napoleon. In virtue of *what* does my self represent Napoleon's self; and what makes immersion possible here? Surely, my living in the 21st century, being born of my actual parents, and so on, are irrelevant features. In fact, *all* my actual physical and psychological properties are bracketed. My self does not represent Napoleon in virtue of having them. In virtue of what then? The answer is: just in virtue of being a self; a subject of thought. But if we are able to bracket *all* my physical and psychological properties and yet say that *I*, the subject, represent Napoleon, then, evidently, we have a grasp of the notion "subject" which does not involve any physical and psychological elements. And since we can entertain the body-exchange scenario, we do not need physical criteria to retain the grasp on the identity of the subject. It is by tracking in imagination the flow of personal time or the flow of consciousness that we hold onto the identity of the subject.

It may be protested that there is another important property which Napoleon and I have in common: having *a* body. I have not shown that a disembodied being is imaginable. So it is still open for the proponent of PBA to claim that we have to think of ourselves both as subjects *and* as physical things. I prefer to avoid the discussion of disembodiment, which brings issues not directly related to personal identity. But there is also no need for me to discuss it, given that the imaginability of disembodiment has been satisfactorily demonstrated by W. D. Hart.[240] We can imagine ourselves becoming disembodied ghosts. My self will represent the ghost's self just in virtue of being a self. My point stands.

Finally, we should note an important consequence of the thesis that in imagination our self is seen as the imagined person. Contrary to what R. Wollheim has claimed, it is impossible to centrally imagine somebody else.[241] Of course, it is possible to imagine somebody who we know is not identical to us in fact. But it is impossible to centrally imagine somebody who is not me within the imagined world. Consider Wollheim's case: imagining the Sultan's entering Constantinople as victor. My having this imaginative project involves seeing *myself* as the Sultan. So within the project *I* am the Sultan. I can express this fact from the audience's side: "This is Sultan", or from the immersed actor's side: "I am Sultan". In any case, "this" and "I" directly refer to my self of which I am directly aware. In imagining the situation I projectively ascribe various properties - being a glorious victor, being merciful – to this self. We can test this account against Wollheim's hard case. It seems I could imagine Sultan centrally and me – Milosz – being brought to him. Milosz is being imagined peripherally. I can imagine the Sultan magnanimously saying to his subordinates "*I* order you to free *him*". This is certainly perplexing. Where am I in this world? My response

[240] Hart (1988). See also Hart and Yagisawa (2007).
[241] Wollheim (1984), 74.

is that in this world I am the Sultan. The identification of Milosz as *me* is external to the imaginative project. We are imagining the Sultan, and Milosz should be just described as a man with such and such properties. From such descriptions it does not follow that he is me. And, as the use of personal pronouns in the Sultan's order makes clear, I am playing and I am seen as the Sultan. So I am the Sultan, not Milosz in this world. But what if our project would be defined precisely as "centrally imagining the Sultan and *me* being brought to him"? My answer would be that the function of 'me' in this context is just to pick out a man in the actual world, who will become an element of the imaginative project. But within this project this man will not be presented as me, for within the imagined world "I" and "me" refer to the Sultan. In terms of the theatrical analogy, it is only the playwright, who is not immersed in the play, who makes use of this external project-description, and not the actor and the audience. But we might also observe that in imagination we often shift the perspective. After all, we are just one actor, and to represent complex situations we must switch from role to role. When I imagine me being brought to Sultan this will almost inevitably occur. So I will alternately see myself as the Sultan and as Milosz. Can we have our internal actor play two roles *simultaneously* (somewhat in the manner of a puppeteer having a dialogue with a puppet)? I suppose it is hard for us humans, but in principle possible. This would have to be done in the following way: some mental items would represent Sultan's consciousness; other items would represent Milosz' consciousness; and these items, though *in fact* co-conscious, would be *represented/seen-as* disconnected. Secondly, as usual, "I" as used by Sultan and Milosz will directly refer to my self. How then am *I* seen? There are two options. First, we could say that I am seen as one self with two separate minds. But I think it is better to say that my self would represent two people. After all, one actor, or one piece on the board, may represent a whole army. My self would be referred to and experienced twice over. That it is *the same* self which is so experienced is a fact, but not something which enters into the content of the imagination. Could the same thing be said about the property of being *me*? Could we say that although the self which is seen as the Sultan is *in fact* me, this is something which does not enter into the content of imagination? If so, William's point that imagining being Napoleon has really nothing to do with *me*, would still stand. But we cannot say that for two reasons. First, seeing myself as not myself but somebody else would require me to be able to separate my experience of my self as *a* self from the experience of my self as *my* self.[242] This is impossible. I cannot find any quality of *mineness* (if indeed there is anything like that), if only for the boring reason that I do not and cannot know any other self first-hand, and so I cannot abstract the quality of mineness from the experience of selfhood. Secondly, if it were possible to experience oneself as just a self and not one's self, then upon any act of self-awareness there would still arise the question "Who is it?" This would, absurdly, make self-knowledge impossible. We have to conclude that to experience a self is always to experience it as one's self.[243] Secondly, it is difficult to understand how personal pronouns would work on William's account. If "I" is used in an ordinary way, then it refers to me and it is impossible to say that the imaginative project does not concern me. Then should we perhaps say that I *pretend* to refer to me when I say "I am Napoleon"? Yet it is hard to say how "I" would be understood then. On the best interpretation, acting comes out as *quoting.*

[242] Note that my claim is restricted to *experience* of the self. I do not deny that we have *general concepts* like "self", "first-person perspective" etc. (though how exactly do we get by them is, I admit, not wholly clear to me). So I do not deny that we can *think* of other selves; and that we can imagine what it is like to be them without imagining being them. But for the particular project of centrally imagining the *Sultan*, mere thinking is not enough. Particular form of experience of the self is needed; and here comes the point that we cannot experience a self otherwise than as *our* self.

[243] These considerations lead to an interesting problem about other minds: the seeming impossibility of knowing what it is really like to be another self. Since I cannot compare different selves first-hand, I cannot know how much in my experience of my self is specific to me and constitutes "mineness"; and how much is non-specific and common to every self. In empathizing, I can only put *myself* into the boots of another person; and I can never know to what extent by doing so I misrepresent what it is like for her to be in the situation. An elegant solution to this problem is to deny the possibility of haecceity with qualitative character. My self has no experienced essential qualities other than that of being a self; and its recognition as mine is a matter of perspective or access and not of grasping any of its qualities. This solution accords with (my) phenomenological experience of myself. Secondly, it could be argued that if there were an experienced qualitative property "mineness", then since we could not distinguish it from the property "selfhood" we could not even *conceive* of other minds. This, parenthetically, is another problem for Maddell's Simple View account.

But that is obviously implausible. When we are in the theatre we do not take actors to be quoting the lines of the characters; we take them to say the lines and to be the characters. To conclude, in imagining being somebody else, we do implicitly take ourselves to be identical to the imagined person. But then it is also possible to explicitly imagine that one could be the imagined person. Contrary to William's claim, imaginability of being somebody else has direct consequences for what one can imagine about oneself.

6.2. Conclusions

If we conceived or experienced ourselves as being essentially normal physical things, then we could not properly imagine Body-Transfer and Body-Exchange. We would, as Williams claimed, lose the grip on our identity somewhere in the process of imagining these case. Yet this is not true. The cases are imaginable. We can experience ourselves as mental subjects; and as long as we can trace the flow of personal time or the flow of consciousness, we have a sure grasp of our identity. We do not and need not conceive or experience ourselves as bodies. The belief that people are bodies could yet count as a deep belief if it were presupposed by our attitudes and practices. Our practices of identifying ourselves and others do not, however, presuppose such belief. Nor do the moral attitudes and practices. I have argued that the support they give to PT as against PBA has been seriously overplayed. But the fact that our moral practices can be regarded as *consistent* with PBA does not positively support PBA. For they are equally consistent with the Simple View. Thus it cannot be claimed that PBA is presupposed by them. To conclude, PBA seems to lack any solid intuitive support. The best hope of PBA lies in purely metaphysical considerations. But the Homunculism challenge points to its deficiency in metaphysics as well.

Among many valuable insights in William's discussion of Body-Exchange, there is also the point about individuation of selves. If we allow body-exchange, then, Williams claimed, we do not have in hand any individuating property of selves. This, I submit, is right in the sense that we fail to grasp any *metaphysically* individuating properties. We can imagine being anybody. And, due to the nature of self-experience and of first-person reference, *but not due to the grasp of any individuating features*, we have no problem with identifying ourselves in the imagined world (*pace* Williams). But if it is asked whether it is *really possible* for me to have another life; and *what* makes it the case that the possible person is me, then it is far from clear what we should say. This accords with the conclusions I have reached in ch. 4.

Should the failure to provide clear conditions of individuation and trans-world identity of selves cause embarrassment to a proponent of the Simple View? Not necessarily. On some versions of the Simple View, selves have no components and are among ultimate constituents of reality. Now, to be fair, we should ask whether we have a grasp on individuating features of ultimate constituents of the material world? Hardly. The extent of our grasp of the identity of, say, an elementary particle is very narrow. The best we can come up with is a procedure analogous to the one I employed in connection with selves. We can start with an (indirect) experience of the ultimate thing which enables us to pick it out. Then we can work our way backwards and forward in time tracing the possible histories of the particle. But in the absence of the anchor - the initial experience - questions about trans-world identity of the particle seem perfectly unanswerable. We are in the same position with regard to any ultimate constituents of reality. If selves are among them, we should not be troubled by our inability to grasp their individuating features.

Chapter 7. Fission and the Continuity of Consciousness

In the previous chapters I have suggested that the possibility of continuous consciousness linking experiences of a person is a necessary condition of personal identity. It is time to elaborate this idea. I start with preliminary considerations on the temporality of consciousness and I define several kinds of continuity of consciousness. Then I argue for a particular formulation of a necessary criterion of identity based on potential for continuity of consciousness. The second part of this chapter will be devoted to an argument to the conclusion that Fission cannot preserve continuity of consciousness. This conclusion will be used to argue against Reductionism and in favour of Dualism.

7.1. Five claims about temporality of consciousness

I make five basic claims concerning temporal features of consciousness and its contents:

(1) Specious present (phenomenal "now") normally has temporal extension (duration).[244]

(2) Experiences normally have temporal extension.

(3) Consciousness is changing in time: one specious present is supplanted by another; experiences come and pass.

(4) Experiences are normally experienced as *oncoming* or *yielding.*

(5) Co-consciousness of experiences is identical to their co-presence in a specious present.

The first three claims should be reasonably clear. Let me first comment on my last claim. It is customary to distinguish between synchronic co-consciousness (relation between experiences

[244] Perhaps specious present is not always extended and it is possible to be conscious of an instantaneous happening. I leave this possibility open. Even then co-consciousness should be possible: if one could be aware for an instant, one should be able to be aware of a red patch and a green patch together.

occurring at the same time) and diachronic co-consciousness (relation between contents occurring at different times).[245] This distinction is fishy. In order to use it, we should be clear what we mean when we talk about temporal location and relations of experiences. Do we mean their objective, *physical* location and relations? Or do we mean their *phenomenal* location and relations? I think it is wrong to *assume* that experiences have location in physical time.[246] Secondly, location of experiences in physical time is not a phenomenological fact (as location in objective time is not a purely phenomenal property).[247] Co-consciousness means that contents are experienced together; they are co-present in consciousness. From the point of view of consciousness, all co-conscious contents occur at the same time: "now". Thus co-consciousness is identical to being in the scope of a specious present. The distinction between synchronic and diachronic consciousness is a non-phenomenological distinction made from the perspective of the objective time. The phenomenological fact is that contents within specious present have *some kind* of duration and are *in some way* temporally arranged. This may be necessary in order for contents to *represent* events occurring at different physical times. But again it would be wrong to simply assume that the phenomenal arrangement of contents and represented time-order of events in the world faithfully corresponds to objective time-relations.[248] The "neat" distinction between synchronic and diachronic co-consciousness obscures complex problems of relations between consciousness and time. Let me present a list of temporal-like relations which seem to characterize the realm of consciousness:

1) Temporal order of physical events (objective physical time)

[245] See for example Dainton (2000), 3f, 25f. Tye (2003), 13-21, 85ff operates with an equivalent distinction. It is worth noting that both Dainton and Tye develop accounts where one factor (primitive co-consciousness and singleness of experience respectively) accounts both for synchronic and for diachronic co-consciousness.

[246] Cf. ch. 9, 231.

[247] Phenomenological description will only note that human consciousness is liable to *seem* to an introspecting subject to unfold in the objective time.

[248] Cf. Dennett (1991), 143-153.

2) Co-presence of contents within a given specious presence: co-consciousness

3) Arrangement of contents within a given specious present. Example: I see a car moving from A through B to C. I see it “now”, in one specious present. But there is some “earlier-later”-type arrangement of sub-contents: the arrangement between seeing the car moving from A to B and from B to C.

4) Dynamic features of experiences: experiences are oncoming or yielding.

5) Succession of specious presents.

6) Represented time-order of physical events.

7) Represented order of succession of experiences. This is the order experiences seem to us to have when we reflect on what happened earlier and what later.

Items 2) to 5) make for phenomenal or subjective time. Given the list above, it will often be unclear as to what exactly we mean by “earlier” and “later” when reporting on our experience. Likewise, the exact meaning of “duration” as applied to experiences will be open to question. Even if we limited our attention to the phenomenal, duration could still mean different things: either temporal extension within a specious present or being present in several successive specious presents. I will not try to solve these questions now; these ambiguities will not affect the main argument. My list is tentative and likely incomplete; some items may turn out to be spurious. I do not pretend to have a fully-worked out theory of temporality of consciousness, nor do I think anybody else has it. My point is this: unless we have a theory which does justice to the complexity of phenomenology and explains the relation of consciousness to physical time, using the distinction between synchronic and diachronic co-consciousness may beg important questions and lead to confusions. I will decline to use it. There is no phenomenological difference between co-consciousness among contents temporally arranged within one specious present according to an “earlier-later” relation and co-consciousness among contents within one “specious present” which are not so arranged.[249] So I will speak of co-consciousness without qualification and take it as identical to contents’ co-presence within one specious present. It is time to move on to explaining my fourth claim: that experiences are normally experienced as oncoming or yielding.

7.2. Dynamism and continuity of consciousness

Our consciousness undergoes changes. Secondly, we are often directly aware *of* changes. In particular, we can be aware of changes of our consciousness. These obvious observations lead to a much less obvious claim: how it is with us phenomenally cannot be fully grasped in static terms, without taking note of the dynamic features of experiences. I claim that there is a phenomenal difference between being aware of an experience which freshly starts being experienced and having an experience which is in the process of ceasing to be experienced. This is the difference between *oncoming* and *yielding* experiences.

To get a better grasp at this idea, let us again consider seeing a car traverse some distance. I am aware of its movement. This is a basic experience. Within a given specious present, I have (sub-)experiences as of the car being in several different places.[250] But that would not be enough for me to be aware of movement. For I can experience something like this even when there is no movement (e.g. when I cross my eyes, the same thing seems somehow to be in two places). Secondly, this would not determine the direction of movement. To be aware of movement, experiences of the car in different position have to be somehow arranged temporally within the specious present. Yet a mere static arrangement in a

[249] Such contents can be picked out as those with regard to which we are either agnostic as to which occurred earlier; or which positively seems to us to have occurred “really in the very same moment”.

[250] If the car moves *very* fast I will indeed experience a blur and I will really see the car in several positions at once (not only within one specious present, but within *it* I will see it being in different places simultaneously - “simultaneity” meaning here a relation arranging contents within a specious present). But this does not always happen. With lower speeds, my experiences will have a different arrangement and dynamism. I devised a helpful exercise for observing these differences. Try to quickly waggle a pen back and forth with your fingers. Observe the fast-moving end, the middle, and the tip which you hold. In each case you will see a movement. But the way you will see it will be very different in each case.

(phenomenal) temporal dimension would not account for the way we experience things. We are also aware of the direction of the flow of (phenomenal)time. Thus some experiences will be *oncoming* and some of them *yielding*. More precisely, some given experiences will yield *to* particular oncoming experiences.[251]

There seems to be no discernible middle-ground nor *limes* between what is oncoming and what is yielding in consciousness. Together, the oncoming and yielding experiences constitute our "now". To have a phenomenal "now" is not to be aware of a (static) content of an atomic moment. It is to be aware of something oncoming and something yielding to it. It also seems that experiences could not have this character if there was no *real* change.[252] That is, the experiences which yield promptly disappear from our consciousness. Likewise, we experience as oncoming those experiences which really just started to be in our consciousness.[253] It is hard to imagine how things could be otherwise.

The relation of yielding between experiences provides the ground for defining a kind of dynamic continuity of consciousness. The definitions will be formulated for experiences having temporal extension. Although I prefer to interpret these definitions as involving only the notion of phenomenal time, they could also be interpreted in accordance with the view locating experiences in objective time.

(DF1) For any experiences A and B, A and B are *dynamically successive* iff (i) A starts to be experienced before B (ii) there is a time when A is experienced as oncoming (ii) there is a time when A is experienced as yielding to B.

(DF2) For any experiences A and B, A and B are *dynamically continuous* iff there is a chain of dynamically successive experiences from A to B.

[251] The idea could be equally well stated in terms of *experiencings* rather than experiences.

[252] I leave it an open question whether real change requires A-theory of time. It certainly seems so to me.

[253] Though, perhaps, the experiences which were unattended to could be experienced as oncoming upon turning attention to them.

(DF3) For any two states of consciousness C and D s.t. C is earlier than D, C and D are *dynamically continuous* iff there is experience A in C and experience B in D such that A is dynamically continuous with B.

We have defined a new kind of continuity of consciousness. But it will also be good to have a more traditional notion defined in terms of co-consciousness. The structure of definitions is virtually identical.

(DF1a) For any A and B s.t. A starts and ceases to be experienced earlier than B, A and B are *directly continuous* iff from the time B starts to be experienced, A and B are co-conscious as long as A is experienced.

(DF2a) For any A and B, A and B are *continuous* iff there is a chain of directly continuous experiences from A to B.

(DF3a) For any states of consciousness C and D s.t. C is earlier than D, C and D are *continuous* iff there is experience A in C and experience B in D such that A is continuous with B.

Continuity and dynamic continuity go hand in hand. If A yields to B, then A and B must be co-conscious. Can A continue to exist but cease to be co-conscious with B? This will be a matter of our argument about Fission. For now, suffice it to say that in all normal cases A's yielding means that it promptly ceases to exist, so there is no time for it to grow apart from B. So at least in normal cases dynamic succession will imply direct continuity. In the other direction, direct continuity entails some direct succession. If A ceases to exist earlier than B, there will be awareness of the transition between the state when A is present and when A is not present. It will be experienced that A-co-conscious-with-B yields to B-without-A. Putting the two kinds of continuity together we can define the notion of *strong continuity*.

(DF4) For any A and B, A and B are *directly strongly continuous* iff B is both directly continuous and directly dynamically successive to A.

(DF5) For any A and B, A and B are *strongly continuous* iff there is a chain of directly strongly continuous experiences from A to B.

(DF6) For any states C and D s.t. C is earlier than D, C and D are *strongly continuous* iff D is both continuous and dynamically continuous with C.

We can also define weak or *disjunctive* continuity. Contents or states will be so related when they are related either by dynamic continuity or ordinary continuity.

7.3. Potential continuity of consciousness as a necessary condition of identity

Whether or not we are conscious at all times we exist, it seems undeniable that there are discontinuities in our conscious life. Experiences we have just before falling asleep and those we have upon awakening may seem not to be even weakly continuous.[254] For this reason continuity of consciousness is not a good candidate for a criterion of personal identity. *Potential* for continuity of consciousness is more promising.

First, let us tentatively assume that capacity for consciousness is an essential property of persons. So at every time of her existence a person could be conscious. From this it does not follow that a person could be conscious at all times; that she could have unbroken consciousness throughout entire life. Perhaps it is a necessary property of human persons that

[254] Later I will call this appearance into doubt; ch. 9, 230. For now we can take it at face-value.

they have to sleep sometimes. Therefore it would be too much to demand that it should always be possible to link two stages of a person's life by a continuous chain of experiences. If we take a stage at the beginning of a human life and at the end of it, this may not be possible. So we have to ask for less. Consider then this idea. Even if we have to sleep some nights, we could always stay awake on any particular night. It is certainly true that for any two stages of our conscious life which are no more than 24 hours apart, there could be continuous chain of experiences between them. To be cautious, we can take less than 24 hours. In any case, there will be some minimal continuity-safe period over which it will always be possible to make a link between experiences. This period may be different for different kinds of persons. Now consider a person *x* at t_1 and *y* at t_2. We need not postulate consciousness at the respective times. We can imagine a person who spends many years in a comma, and ask whether the person in the first year of coma and the person in the same body 20 years later, still in a comma, are the same person. We think it is. This person *could* have a normal conscious life, with only short periods of unconsciousness necessary for sleep. So she could have experiences over the 20 years in question. And for every minimal continuity-safe period within these 20 years, the experiences which she could have within this period could be linked by continuity. By iterating modalities in this way, we obtain a weak, but intuitive and plausible necessary condition of personal identity.[255]

One more refinement needs to be added. It is not always possible to have a continuous link between two *particular* experiences which are actually dis-continuous, even if they occur no more than 24 hours apart. First, if causes have something to do with the individuation of experiences, and introducing the continuity would require altering causal processes, then we could not have the very same experience at the end of the chain. Secondly

[255] At the beginning we mentioned the assumption that capacity for consciousness is essential to persons. This view may be contested, but it is not required for our criterion. *Possibility* of having consciousness at a given time is enough. Arguably, even a fetus in early stages has it. For by a miracle in Lewis's sense, or by Unger's "statistical miracle", it could have developed at an incredible speed, so as to gain consciousness at the time in question.

and prosaically: if we experience *awakening* from a dreamless sleep, then we could not have *this* experience, nor anything remotely similar, at the end of a chain of continuous experiences.[256] So we should understand the potential for continuity of consciousness in a different way. Given two times t_1 and t_2 the person will have a potential for continuity of consciousness over this period if and only if there will be *some* possible experiences: C at t_1 and D at t_2 such that C and D are continuous.

We are now in the position to formulate our criterion:

(C) For any persons *x*, *y* s.t. *x* exists at t_1 and *y* exists at t_2 ($x = y$ only if there is a *possible* (or actual) series of conscious states (A,..., B) such that:

(i) *x* has A at t_1 *&* *y* has B at t_2 & the states between A and B occur at times between t_1 and t_2

(ii) every two successive states C and D in the series (A,..., B) *either* are (weakly) continuous *or* there is a possible series of (weakly) continuous experiences (C*,..., D*) s.t. C* is had by the person who has C and occurs at *t*(C) and D* is had by the person who has D and occurs at *t*(D).)

Let me show how this formula works. Suppose *x* and *y* are identical and check if (C) holds. Call the actual states of *x* and *y* A and B. We have three possible cases. In the first case, the *actual* states are linked by continuity of consciousness. Then there is a series (A,..., B) meeting the requirements. Secondly, A is, let's say, the experience of falling into sleep and B - that of awakening, separated by a few hours. Then A and B cannot be continuous. However, *x* could have stayed up that night, so there is a possible world w_1 in which the time-wise counterparts, so to say, of A and B (call them A* and B*) are continuous. So there is a possible series (A*,...B*) meeting condition (i) and the first disjunct of (ii). Finally, suppose

[256] In my terminology, in experiencing awakening we have some experience as *oncoming*, but without any prior experience yielding to it. Symmetrically, in falling asleep we have some experiences yielding but not to any further experiences. Or so it seems. If there was continuity, our experience would be different: we would experience contents continuously yielding to other contents. If dynamic features of experiences are phenomenal features, as I claim, then we could not have exactly *similar* experiences at the end of a dis-continuous chain and at the end of a continuous chain.

A and B are separated by a long stretch of time. In some possible world w_1, *x* should be able to live a normal human life - punctuated by short periods of unconsciousness - from t_1 to t_2. So there will be a series, say, (A*, C*, D*, E*, F*, B*). Suppose that A* and C* are continuous, as are D* and E* and F* and B*; but not C* and D* and E* and F*. But to the non-continuous couples we can apply the same reasoning as to our case two; so there will be w_2 with a continuous series (C**,..., D**); and w_3 with a continuous series (E**,..., F**). So the series (A*, C*, D*, E*, F*, B*) satisfies (i) and the second disjunct of (ii).257

It will be handy to say that whenever the condition imposed by (C) is met there is *potential continuity of consciousness* between *x* and *y* (or their states). Potential continuity of consciousness is a necessary condition of personal identity. Is it also sufficient? I will argue that it is, but that it does not give an *informative* analysis of personal identity. As some issues need to be settled before that, I defer the treatment of this issues to chapter 10.

My criterion is very close to the one proposed by John Foster.258 Foster's criterion rests on the idea of hypothetical extendibility (logically or nomologically grounded) of an earlier stream to the beginning of the later stream of consciousness (or backwards extendibility of the later stream to the ending of the earlier one). The chief difference is that by operating with counterpart states and iterated modalities (C) takes care of cases where the actual states could not, by their very nature, be joined by continuity of consciousness and cases where *x* or *y* is unconscious at the specified time. It is not committed to the idea of "joinability by extension" of the actual states (we can say that while at least one of the states: *x*'s or *y*'s must be held constant in Foster's criterion, they are variable on my account; indeed, (C) makes no essential use of the actual states of *x* and *y*). Finally, (C) is noncommittal with regard to grounds of potential continuity of consciousness; it operates with pure possibility.

There is one more difference between Foster's general account and mine. Foster

[257] In terms of tracking identity: we track *x*'s identity along the series (A, A*, C*, C**, D**, D*, E*, E**, F**, F*, B*, B). Shifts in the number of stars signal shifts between possible worlds.

[258] Foster (1991), 251, 260.

tentatively accepted the possibility of division of consciousness (but not of the subject).259 Using the principles which, I think, Foster himself would accept, I am going to show that this is impossible.

7.4. Fission and continuity of consciousness

In this section I will argue that Fission cannot preserve continuity of consciousness.[260, 261] Details of Fission do not matter. Fission is assumed to result in the situation where there are two centres of consciousness when previously there was one. Any process leading to such result, I will argue, cannot preserve continuity of consciousness.

Let us have three traditional names "Pre-Fission person", "Lefty" and "Righty" to call persons involved in a Fission situation. Imagine that for some time before Fission, Pre-Fission person was subjected to the following experience: she was seeing only white light with both eyes. Call this experience U. Now at the very moment of Fission we will have the white light abruptly change to green light shining on the right eye (but not visible through the left eye); and a red light for the left eye (but not for the right). For simplicity, suppose that Lefty sees only with the right eye, and Righty with the left eye. So Lefty will have an experience of green light, call it G. Righty will have an experience of red light, call it R. Now, the situation with U, G and R is this. R and G are *not* co-conscious. For they occur in two separate streams of consciousness (call them Lefty- and Righty-streams). But U is supposed to be continuous with G and with R. This is what it means for Fission to preserve continuity of consciousness. And this I will show to be impossible.

[259] Foster (1991), 258ff.

[260] See Unger (1990), 187-191 for an argument to the contrary.

[261] The argument applies, *mutatis mutandis*, to Fusion too.

Argument Against Division of Consciousness:

1. R and G are not co-conscious at any time.

2. U is strongly continuous with R and U is strongly continuous with G.

3. Therefore either U is directly strongly continuous with R and with G or there are chains of strongly directly continuous experiences linking U to R and U to G.

Case 1.

1.1. U is experienced as yielding to R *and* to G (by assumption)

1.2. If U is experienced as yielding to R and to G, then this experience of U involves co-consciousness of U *and* R *and* G.

1.3. Therefore, there is a time when R is co-conscious with G.

Contradiction (1.3, 1)

Case 2.

2.1. Call the first members of the chains linking U to R and U to G: R* and G* respectively. R* and G* are either co-conscious or not.

Case 2.1.

2.1.1. If R* and G* are co-conscious, then all later members of one chain are co-conscious with co-existing members of the other chain, by the application of an argument analogous to that presented in

Case 1. (consider compound experience R*+G* yielding to successive experiences and so on).

2.1.2. Therefore R is co-conscious with G.

Contradiction (2.1.2, 1)

Case 2.2.

2.2.1. R* is not co-conscious with G*.

2.2.2. U is strongly directly continuous with both R* and G*.

(by assumptions)

2.2.3. By argument analogous to Case 1: R* is co-conscious with G*.

Contradiction (2.2.1, 2.2.3)

Contradiction (1,3)

QED

Let me note what this argument presupposes and what it does not. The argument exploits the notion of *dynamic* continuity. The crucial idea is this: we should ask not only who is aware of what contents in Fission, but above all we should ask about the experience of *changes* - in our example, the changes in the colour of light. If the change of consciousness is continuous, then there must be experience of white light turning to green *and* to red. For *such* change is occurring. But by assumptions of Fission, such experience is impossible. Once these basic ideas are in place, we could run the argument within various conceptual frameworks. Even though I prefer not to use the term "diachronic co-consciousness", my ideas could be expressed in this idiom. Secondly, the argument does not presuppose any particular view about the specious present: whether there is repetition of content in successive specious presents, or whether specious presents "overlap" etc. I claim only that there must be a specious present encompassing white, green and red lights together. Otherwise, the change is not really continuous.

The idea of the argument becomes clearer when we bring in subject-talk. When we think of subjects, many of us have a strong intuition that something unimaginable seems to happen at the moment of Fission. Consider the following principle: if consciousness changes continuously, then *someone* must experience the change. Continuous changes are experienced. And at no time can consciousness be ownerless. We have to admit this if we want to take subject-talk seriously. So *who* in Fission can experience the change? Suppose it is the Pre-Fission person. But then, he will experience the white light changing to green *and* to red. But since one subject has this experience and the change is continuous and co-consciousness-preserving then just in *what* sense would R and G fail to be co-conscious? They have to be co-conscious, contrary to assumptions. Moreover, if Pre-Fission person can experience the transition to R and G, he can just as well go on to successive experiences and be co-conscious of all of them. It follows that the lack of ordinary neurophysiological connections would not disturb the unity of consciousness. Dualism looms. So perhaps Righty and Lefty experience the change? No, since in experiencing the change Righty would experience white light turning to red and green. And awareness of green light was supposed to be Lefty's only. So it follows that *no one* can experience the continuous change of consciousness involved in Fission. So either we reject the common-sense principle that someone has to be aware throughout the continuous change of consciousness, or we admit that no one could experience the change in Fission and so it is not continuous.

What remains to be explained is the illusion of possibility of fission with continuity of consciousness. On the formal level, this illusion is due to operating with poor apparatus of relations. If we operate only with synchronic co-consciousness and diachronic co-consciousness between experiences occurring at different times, then we are prone to miss the crucial ideas spelled out above. We may fail to see a reason to rule out the possibility of R being diachronically co-conscious with U and G being diachronically co-conscious with U, but R being in no way co-conscious with G. The situation on the formal level is due to insufficient attention on phenomenological level; inattention to dynamic features of experience and consciousness. Let us then consider our attempts to imagine the process of fission in all its experiential details, including the dynamic features. It seems compelling to move along the following line. Up to the moment of division, there is the experience of white light in the Pre-Fission stream of consciousness. Then, at the moment of division, in Righty-stream there is the experience of pure white light yielding abruptly to oncoming experience of red light. At the same time, in Lefty-stream, there is the experience of white light yielding abruptly to oncoming experience of green light. Didn't we just give a perfectly sound phenomenological account of continuity-preserving fission of consciousness? Contrary to appearances, no. The account is fatally ambiguous about the identity and individuation of experiences. Consequently, it makes unwarranted assumptions regarding the possibility of occurrence of certain phenomena. Is the experience of white light as yielding to red in Righty-stream, call it U*, *the same* experience of white light as before division (U) albeit in a different mode? Well, if yes, we should note that the same U is simultaneously yielding to green light in C-stream. So if it is the same U, it will have to be experienced as yielding to green light in Righty-stream.[262] So we have our argument going, and Righty- and Lefty-streams come out as one stream. Suppose then U* is different from U. But then, what makes this experience *continuous* with U, rather than just being in some respects similar to U, and such as to make the person who has it *think* she just had had an experience of pure white light (i.e. U)? U* either has to be co-conscious at some time with U or there has to be an experience of U yielding to U*. So again my argument would apply, and we could repeat the process over and over again, however finely we wanted to cut up experiences mediating between U and R and G. To sum up, the attempt to imagine Fission from which we started is at fault at several subtle points. It does not consider identity of experiences. This leads to taking *qualitative* similarity for continuity. What we imagine in fact is that in Righty-stream and Lefty-stream there would be experiences *such as* would be gotten by a normal process of continuously proceeding consciousness.[263] But we do not imagine properly the continuous process itself. We just jump from the experience U to experiences in post-division streams without attentively considering how the process of change actually looks. So, after all, we are prone not to be attentive enough to the dynamism of consciousness.

[262] Unless we want to say that one experience is parallelly experienced in two different ways. I do not find this idea intelligible. How can one distinguish the experience-in-itself and the way it is experienced? To put the point differently, the phenomenal features (such as being co-conscious with something) are *essential to* and *constitutive of* a (phenomenal) experience; they are a part of what the experience *is*. So it makes no sense to say that one experience could have incompatible phenomenal features. If, say, being co-conscious with A would be essential to experience B, then saying that B can be also experienced without experiencing A amounts to saying “An experience which is essentially experienced co-consciously with A is not experienced co-consciously with A”. Cf. Dainton (2000), 107 and Dainton (2008), 260f for arguments against intransitivity of co-consciousness.

[263] Many authors think only in terms of qualitative similarity and underlying continuity of causal mechanism, taking these to be sufficient for continuity of consciousness, or indeed, all there can be to continuity of consciousness. Making such *assumptions* renders the argument from Fission against the doctrine of indivisibility of consciousness obviously question-begging. See Unger (1990), 187-191.

7.5. Dennetian challenges

The view of unity of consciousness presupposed in my discussion is moderately popular, but not universally shared. Here I want to consider challenges coming from a very different approach to consciousness and its unity.

If one takes inspiration from Dennett, one will suspect that continuity of consciousness is to a large extent a fabrication. One need not buy Dennett's eliminativism and verificationism. It is enough to say that the appearance of continuity is mostly a matter of reconstruction, memory-links and other processes of representing the past and construing a coherent story about one's experience. Let us then explore a Dennett-inspired realist view of consciousness. On such view, our case of Fission is not very disturbing. The resulting persons would retain the memory of the white light; this memory would be utilized in construal of their experience (or the story thereof). The end result is that it would appear to post-Fission persons that they witnessed the change of light from white to red or green respectively. The weak point of this story is that it implicitly presupposes an atomic view of consciousness. A point in time is singled out at which representations of other times occur. And then there's another point in time at which other representations occur. Consciousness is enclosed in such moments. These assumptions are false. The following questions should be considered:

(i) How is it possible that we are immediately aware of changes of our own consciousness?

(ii) Can the *change* of consciousness and *experiencing* the change really be distinguished?[264]

I will address here only the first question. I take it as plainly true that we are immediately aware of changes of our consciousness; as immediately as we are aware of its contents.

[264] It may be helpful to entertain a radically non-atomist view. Call it the Process View. Consciousness is a process. What we are immediately conscious of (the contents) are changes which constitute the process which consciousness is. Experiencing is not distinct from this process. This view may be bizarre and hard to grasp; but I think it is close to truth. Zahavi discusses similar ideas formulated in phenomenology, Zahavi (2005), 58-72.

Immediate awareness means awareness without aid of a representation distinct from the object. Now, if we are to be immediately aware of changes in our own consciousness, our point of view cannot be limited to a point of view from a single point in time. In some way we have to occupy *both* the perspectives from particular times *and* a perspective which encompasses them. Here Dennett's Multiple Drafts may come to mind.[265] What may account for this multiplicity of perspectives is that smaller-scale drafts and changes in them figure in larger-scale drafts. This idea is fine, but it tells us little about what happens at the phenomenal level. Let us try to square a realist approach to phenomenal consciousness with the Multiple Drafts model. Each draft (upon meeting some conditions) will get to constitute our consciousness. The story of Fission may then run like this. There is the white-seeing draft in the brain. Then there are red- and green-seeing drafts. What about the larger draft that traces changes in smaller-scale drafts? It's hard to say *a priori.* Are its mechanisms distributed in the whole brain? If so, how was it affected by Fission? Or were there two drafts: one for each hemisphere? Is there a continuously updated general draft tracing changes of consciousness, or are there many; for instance, one draft for the persistence of white-seeing, terminated at Fission, and then two separate drafts tracing the changes from white-seeing to red- and green-seeing? It is hard to speculate, but let us take an option which seems most likely to give us something like a division of the stream of consciousness. Suppose that after Fission each hemisphere has its own larger-scale draft. Each traces the change from the white-seeing draft to a colour-seeing draft. All drafts give rise to consciousness; each presents a peculiar perspective. This, it may be claimed, is how continuous consciousness gets divided. In fact, however, questions are just relocated. How are different drafts integrated into a single consciousness? Is there a single subject who enjoys consciousness from different perspectives? Note that on Dennett's Multiple Draft model there is no specific mechanism

[265] Dennett (1991), 111-138.

linking all drafts together; no Central Theatre. That drafts are created in different areas of the brain at different times, without any overarching linking mechanism, is wholly irrelevant to *how things appear* to us. If we combine this model with realism about consciousness and the subject, we get the following view: multiple drafts give rise to a single self with an integrated, though multiple-perspectival consciousness. Physical relations between drafts are irrelevant to the way they are integrated at the phenomenal and representational levels. But then, why should cutting the brain matter? If, in normal life, activities in un-related areas of the brain get integrated into single consciousness, then why would it not happen with drafts created in cut-off hemispheres?[266] For a realist Dennetian, the plausible thing to say is that all drafts in our Fission case get integrated into a single consciousness of a single self. This self will experience seeing white light and seeing red light and seeing green light; and it will experience the change from seeing white to seeing red light and from seeing white to seeing green light. True, instead of having 3 perspectives (as in normal cases), this self will have 5 perspectives. But given that we agreed that a single consciousness may be multi-perspectival, this is not a problem. It will also be true that such self will not have a perspective according to which white light turns both to red and to green lights. Instead, it will have two perspectives telling only one side of the story. But this does not impugn on the unity of consciousness.[267] It is rather like the case of visual experiences coming apart when we cross our eyes. We find it hard to integrate these experiences into a single framework or perspective, but the experiences are plainly co-conscious. The realist quasi-Dennetian model, unlike my model, allows for there not being experience of white turning both to green and to red.[268] But even so, it leaves the unity of consciousness intact. Consider now two other

[266] Cf. Tye (2003), p. 127f.

[267] Remember that perspectives do *not* account for the unity of consciousness. Co-existence of different perspectives within a unified consciousness accounts for the immediate awareness of changes.

[268] This, and multi-perspectivalism, may be advantages of this model over the view I proposed. On the other hand, the integration and unity of consciousness and the immediacy with which a "higher" perspective takes in what is going on in "lower" perspectives are left unexplained.

Dennetian options. One could adopt non-integrative realism: drafts give rise to consciousness, but there is no integration at the level of consciousness. Each draft is a world unto itself. This kind of Homunculism is obviously implausible. If we agree that there is phenomenal consciousness at all, we should agree that we are immediately aware of its existence and of its changes. Homunculism denies that: since there is no integration, the change-tracing draft can only give rise to indirect representation of conscious events in different homunculi. This position is unprincipled and absurd. The only position left is Dennett's eliminativism about the phenomenal. Now this is a position in a quite different game. It does not say that there is continuity of consciousness which allows for division. There is nothing like continuity of consciousness, for there is nothing like phenomenal consciousness to start with. Moreover, Dennett seeks to undermine the point of departure of the whole debate: introspective reports on the unity of consciousness. The only thing to do about such radical disagreement is to see which approach will manage to live and solve various puzzles better. Let it suffice to say here that Dennett's eliminative model seems to create more problems than it solves.[269]

7.6. Reductionism and Dualism

Let us now examine the consequences of my argument. The impossibility of fission with continuity of consciousness seems troublesome for Reductionism. For all the *physical* mechanisms supporting continuity of consciousness seem to be present. Likewise, all the *physical* mechanisms supporting singleness of stream of consciousness are absent. But it is *logically* impossible for continuity to be preserved in the absence of singleness of consciousness. So whether continuity and singleness are both preserved or whether both are disrupted, there occurs something which is inconsistent with Reductionist principles of dependence of consciousness on physical processes.

[269] For a balanced estimation of Dennett's views see Seager (1999), 85-131

If my argument is somewhat troublesome to Reductionist, it is certainly highly beneficial to the Dualist cause. It is often claimed that Dualism can offer only a wholly arbitrary answer to the question of what happens at Fission. But now we are able to offer a quite plausible account. It transpired that what could happen with consciousness at Fission is altogether different from what we would be led to expect by observing the mechanisms normally supporting consciousness. We would be led to expect continuity - but this is impossible. This cries out for explanation. Dualism affords a very simple one: there is a rupture in the continuity of consciousness because the old self dies and two new selves arise. The self could not survive because it could not have two separate centres of consciousness. *Pace* Parfit, it is the belief in the principle “one subject - one consciousness” that is the main reason we are all inclined to think that Fission cannot preserve identity. So Dualism gives the intuitive answer *and gives it for the right reason.*

Still, we can imagine the following situation: the Pre-Fission person goes one way. Say, it survives as Lefty. There is continuity of consciousness between Pre-Fission and Lefty-stream. But Righty-stream is not so continuous. Righty is a totally new self, perhaps created by God at the moment of Fission. This scenario seems possible by Dualist lights. So the Dualist owes us an explanation why this would not happen at Fission.

We should note first that the Dualist has no reason to accept *anomic* embodiment. Every sensible Dualist will want embodiment to be a lawful affair. The Dualist can convert the most common-sensical Reductionist criteria of personal identity to criteria of persistence of embodiment relation.[270] The Reductonist criteria of identity may be to some extent shaped by the belief that we are material things, and such influence will have to be discounted. But the criteria are mostly shaped by intuitions of the following kind: “these processes seem very similar to processes *normally* underlying our mental life - so it is natural and plausible to

[270] There are no reasons of principle why Reductionists and Dualists should disagree about the time of *birth* and *death* of persons.

suppose that our life will flow on". Such intuitions are metaphysically neutral and available to the Dualist as well. In this spirit, we can advance the following proposal:

(E) I am embodied as long as my brain continues to exist and to have the capacity to support unified consciousness.

We could say that I continue my embodied life as long as my properly functioning brain persists.[271] This seems common-sensical and close to what we think about death, irrespective of whether we are Dualists or not. Now if my embodiment obeys such law, then it follows that at Fission I become dis-embodied. For my brain ceases to have the capacity to support unified consciousness. So there is a good reason for the Dualist stance on Fission.

The Dualist may further speculate that the facts about embodiment are relevant to *individuation* of selves. If so, it may be my necessary property to be embodied in a certain way (or being embodied in a certain way as long as I am embodied at all). Surviving as one of the Fission offshoots will be inconsistent with being so embodied. So although I can imagine this, it may not be metaphysically possible for me. This strategy is open to the Dualist, but it is not required. It is good enough to say that if things go the way they *normally* do, if no miracle occurs, then a person would not survive Fission.

Now the Dualist can turn the tables on the Reductionist. The Dualist can provide an answer to what happens in Fission which is *definite, non-arbitrary,* and *consonant* with our intuitions. On the other hand, the Reductionist (and, as far as I can see, non-Dualist non-Reductionists) will have to embrace endlessly *unintuitive indeterminacy* of identity and provide more or less *arbitrary* answers. Admittedly, the Brain Theorist could have an answer parallel to that I gave as a Dualist. For, by and large, I took over MacMahan's criterion of

[271] Cf. McMahan (2002), 92. The Dualist, unlike McMahan, need not postulate the existence of a new "brainy" object over and above the ordinary brain.

personal identity and turned it into a criterion of embodiment. Yet, the Brain Theorist would still face the puzzle of how to explain the impossibility of preservation of continuity within the Reductionist Framework. And we saw earlier that the Brain View is implausible for other reasons.[272] All other Reductionist accounts seem to have only answers plagued by unintuitive indeterminacy and arbitrariness. In comparison, the Dualist has a much better answer. Fission supports Dualism.

[272] Ch. 5, 130-134.

Chapter 8. Anticipation and the Self

In this chapter I will investigate the links between anticipation and identity. As I stated in chapter 4, the claim that anticipation of having experiences presupposes identity of the anticipator and the future experiencer is important for showing that Substantialist SV is not devoid of practical implications. Now I will defend this claim. Moreover, the consideration of the link between anticipation and identity will enable us to present a serious argument against Reductionism in general. I will argue that Reductionism undermines the possibility of genuine anticipation. This means that Reductionism is no more plausible than the No Self view and may even collapse into Eliminativism. I will focus primarily on the Logical Construction Reductionism (Parfitian Reductionism) for two reasons. First, this theory has perhaps been least touched by the objections to Reductionism I developed so far. Secondly, it throws into high relief the features of Reductionism which may be downplayed by other views.

The best way to approach the issues of anticipation is to think of Fission. Can we anticipate experiences of post-fission persons? Can we anticipate having experiences of post-fission persons even if they are *not* identical to us?

8.1. Fission, indeterminacy and anticipation

Fission has a paradoxical flavour partly because it apparently involves *indeterminacy* of identity. It is often suggested that the air of paradox is due to special features of the first-person perspective.[273] This is not quite right. The fission of Theseus' ship was troubling philosophers ages before the idea of personal fission transpired. The gist of the problem lies in the fundamental metaphysical intuition that in nature there is no place for indeterminacy of

[273] Blacburn (1997), 181; Velleman (1996), 172, 200ff.

identity and existence.[274] Identity and existence are all or nothing. This idea, when applied to my own self, results in the principle "For every time *t*, either I exist at *t* or I do not exist at *t*". It is hard to see how this application of the excluded middle principle could fail and leave a grey area in some cases. Moreover, it seems that this principle entails that for every future time *t*, I can either anticipate having some experiences at *t* (if I happen to be conscious then) or that I cannot anticipate having any experiences at *t* (due to not existing at that time). So when I am told that it is indeterminate whether I exist at some time, this is not only metaphysically troublesome, but it seems to make no sense from the first-person perspective.

Parfit's treatment of fission, indeterminacy and anticipation will be a convenient starting point for our discussion. Parfit makes three crucial claims.

(1) The question "Will I survive?" does not have a determinate answer in Fission.

Our concepts do not determine the choice between alternative descriptions of what happens in Fission. From (1) it logically follows that:

(1a) The persons resulting from my Fission *could* be regarded as me.

Since our concepts do not determine the description, it cannot be *absurd* to regard a resulting person as me; even if it is not the *best* description of the case overall. Parfit's best description is not *the true* description. There is no "fact of the matter" about my survival in Fission. Hence indeterminacy and the need to make a conventional decision about descriptions.

(2) For the fissioning person, Fission is better than death. It is about as good as ordinary

[274] Cf. Shoemaker (1997), 144: "No ordered pair of entities in the world can be such that it is indeterminate whether its first member is identical to its second. But the truth of the statement of identity can be indeterminate owing to indeterminacy in the reference of its terms".

survival.

(3) The questions "Will I survive?" and "Can I anticipate having experiences after Fission?" are *equivalent*.

It seems that for Parfit this is an obvious *conceptual* equivalence.

The conjunction of these three claims leads to discrepancies in Parfit's position. From (1) and (3) it follows that there is no determinate answer to the question "Can I anticipate having experiences after Fission?" This is something which is terribly difficult to make sense of. If I live on, then surely I am justified to anticipate having my future experiences. If I die, then just as surely I cannot anticipate my future experiences, for there will be none. But in the case of indeterminacy I am told that I neither *can* nor *cannot* anticipate the future experiences. So indeterminacy is neither like life nor like death. From the first-person perspective it is an unknown and un-imaginable *tertium quid*. Now, if I cannot confidently anticipate post-Fission experiences, then this situation is for me radically *unlike* my ordinary survival. Therefore Parfit is wrong to claim that Fission is in relevant respects *like* ordinary survival and *not like* death. And this false claim is central to his position on the importance of identity.

Of course, we can *give* an answer to the question about survival. This will be a conventional answer. According to Parfit (and many following him), the best answer is that I do not survive Fission. But then it follows from Parfit's claim (3) that I *cannot* anticipate post-Fission experiences. So according to the best answer, Fission is for me *remarkably like* death. Claim (2) is false.

The conjunction of Parfit's three claims apparently leads to a contradiction. We need to retract our steps to see what can be mended. What seems strikingly implausible is that the answer to the question “Can I anticipate such-and-such experiences?” could be resolved by adopting a convention. Rather, it is just the intrinsic relation to a future person which determines whether it is possible or appropriate for us to anticipate having her experiences.[275] At this point we need to attend to claim (1a). This is a rarely emphasized, but important claim. It is because I can regard the person resulting from Fission as me - while it is equally possible not to regard him as me - that, first, we have indeterminacy and, second, that Fission seems remarkably unlike death. If I could not regard the resulting person as me, the situation would be straightforward: Fission would amount to my certain death. Now, it seems to follow from the fact that I could regard the person as me that I could anticipate his experiences! For if I could not anticipate his experiences, then there would be no way that I could regard him as me.[276] This reasoning seems persuasive. Now, if I can anticipate having the experiences of a post-Fission person, then claim (2) becomes quite obviously true. If I may justifiably look forward to having experiences after Fission, then obviously this situation is not like death, and very much like ordinary survival. The problem of making sense of Fission from first-person perspective is also greatly alleviated.

So it seems that it was Parfit’s claim (3) which was making trouble for the Reductionist. This claim, unlike claims (1) and (2), is relatively insulated from the network of central Reductionist ideas. By rejecting this claim the Reductionist can dispose of many troubles with one stroke. This is what some Reductionists do.[277] Following the dialectic inherent in Parfit’s account, we come close to the position developed by Raymond Martin.

[275] Most of us are probably disposed, like Parfit, to accept claim (3) i.e. the equivalence between questions about identity and anticipation. Non-conventionality of anticipation partly explains why the rejection of the “Only *x* and *y*” principle of identity seem so implausible.

[276] This proposal gives two simple equivalences. I can regard a future person as me iff I can anticipate having his experiences. I cannot regard a future person as me iff I cannot anticipate having her experiences. Parfit’s original position involves more complicated equivalences: that I can regard a person as me is equivalent to it not being determinately false that he is me; and that in turn is equivalent to it not being determinately false that I can anticipate having her experiences. Analogously for negations. Finally, we could say that I should regard a person as me iff it is determinately true that he is me; this is also equivalent to it being determinately true that I can anticipate having his experiences.

[277] Martin (1998), 32-50; Velleman (1996), 200ff.

8.2. Anticipation, identification and continuation

Martin’s central account can be summarized as follows. Human anticipation tends to trace the continuity of mental life. But continuity of mental life and personal identity can go apart e.g. in cases of Fission or Radical Transformation. In such cases, people tend to anticipate having experiences of people not identical to them. In Martin’s terms, people surrogately *self-identify* with their future *continuers.* Self-identification is explained in terms of *appropriation.* Appropriation consists in adopting particular *affective, cognitive and behavioural* dispositions towards a future experience imagined “from inside”. The link between anticipation, self-identification and appropriation is presumably necessary since anticipating having experiences seems to consist in imagining future experiences and appropriating them. But the claim that our anticipation is “continueristic” does not purport to state a necessary truth. It is merely an observations of human tendencies. Martin intends to offer a *descriptive* account of anticipation. The issues of rationality of anticipation are addressed only shortly in response to objections.[278] Martin seems sceptical about the possibility of establishing *a priori* substantial restrictions on the scope of rational anticipation.[279]

There is much to recommend the view that anticipation primarily traces a continuity relation. The claim that anticipation presupposes *some kind* of self-identification sounds very reasonable. It can be said that this link explains why it seems that anticipation presupposes identity. But when it is seen that anticipation presupposes identification rather than identity, the conceptual space opens up for the divergence between anticipation and identity. Thus the view can explain away opposing intuitions. Now consider the claim that we can identify with our continuers. Even if my continuers are declared to be distinct *persons* from me, there is a genuine continuity between our *lives.* It even seems natural to say that

[278] Martin (1998), 39-50.
[279] Martin (1998), 43, 48.

they take up *my* life; that there is one life involving a switch-over of persons.[280] This is an intimate relation, which certainly gives ground for some kind of identification. We can also observe that we extend empathy, and adopt other dispositions which are markedly similar to self-concern, to persons if (and perhaps in proportion to) our relation to them is similar to that between me and my continuers.[281] [282]Such people as one's loved ones, friends, disciples and especially children, share many features with one's continuers.[283] The hypothesis that our identificatory tendencies are, as Martin puts it, "at most continueristic" can explain the extension of empathy and other dispositions as a natural extension of our self-identification and self-concern. This also explains the feeling that we relate to special people in a very similar manner as to ourselves; we regard them as our alter egos or we do not treat the two of us as fully separate beings. On the other hand, the hypothesis that anticipation and appropriation (and consequently self-concern) traces identity seems to face difficulties in this area. For it makes precious little sense to say that my relation to certain others is somewhat like identity. To sum up:

(i) anticipation seems to presuppose self-identification

(ii) self-identification arguably consists in the appropriation of experiences and actions

(iii) there is continuity between appropriative attitudes we have towards ourselves, continuers and yet other people

[280] It is also possible to think of a family as having a single life (or history) in this sense; or think of family members as sharing one life. It is especially tempting to think in these terms about parent-child relationship. By *passing on* life, the ancestors may be said to escape death. Aristotle's ideas on this subject could be viewed from this angle.

[281] It is worth noting that the ancients used a single technical terms *oikeiosis* which apparently covers all empathic and identificatory dispositions and attitudes. The relation of *oikeiosis* applied both reflexively, to the individual itself (what we would call self-identification and self-concern) *and* to others (what we would call empathy, other-concern, broad egoistic concern, and a kind of identification). The two types of *oikeiosis* are not merely similar: they may be taken to be essentially the same. See Annas (1993), 262-290 and Sorabji (2006), 43f, 104f.

[282] I do not claim that *all* dispositions involved in anticipation may easily take other people as objects. Martin lists dispositions which do not seem easily extendable in this way; Martin (1998), 112-118.

[283] Cf. Whiting (1986).

(iv) it is not absurd to extend appropriation and by that token some kind of self-identification and (by that token?) anticipation to continuers and perhaps yet other people.[284]

Finally, when we notice that the continuity of life can go apart from identity, it is not obvious that anticipation (and memory) trace identity rather than continuity of life.[285] And that is enough to undermine Parfit's claim (3): the questions about anticipation and about identity are not straightforwardly analytically equivalent.

Martin's theory can be yet improved. Since Hume, Psychological Theorists tried to substitute causation and resemblance for the flow of mental life. But if we want to stay on the intuitive foundation, it is better to hold to the natural idea of the flow. To remember is to remember where we come from - and to anticipate is to anticipate where we are going. To put it non-metaphorically: in remembering something past, I remember something that was present to somebody; and in anticipating I anticipate something that will be present to somebody. Now, we are immediately aware of our "travel" through time by being aware of the flow of our conscious life. This flow can be understood in terms of dynamic succession and co-consciousness. Yet we can (so we think) continue to exist (and "move forward in time") also when unconscious. I claimed that potentially continuous consciousness is a necessary condition for identity of the person existing at these times. With equal reason we can say that it is a necessary condition of persistence of a person's life. As such, it will also be a necessary condition for appropriateness of anticipation under the current account.

This proposal seems to have the best intuitive support among the proposals on the table. First, it is a re-formulation of a very widespread intuitive idea: that anticipation

[284] Note that (iv) does not follow from previous claims. Even if self-identification could be extended to other people, anticipation perhaps could not: it could require something more than self-identification. But it is characteristic of Martin's approach that he thinks there are no substantial restrictions on the scope of anticipation.

[285] In Fission it seems natural to say I can anticipate the life of my continuers and that they can remember (not just "remember" in some quasi-sense) my life.

presupposes the possibility of “getting” to experiences (or to a time when they occur). Secondly, since under this proposal anticipation traces persistence of a person’s life, it can claim the same intuitive support as Martin’s proposal. But it remains closer to our actual everyday way of thinking. Unlike Martin’s, this account is committed neither to the artificial Humean re-interpreation of the idea of the flow of mental life nor to any controversial claim about anticipation being “continueristic rather than egoistic”.[286] Although I put forward this proposal now, it is *not* presupposed by my arguments against the Parfitian Reductionism in the next section.

8.3. Appropriateness of anticipation and the Argument from Excessive Extendability of Anticipation

Martin’s account, and my consciousness-based account, provoke the following question:

(Q) Could I anticipate having experiences of *anybody*?

Given that the scope of my anticipation is not restricted to *my* future experiences and actions in the way suggested by Parfit’s claim (3), is it restricted in some other way? Martin shortly addresses this issue by asking about the limits of *rational* anticipation. Yet it is unclear what is meant by “rationality” of anticipation and how one would go about arguing for or against rationality of particular acts of anticipation. Martin says that the distinction between rational and irrational anticipation is the analogue of the distinction between genuine and seeming memory; but the way he goes on to treat rationality makes the two distinctions in no way

[286] In and of itself this proposal does not determine the solution to the question whether we can anticipate having experiences of people not identical to us. Still, it provides a clear criterion for appropriateness of anticipation. I cannot appropriately anticipate having experiences of any random person. There has to be a possible link of conscious experiences between us. It is clear that *at most* my continuers could be so related to me.

analogous.[287] The unclarity of what is meant by "rationality" here may well be one reason why Martin does not see a way to seriously restrict the scope of anticipation. It may also be the reason why debates on "what matters in survival" into which discussion of anticipation tends to be embedded are so muddled. The debates on "what matters in survival" tend to be focused on egoistical concern and its justification. Anticipation is certainly central to our egoistical concern.[288] I think, however, that the real importance of anticipation for matters of identity is obscured by the context of "what matters in survival" debate. This is what happened with Jim Stone's argument, which I take as the starting-point for addressing the question about the limits of anticipation.

In *Parfit and the Buddha,* Stone presses a traditional problem for Reductionism: whether egoistical concern, and other seemingly identity-presupposing attitudes (like remorse, pride, anticipation) are justified on the Reductionist view.[289] In the case of anticipation, the argument goes as follows:

> **(1)** Personal identity is conceived as a relation which *justifies* my anticipation of having experiences of a future person.
> **(2)** Personal identity as analyzed by the Reductionist does not justify anticipation.
> **(3)** Therefore the Reductionist analysis of personal identity is inconsistent with the way we conceive personal identity.
> **(4)** Therefore if Reductionism is true, there are no persons.[290]

[287] Martin (1998), 44: "when it comes to anticipation, the only distinction we want to preserve that might be analogous to that between genuine and mere seeming memory is that between rational and irrational anticipation. And there are many possible ways of preserving the latter distinction. For instance, we can preserve the distinction between rational and irrational anticipation simply by saying that anticipations are irrational if anticipators should have known they had insufficient evidence that the experiences would occur". Where is the analogy? I see none. I have no grasp of what is meant by "rational" here; and I see no such distinction anywhere at work in our everyday thinking. There is, of course, a distinction between anticipating probable and improbable events; but this is irrelevant to our topic. The relevant distinction has to be introduced either by means of an overt definition or by means of an analogy to the memory-related distinction. Martin does neither.

[288] Velleman (1996), 194f: "What we most want to know about our survival, I believe, is how much of the future we are in a position to anticipate experiencing. We peer up the stream of consciousness, so to speak, and wonder how far up there is still a stream to see."

[289] The claim that Reductionism makes egoistical concern unjustified has been called by Parfit the Extreme Claim. It has been clearly formulated by Butler; see Butler (1736/1975), 102. The best defenses of Extreme Claim can be found in Stone (1988) on the Eliminativist side; and Haksar (1991), 158-185, on the Simple View side.

[290] Cf. Stone (1988), 529.

I will call this argument *Stone's Reductio.* Obviously, one can debate whether premiss (2) is sound. But there are more important prior questions. The most important of these is: what "justified" means in this context? Parfit and Stone think of justification of anticipation along the same lines that they think of justification for holding people responsible for past actions or of justification of special egoistic concern.[291] So the problem is whether an attitude is *rational* or *ethically justified.* But on such reading Stone's Reductio does not pose much of a problem to Parfitian Reductionism. Suppose that the argument is sound - a possibility Parfit himself is willing to entertain. On this reading of "justify", the argument does not really show that Reductionism is *false*. The most that follows from the argument is that on the Reductionist analysis, some of our attitudes come out unjustified; and that we deeply believe in anti-Reductionism about selves. But then, Parfit (unlike many other Reductionist) admits at the outset that we have ingrained Cartesian intuitions. That is also why he is a revisionist in ethics. For on the one hand, the true view undermines our old attitudes. On the other hand, it may be necessary to change our practical attitudes in order to be able to believe the true view. Now Stone apparently wanted to block such line of defense by claiming that we *define* persons as subjects of responsibility and rational anticipation:

> "as Locke puts it, "person is a forensic terms, appropriating actions and merit". Persons are *conceived* as responsibility bearers, beings which carry rights and obligations through time, capable of rational hopes, fears and regrets. Persons are *essentially* morally interesting".[292]

He then tries to show that even the basic Lockean definition would not be satisfied under Reductionism:

> "If at t1 I recognize that the intelligent being at t2 is identical to me, then if I know he will have an experience E, then it is rational for me to anticipate experiencing E. If Reductionism is true, the consequent is always false..."[293]

[291] Parfit (1984), 312; Stone (1988), 529f.
[292] Stone (1988), 530.
[293] *Loc. cit.*

Stone locked on a central issue here, but the framework in which he operated vitiated the argument. The Reductionist has good lines of defense here. First, if the relevant meaning of “rational” is the same as in the case of rationality of responsibility, then the Reductionist could agree that anticipation is irrational. But what, on such reading, is Stone’s argument for the validity of the conditional: “if I recognize that x is identical to me and x will experience E, then I am rational in anticipating E”? Stone offers no argument for this. A consistent Reductionist could simply deny the truth of the conditional. Thus Stone would have no convincing argument to show that the antecendent is false, and that thereby the Lockean definition is never satisfied. But then, why did Stone not offer an argument for the conditional? There may be two reasons. First, the conditional may be a consequence of his claim that persons are *essentially* morally interesting. But then, Stone offers little support for this claim. A hard-headed Reductionist can reject it and stick to a morally neutral definition. But it is more probable that Stone took the conditional as *self-evident*. Yet then the Reductionist has an even better response to Stone. The conditional is indeed self-evidently true. But that is because anticipating the experiences of a person identical to me is rational *by definition*. Anticipation and identity are conceptually linked.[294] One could even take Johnstonian line and say that the analysis of “identity” and “person” is altogether irrelevant to the rationality of anticipation: whatever the analysis of the concept, anticipation locks on the thing which satisfies the concept. Note, finally, that even if our actual concept of persons would include substantial moral elements, the fact that nothing would satisfy *this* concept, would not yet mean a final defeat for Reductionism. For, if our concepts turn out to have uncongenial features, we can modify them; we can forge new concepts like person*.[295] And it is not obvious that our moral concerns are so predominant that if they could not be satisfied,

[294] This need not mean that rational anticipation always traces identity. If, for example, rational anticipation is continueristic, then since my future self is my continuer, it is rational to anticipate his experiences. Or if rational anticipation *co-constitutes* identity, then again, anticipating experiences of my future self is automatically rational.

[295] See ch. 2, 21.

we would abandon the talk of persons or at least persons*. We have classificatory, metaphysical, epistemological, religious, psychological and practical interests as well. For all these reasons, we could still discern in the Reductionist world certain wholes and call them "persons" or "persons*", while giving up some of our ethical ideas.

On the normative reading of "justification" of anticipation, Stone's Reductio fails to present a fatal problem for Reductionism. However, the normative reading is not the right reading. In the first place, the relevant analogue to anticipation is not responsibility or egoistical concern or any other practical attitude; it is *memory*. Now, there is a distinction between correct and incorrect memories. Even more importantly for present purposes, there is another distinction: apparent memories may either be *genuine* memories or *seeming* memories. Anticipation can also be correct or incorrect. And corresponding to the second distinction, there is the distinction between *appropriate* anticipation and *inappropriate* anticipation (or *seeming* anticipation). It is this distinction, I claim, which is really relevant for Stone's Reductio.

Let me elaborate on the analogy between the memory- and anticipation-related distinctions. Genuine memory of an experience is distinguished from any old imaginative state about the past by the presence of a relation, call it Q, obtaining between the rememberer and the person who had the remembered experience. How Q is to be analyzed is a matter of debate which will not concern us here. If the distinction between appropriate and inappropriate anticipation is to be analogous to that between genuine and seeming memory, its logical *form* should be the same. Thus I say that an appropriate anticipation of having an experience is distinguished from any old imaginative state about the future experience by the presence of a relation, call it S, obtaining between the anticipator and the person who will have the experience. Of such S we will say that it is *traced* by anticipation and that it *restricts* anticipation. Tracing S is essential to anticipation. For it is the presence of S which distinguishes genuine anticipation from any old figment. However, we should not think that in anticipating we must think in terms of S or imagine S holding between us and the imagined person. We could have a tendency to imaginatively trace a different relation S* which, however, could not diverge from S where the impossibility involved would be of a conceptual kind.

The question which needs to be posed now is whether Parfitian Reductionism can account for the distinction between appropriate and inappropriate anticipation. In other words, whether it can provide a credible candidate for the relation S that restricts the scope of anticipation. I will argue that it cannot, and this reduces this view to absurdity. But first, let me state explicitly the requirements I put on any account of S.

(i) The existence of competing accounts indicates that the claims that the account of S follows from a straightforward analysis of the meaning of "anticipation" are not credible.

(ii) Nonetheless, the selection of S and the distinction between genuine and non-genuine anticipation should *not* be purely *conventional*. It should not be possible to get a different scope of anticipation by trivial tinkering with definitions or conventions. This leads to the third postulate:

(iii) It should be *logically impossible* to extend anticipation beyond its proper scope. Consider this. Could *I* anticipate now the experiences *you* will have in 2 minutes from now? The idea is strikingly *absurd*. Indeed, it seems as absurd, as absurdity can get. The account of anticipation should capture this feature (or at least explain it away in a strikingly plausible manner). It should rule out the possibility that we could have an attitude which would be just like anticipation, but would have a different scope and would not get the label "anticipation".

There must be a difference *in kind* between genuine and non-genuine anticipation; a difference due to the nature of S.

(iv) Finally, the selection of S should not be *arbitrary* in the sense that it should not be wholly irrelevant from the standpoint of phenomenology and actual practices of anticipation. Proposing, say, that “having the same DNA” is what restricts anticipation, would violate this requirement even if it satisfied the remaining ones.

The stage is set for the argument:

Argument from Excessive Extendability of Anticipation

(1) If anticipation is just a matter of adopting a particular emotional/practical attitude towards a future person, then it can be extended without restriction.

(2) Identity cannot restrict anticipation in Parfitian Reductionism.

(3) No psychological or physical relation can restrict anticipation in Parfitian Reductionism.

(4) Therefore Parfitian Reductionism can offer no plausible candidate for the relation restricting anticipation.

(5) If anticipation is unrestricted, the appropriate-inappropriate distinction collapses.

(6) Therefore my relation to my future self, as far as anticipation goes, is the same as my relation to *any* other future self.

(7) Therefore the concept of genuine anticipation is inapplicable under Parfitian Reductionism.

□

This is my general argument. Now, the premisses need to be argued for.

On Martin’s account, the scope of my anticipation is only contingently limited by my contingent psychological tendencies for adopting the emotional/practical stance of appropriation towards future experiences of some persons and not others. As Martin suggests himself, it is hard to see any logical limits to extending anticipation when conceived this way. There are three considerations showing that this suggestion is correct. *Identification* plays a crucial role in each, since, as we remember, it is the identification which underlies anticipation on Martin's account. First, the reason people may feel inclined to surrogately self-identify with their continuers, and hence anticipate their experiences, is that the continuers, even when they are not linked by strong psychological continuity, are intimately related to the anticipator.[296] As I suggested, they are imagined to continue the person's *life*.[297] But then, my children, my students and countless people living after my death can also be intimately related to me, can continue my work and carry on my ideals, thus continuing my life in a way not inferior to that of a continuer. Why then should I not anticipate having their experiences too? Secondly, consider identification with all living beings or with the whole of reality; a recurrent theme in many philosophical traditions: Stoicism, Taoism and Buddhism to name a few. If people ever attain sagehood or enlightenment, then adopting the same practical stance towards all future mental events is not only a theoretical, but a practical possibility. Now, if there is nothing more to anticipation than the stance of appropriation, then it follows it is possible to extend one's anticipation to all future mental states. Thirdly, consider an object composed of *all* mental items occurring anywhere and anytime in the Universe. Call it *Cosmic Consciousness*. Cosmic Consciousness is a perfectly respectable object by Parfitian Reductionist lights. It is a logical construction on a par with (ordinary) persons. I can identify with Cosmic Consciousness. I can regard the events I am conscious of here and now as but a part of a larger whole which is me. But if I identify with Cosmic Consciousness then I can anticipate all future mental events. The Parfitian Reductionist cannot claim that such identification is defective. One way to argue against identification would be to claim that "I" can refer only to persons and that Cosmic Consciousness is not a person. I see no good reasons to accept this claim about "I".[298] It would be plausible if self-

[296] Martin (1998), 91f, 127f.

[297] Life is a flexible and comprehensive concept as noted (see ch. 4, 95f). Life can be continued on the account of general family resemblance even if strong psychological continuity fails.

[298] Cf. Olson (2007), 37ff.

consciousness were taken as sufficient for personhood. But then, Cosmic Consciousness is self-conscious: in the acts occurring in the minds identifying themselves as parts of it. Moreover, identification taken as emotional/practical stance does not seem limited to persons defined in any robust way. I could believe that if I become very senile and Alzheimer-ridden, I will no longer be a person. But there will be my sub-personal continuant with whom I can identify and whose experiences I clearly can anticipate. Another reason to criticize the identification with Cosmic Consciousness is its lack of the unity of consciousness. But this is not a factor for the Parfitian Reductionist. I can identify with a whole composed of disunified streams - this is the central argument in Parfit's *My Physics Exam* case. Likewise, it is not *absurd* to regard the offshoots of Fission as one person. The *Branch-Line* case is finally most instructive. What is most noteworthy about Parfit's treatment of Branch Line is the claim that there is no significant difference between my relation to my *Replica* in the future and my relation to *myself* in the future. This is a telling illustration of Parfit's view of persons. Persons are logical constructions. I could regard my self in the future as a different person from me-now; and there would be no *fact of the matter* to prove me wrong. Nor could I be wrong in identifying with a whole larger than a human mind. This is a consequence of Parfit's fundamental view: there are no deep facts about the unity and separateness of persons. There are just so many *discrete* mental events which are interrelated and can be *regarded* as a unity. On this general view, anticipation apparently can *only* be a matter of our imaginative and affective dispositions. Its scope and limits are not fixed by any deep facts, but only by our contingent, and perhaps revisable, psychological tendencies. If *identification* (and hence anticipation) were robustly restricted by *facts,* then so would be identity; and that would undermine the whole approach.

Consider now the prospects of restricting anticipation by identity. Given that anticipation is not restricted by any deep facts of the matter, the view that anticipation and identity are linked *analytically* is all the less surprising. As we saw, this view is not above doubt. Furthermore, it creates inconsistencies in Parfit's account.[299] And if this view were right, the Branch-Line case could be immediately turned against Parfit. If anticipation is analytically tied to identity, then it is impossible for me to appropriately anticipate the Replica's experiences. But my relation to my future self is not importantly different from my relation to my Replica. Therefore I cannot appropriately anticipate even my own future experiences. Finally, even if anticipation was tied to identity, the possibility of identifying with Cosmic Consciousness would still remove any restriction of the scope of anticipation.

The Cosmic Consciousness argument applies in fact to any relation between mental events: since the Cosmic Consciousness contains all mental events, there are no events which are un-related to it. This argument may not persuade everyone. But there are other decisive considerations. Suppose that a psychological or physical relation R is proposed as the relation that restricts anticipation. How can this claim be established? Anticipation cannot be tied to R analytically - if anticipation were analytically tied to anything, it would surely be so tied to identity. So the claim would have to be established by the investigation, phenomenological or otherwise, of our actual acts of anticipation. Now, suppose that the investigation revealed that in anticipation we tend to track some kind of psychological continuity. But would that determine any definite relation? This is not very likely. There are many definitions of "psychological continuity"; and whenever an author tries to be even slightly precise (Parfit is one of the few examples), the definition is patently and overtly stipulative. What is the consequence? The consequence is that any claim that a particular well-defined relation R restricts anticipation will unavoidably be *arbitrary*, as there will always be indefinitely many similar relations with equally good claims. This means that my requirements (ii) and (iii) will be violated. Finally, one cannot argue in the following manner:

[299] See above, 175.

“Anticipation must be restricted by something. It seems that in anticipation we tend to track some kind of psychological continuity. Therefore anticipation is restricted by some psychological continuity relation, only we cannot say exactly which.” This kind of inference to the best explanation could perhaps be allowed if there were no other plausible candidates. This is actually not the case. But even if psychological continuity was the best contender, its claim would still be undermined by the following consideration. Psychological continuity is always a matter of *degree*. And this makes for a slippery-slope. If it is not absurd to anticipate experiences of continuers strongly psychologically continuous with one, it is not absurd to anticipate experiences of continuers only weakly continuous with one. And if it is not absurd to anticipate experiences of such continuers, it is not absurd to anticipate the experiences of one’s Replicas and children and so on. My point is that any cut-off point in this spectrum a Parfitian Reductionist would wish to make will be *arbitrary* and *conventional* in a noxious sense (i.e. the requirement (ii) will be violated). Nor will she ever be able to explain the absurdity inherent in the excessive extension of anticipation (requirement (iii) violation).

We are now done with the first three premisses of my main argument. Are there any available candidates for S left? There are: continuity of consciousness and potential continuity of consciousness. But these, I take it, are not available to the Parfitian Reductionist. If one goes for a Consciousness-Based account, there is no reason to keep to logical construction ontology. The unity of consciousness furnishes quite solid facts of the matter as the basis for identity and anticipation. Consciousness-Based fits with Constitution View PT, or with Mixed Psycho-Physical Approach, or with the Simple View. Thus, given the truth of premisses (1)-(3), claim (4) is established.

What, then, happens if there is no S to logically restrict anticipation? The possibility of its excessive extension makes nonsense out of our idea of anticipation. If I can anticipate future experiences of just anybody, then there is no distinction between genuine and seeming anticipation. I claim that this lack of distinction entails *inappropriateness* of all anticipation. First, note the absurdities to which unrestricted genuine anticipation would lead. Extending the scope of anticipation offers a cheap way to cheat death. If I can anticipate *having* various experiences occurring after the demise of my physical body (e.g. by identifying with Cosmic Consciousness), then, I submit, it is just not true that I will die. On Naturalist assumptions this consequence is properly absurd, if anything is. This shows again, and that is the second reason for my position, that there is a deep difference between the anticipation of having *my own* future experiences and "anticipation" of experiences of other people.[300] But Parfitian Reductionism has no place for such difference. It assimilates all anticipation to the second model. Neither in attitude nor in reality is there a deep difference between my relation to my future self and my relation to my Replica. In both cases, there are just discrete events which can be regarded as unity or not; there is no deeper fact which would make my future self significantly closer to me than the Replica. Finally, and that is the decisive consideration, the distinction between genuine and seeming anticipation is as *essential* to the concept of anticipation as the analogous distinction is to the concept of memory. Without the distinction, anticipation ceases to be the analogue of memory which it manifestly is. Indeed, it ceases to have any use whatsoever. It is not that we simply *want* to have such distinction. We want to have it because we would not have the concept without it.

8.4. Objections to the Argument from Excessive Extendability of Anticipation

A complex response to my claims could run somewhat like this. "The relation between anticipation and identity is perhaps not trivially analytical. But your argument, if nothing else, shows the danger of abandoning identity as the restricting relation. Taking identity to restrict

[300] It could be said that this is just a deep belief we have; and Parfit is ready to admit that we have anti-Reductionist deep beliefs. But then, this deep belief is special. Its truth is a presupposition of the appropriateness of anticipation. So if it is wrong, all anticipation is inappropriate. And then, as I argue, by *Parfitian* principles, there are no persons. Let me put it this way. The concept "person" could survive the demise of the idea of *responsibility* or perhaps even *agency*. But it cannot survive the demise of the idea of *anticipation.*

anticipation seems concordant with our actual practices. So why not say that identity is the *best candidate* for what restricts anticipation? Secondly, from the fact that identity, identification and anticipation are not rigidly restricted by facts - there are cases of indeterminacy - you cannot infer that they are not restricted in *any* way. They are restricted by practices and, yes, conventions, but these are not wholly arbitrary in that there are definite limits to tinkering with them. What are these limits? First, the patterns of our concern, identification and anticipation may be contingent, but they are evolutionarily entrenched. They cannot be changed at will. These patterns are not set by shallow conventions. Apart from their being evolutionary hard-wired, it may be the case that the whole system of our practices, morality etc. requires the existence of certain patterns and certain conventions. Our practices an concerns would crumble if stretched too much. Thus although the actual pattern of our identifications etc. may be contingent, it is nonetheless a *deep fact* in a sense. Thirdly, the reason why the scope of "anticipation" cannot be extended too far is not very different from that why the scope of "heap" cannot be extended too far. That's how we use our words, for such-and-such practical reasons. The point of inadmissible absurdity is reached when the concepts and practices crumble. So what should we say of Cosmic Consciousness and the associated "anticipation"? Well, these are so remote from our ordinary ideas about identity, self and anticipation that we hardly have any grasp on what is meant. Of course, one can always introduce quasi-selves and quasi-identification and quasi-anticipation which are in some way like normal selves and anticipation, but so what? They do not satisfy *our* normal concepts "self" and "anticipation". On the other hand, if we have an analysis of what we mean by "identity" and "anticipation", and we have a good candidate for the relation restricting anticipation in the actual practice (identity), then what is lacking? Due to indeterminacy of identity, your conditions (ii) and (iii) cannot always be met - but this only shows they are too strong. But requirements (i) and (iv) could be met; as well as *revised* requirements (ii) and (iv) forbidding extension of anticipation beyond the bounds set by the point of collapse of our concepts and practices."

I do not think this responds to my arguments. First, let me note how my slippery-slope arguments differ from the ordinary Paradox of the Heap. The answer "that's how we use our words, for such-and-such practical reasons" is good in the case of the heap for two reasons. First, "heap" is a vague and conventional notion; and it is made to be such. This reflects the nature of practical reasons which motivate the use of this concept. These practical motivations are *not* just the desire to use the notion like "heap". Now, "anticipation" is wholly different in these respects. The assumption that "anticipation" has definite limits is unavoidable. Indeterminateness of anticipation is unintelligible. That's how the concept is made. Secondly, there is no more fundamental reason to use the concept "anticipation" than to reflect, well, the practice of anticipating. Two consequences follow.

First, if the concept is made to be sharp and not adjustable by convention, then it is *reason-responsive*. What I mean by that is that our belief about the scope of the concept will be responsive to the reasoning of the form: *x* is A, *y* is relevantly *similar* to *x*, therefore *y* is A. In other words, concepts like "anticipation" do not allow for distinctions without a difference. The slippery-slope arguments show that on Parfitian Reductionist assumptions limiting the scope of anticipation will involve distinctions without a difference. Since "anticipation" is made to be well-delimited, non-conventional and reason-responsive, my requirements (ii) and (iii) are justified.

Secondly, consider what is *explanatorily* prior: the acts of anticipation and their limits, or the system of our linguistic and moral practices? It is clear that the acts are prior. Suppose someone said: "It is true that you can have an attitude to certain experiences which is remarkably like anticipation indeed. But, our moral and linguistic practices do not allow us to *call* it anticipation. So you cannot anticipate them". Well, how much do I care about preserving moral and linguistic practices? In comparison with anticipation - very little indeed. If I thought it makes any sense to extend the scope of anticipation, I would certainly do so, in order to push away the prospect of death. I think most people are like me in this respect. It is because we have fundamental intuitions about death, the scope of anticipation and separateness of persons that our moral and linguistic practices have the shape they do and not the other way round.[301]

We should be clear about the datum we want to explain (or explain away). This is: why does it seem *absurd* to extend anticipation beyond its normal scope? Compare it with the question: why does it seem *absurd* to say that 2+2=5? Now, saying: well, you just very strongly tend not to anticipate some experiences (or count 2 and 2 as 5) is *no* answer. I know I tend not to, but I ask why I *could* not. One could say that the sense of absurdity *consists* in the practical (or psychological) impossibility to adjust the practice. Well, this is as credible as the old psychologistic account of the necessity in mathematics. It is moreover plainly false, because I *can* consider changing my practices of anticipation - I just feel it would be dead wrong, and necessarily so. A serious attempt at explanation would be something like this: "In mathematics, you can define a language and models such that a sentence "2+2=5" in this language is true in these models. But you cannot do it while holding to *our* meaning of the terms. If you hold to our meanings, then any attempt to say that 2+2=5 will land in outright inconsistency. Now, psychology and morality are not as exact as mathematics. There may be some leeway. But still, you cannot hold on to *our* meaning of the terms while radically changing their scope. You cannot coherently and intelligibly say that anticipation could be

[301] Consider also the inverse situation: non-standard *limitation* of the scope of anticipation. If the acts of identification and anticipation were conceptually (and ontologically) prior to personal identity, then I could avoid future pain by dis-identifying with the future person (and given that the tortured person's psychology may change dramatically, I may be entitled to do so by PT lights). But here Unger's Avoidance of Future Great Pain Test has rightful application. I think it is obvious that we cannot get off the hook so easily. It is true that we might say "this person will be so different from me that I cannot anticipate her experiences". But this inference is enthymematic. What fully expresses our intuition in such case is rather this: this person is very different from me *therefore* he is not me *therefore* I cannot anticipate her experiences. The assessment of facts of personal identity is logically prior to the judgment about the possibility of anticipation. I know of no other meaningful way to argue whether anticipating having a certain experience is possible or not.

extended *too far*". This is a fair attempt, but it fails. Consider another mathematical analogy. Suppose that we normally mean by "numbers" objects in the standard model for arithmetic. But there are non-standard models too, which may *slightly* differ from the standard model. Are objects in a non-standard model numbers or not? It is natural to say that they *are* numbers, because there is *not much difference* between them and ordinary numbers (the concept "number" is reason-responsive). The case for extension of the scope of anticipation is like that. One purpose of the slippery-slope arguments is to show that we do not lose *intelligibility* in extension, because we proceed by steps of negligible differences between attitudes and their objects. Why then should the result be claimed to be unintelligible? There is no logical *incoherence* in the alternative practices. As the analogy to numbers shows, the mere fact that the scope of "anticipation" is different is no ground for the claim of unintelligibility. So in the end we are left with the *feeling* that somehow we went wrong - but that is precisely the feeling of absurdity which was supposed to be explained.

It could be claimed that alternative practices are not intelligible in that they undermine the whole system of our concepts and moral practices. I already noted this is explanatorily defective. But it is also just false. Consider a less extreme case than Cosmic Consciousness: a family. Imagine a society in which the family is the basic moral unit. Moreover, members of the families never separate, always deliberate and take decisions together. If separated, they cannot act effectively. They emotionally identify with the family. Finally, they have just one personal pronoun where we have "I" and "we". For all that, these people have the capacity to think, feel and behave just like us. This is a not-too-fanciful extension of actual cases.[302] Now suppose the members of the family take it not only as the moral unit, but also as *anticipatory* unit. Now, their conceptual and moral system is different from ours; but it seems perfectly intelligible and practically possible. So where is the

[302] Rovane argues for admission of group-persons into the Reductionist ontology on similar grounds; Rovane (1998) 137-142.

unintelligibility? Where the *absurdity* of the family-oriented acts of anticipation in the Reductionist framework?

8.5. The Argument from Experiential Projection

I now want to present another argument against Martin's version of Parfitian Reductionism. I call it the *Argument from Experiential Projection.* At the end of his book, throughout which he strives to show that anticipation does not presuppose identity, Martin adds a delightful twist. After discussing some singularly important ways of experiencing the self - in particular, what he calls "the perceiver-self phenomenon" - he returns to the issue of distinguishing anticipation from mere imagination. The results are surprising:

> What I am suggesting is that the three identificatory dispositions [...] might take a continuer as their object, even though the person who has this dispositions might - at the level of theory - regard this continuer as an other, *because at the level of experience,* which for the operation of these dispositions might be the level that matters, *the person whose dispositions these are experiences this continuer as himself.* If that were true, it would explain something that is otherwise puzzling, namely, why normally self-regarding dispositions suddenly cross the boundary between self and other, taking an other as their object.[303]

> Our ability to project ourselves at the level of experience [...] could also explain what our anticipation of having humdrum experiences consists in. [...] That is, it may be that in someone's anticipating having the experience of brushing his teeth, he imagines that the same subject of experience that will be involved in the teeth brushing is the one he currently experiences.[304]

This hypothesis of "experiential projection" seems thus explanatorily powerful, and Martin admits that he cannot rule out its correctness. But even if it were right, this would be a pyrrhic victory for traditionalists, Martin says.

> For on the account under discussion the ultimate basis for our so-called survival values is an experiential illusion. The perceiver-self phenomenon involves our taking part of the content of our experience - what we have called the "observed" - and treating it as if it were a subject of our experience - the "observer". It isn't the subject of our experience.[305]

[303] Martin (1998), 150.
[304] *Ib.*, 151.
[305] *Ib.*, 153.

Martin fails to draw an obvious conclusion. If anticipation *consisted* in experiential projection, and this projection were illusory, it would follow that *no* anticipation is ever rational and appropriate. Or, to put it differently, since *all* anticipation involves illusion, there would be no difference between my anticipating *my* future experiences and my anticipating *yours* future experiences. Both acts equally involve an illusion. There is no way to draw the appropriate-inappropriate distinction. This is the Argument from Experiential Projection.

I think Martin is wrong in claiming that "experiential projection" involves an illusion. But he is right in presenting the "experiential projection" hypothesis as highly plausible. This hypothesis is not very different from my view on imagination and on sentences of the form "I will *F*". I maintain that we implicitly ascribe an imagined state of mind to the very subject who presently does the imagining.[306] [307] This is tantamount to implicitly imagining the imagined subject being *me*. Since I claim that this ascription occurs both in imagination and anticipation, I need something else to distinguish anticipation from imagination. And this may simply be the thought (perhaps implicit) that the imagined situation will or could (probably) *really* happen to *me*. Martin appears to say much the same thing. Since the relevant thoughts may be *implicit* in the act of imaginative anticipation, they may well occur on what Martin calls "experiential level", as opposed to the level of overt theoretical beliefs.[308] Since identification-judgment occurs on this level and it is a precondition for adopting the relevant practical attitudes toward a future person - and we do in fact adopt such attitudes - it follows that it belongs to the sphere of *deep beliefs*. We deeply believe that any person whose experiences we anticipate is identical to us.

[306] While anticipation can be regarded as "looking into the future" it can just as well be regarded in Augustinian spirit as "making the future *present*". The ascription of the imagined state to the present subject explains how the future is made present. The present self can be regarded as an actor re-presenting the future situation.

[307] The fact that "I" in "I will *F*" manifestly refers to the present subject does not prejudge issues in personal identity or philosophy of time. For the locus of the debate can shift to the meaning of "will" in "I will *F*".

[308] The thoughts *need not* be implicit. Why could not the anticipation of a humdrum experience consist simply in the thought expressible by "I will brush my teeth?".

One virtue of my account of personal identity is that it brings together two attractive accounts of anticipation: the "flow-of-life" consciousness-based account and the "experiential projection" account. On my account, the conditions "this could really happen to me" and "this could really be linked by a flow of consciousness to my present state" are materially equivalent (and necessarily so). But which of these accounts is more true to the phenomenology of anticipation? I am inclined to favour the experiential projection. The extension of special concern to others can be explained by the fact that the same mechanism of self-projection operates both in anticipation and in empathy.[309] But it also seems true to me that when we *reflect* on our existence in time and try to spell out the connection between our present and future selves, we are naturally drawn to the idea of continuity and flow of life. The considerations of continuity and unity of life may also play a role in shaping our attitudes to others.[310] Since there is no real need to choose between the two accounts, I leave the matter at the note that consciousness-based account of identity has good prospects of providing a comprehensive and nuanced account of anticipation.

Let us now go back to Martin's illusion problem. There is no illusion involved in the projection. The projection consists in no more than ascribing tensed properties to a subject; or in imagining the future flow of a consciousness.[311] There seems to be no room for illusion here, unless *all* thoughts about the future involve an illusion. Now, Martin does not land in this absurdity. The culprit in the "experiential projection" team is the *experience*, not projection. It is our sense of self which from the beginning involves an illusion. But if this is

[309] Cf. an interesting passage in Hazzlit: "The imagination, by which alone I can anticipate future objects, or be interested in them, must carry me out of myself into the feelings of others... by which I am thrown forward as it were into my future being and interested in it. I could not love myself, if I were not capable of loving others. Self-love, used in this sense, is in its fundamental principle the same with disinterested benevolence"; Hazzlit (1805), 3.

[310] It would be fruitful to consider in this perspective the ambiguous consolation that leaving children, works and fame provides in the face of death.

[311] Again, I do not prejudge issues in philosophy of time. Ascription of tensed properties may be taken to be semantically superficial. Still, modern detensers agree that the *meaning* of tensed statements may not be translatable into a de-tensed language, even if truth-makers of tensed statements do not include anything like real tensed properties.

the case, then it seems that Martin should be a straightforward No-Self theorist.[312] For certainly it is because we have the experience of the self that we are interested in forming and keeping the concept of person. Once again we see that a seemingly Reductionist account leads to No-Self theory.

Martin asserts that the "perceiver-self phenomenon" - our ordinary self-experience - involves an illusion without giving a good argument. Suppose that he is right in claiming that we identify a *content* of experience with the subject of experience. This, says Martin, cannot be right. But why not? There are philosophical traditions that maintain that the self knows itself *directly* and not through a mental *image* of itself.[313] It makes perfect sense to be a *naive realist* about self-perception. This means that the properties present in experience (say, "being a perceiver") are strictly identical to the properties of the object (the self). Since the substance is inseparable from the properties, this means that indeed, the self is really contained in its own experience. This may sound paradoxical, but there is no logical inconsistency here. But even if one rejected naive realism about self-perception, there is still no reason to accept Martin's claim. A metaphysically neutral and phenomenologically correct report of our self-experience would say only this: the self is *present* in experience; much like the external objects are present in self-experience. Naive realism is not a part of our phenomenology; it is a metaphysical theory.[314] It may be a wrong theory. What of Martin's claim that our self-experience *consists* in ascribing the role of the subject to a part of our experience? I simply deny its accuracy. If we want to stay on a metaphysically neutral ground, we should say that a part of our experience *represents* the subject in the same way

[312] Martin is sympathetic to Buddhism and seems to hint that this view is closer to truth than Reductionism.

[313] To mention just a few examples. Among schools of classical Indian philosophy such doctrine is held at least by samkhya-yoga and vedanta. It was also clearly formulated by the Neoplatonists. Augustine argues forcefully for this position in *De Trinitate*. At the birth time of analytical philosophy, Bradley and Russel held that the self must be directly acquainted with itself.

[314] So, for example, Augustine starts with the phenomenological claim that the mind is present to itself and goes on to argue that this could not be so unless it knew itself directly and not through an image distinct from itself; *De Trinitate*, X.3.5-10.16.

that other parts of our experience *represent* external objects. If in using “this” in pointing to the subject we would point to the experience, thus confusing the representation and the object, then exactly the same confusion should always be involved in pointing to external objects. The situation is symmetrical. Martin does not give us any reason for thinking that self-perception involves illusion while ordinary perception does not.

As far as I can see, there are two motivations for Martin’s view. First, there is his analysis of dissociative experiences which he calls “many-selves experience”. I may have the following experience: I may be acutely embarrassed and at the same time observe this emotion coolly (or even with amusement) from some higher vantage-point. In such case I experience a division between a “public-self” and a “watcher-self”. This division is explained by Martin as “partitioning of a subjective psychological space”. Thus the public-self is a *part* of the experience; and so is the watcher-self. But it is this “hidden observer” which we experience as the subject whenever the experience of the subject is present. It follows that we identify a part of our experience with the subject. And this, says Martin, is an illusion. Now, it is quite easy to see that this conclusion has been reached due to a very tendentious description of the dissociative experience. You will note that in giving *my* description of such experience I have not introduced two or more selves. To the contrary, the form of the sentence I used indicates that I take myself to be the subject both of the emotion and of the observation of it. True, there will be some feeling of distancing oneself from the emotion and objectifying it and myself-as-subject-of-this-emotion. But all this can be well expressed by saying that the self can look at the world and on itself from two (or more) perspectives at once. This, indeed, is a striking fact. But for all that, this is *one,* multi-faceted self. Now, if anybody really thinks that there are two selves in this experience, not taking it merely as a figure of speech, then either he means by “self” something else than me, or indeed he suffers from a delusion. But most of us, I believe, realize full well that in such experience one self is *represented* in experience in two different ways. Moreover, it being represented by a part of our experience does not preclude its *transcending* the experience in the sense of having many more properties than are represented in the current experience. And, indeed, the sense of "hiddenness" of the self is the evidence that we do experience the self as transcendent in this sense.[315] If so, we do not take a part of our experience to be simply identical to the self.

The second motivation for Martin's view comes from meditative "no-self experiences" of Buddhist practitioners.[316] In such experiences, there is no experience of the division between the observer and what is observed: no "partitioning of psychological space" occurs. Consequently there is no experience of the self of any sort. In this state, lucid and methodic introspection is available. Now, all that this shows is that it is possible to be clear-headed and not to have an experience of the self. But the fact that somebody does not experience her self does not show that she has no self to be experienced. We would have to have an independent reason to believe that if the self existed, it should be experienced by any clear-headed introspective person. But we do not have such reasons.[317] On the contrary, there are good reasons to believe this is not so. First, it is not clear how to understand the workings of meditative states in question. They certainly involve increased focus of attention. But then, these states may after all be similar to what Martin calls "common no-self experiences", i.e. the states when we are so engrossed in an activity or feeling that we are not self-aware. It is true that the meditation practitioner is not bound to a particular activity or feeling. Still, at any time, the state may consist in being completely focused on a particular thing or experience. And in the case of Buddhist practitioners, the beliefs they hold would predispose them *not* to achieve self-experience when meditating. This brings us to the second, more important point. We should not think naively that self-experience is any less interpretation-dependent than

[315] This is fully consistent with naive realism about self-perception. The naive realist is not obliged to take a direct experience of an object to be an *exhaustive* experience of it.

[316] For a clear and unbiased discussion of various meditative experiences see Forman (1999) and Shear (1999).

[317] Except for metaphysical arguments purporting to show that the self does not exist. But if one had such arguments, the appeal to experience would be superfluous.

experiences of the external world. Now note that an advanced Buddhist practitioner may *experience* the world as containing nothing but bundles of elements; such things as “cars” or “trees” being just conventional names without proper referents. Does this show that cars and trees do not exist? Hardly. The same goes for the self. The parallel goes unnoticed because of the popular misconception that the self is something special, altogether separate from the experience; and that there should be a special separate experience with the label “self”.[318] Well, nothing of this sort is true. What happens is that we experience our mental states *as* being states of a mental subject. Some people may experience them differently: *as* being just an ownerless bundle. If you like, this *seeing as* is an interpretation. But this says nothing about which interpretation is right and which is wrong. It is for metaphysics to decide. The fact that one can experience the world *as if* it did not contain substances, does not solve the metaphysical dispute between substantialism, bundle theory of substance and nihilism. This holds true for the special case of persons as well.

8.6. Collapse into the No Self View

Let me summarize the main argument so far. The discrepancies in Parfit’s account of Fission, as well as independent considerations, have led us to the idea that anticipation may track not identity but a continuation relation. This generates, however, a worry about the scope of anticipation. Parfitian Reductionism has to provide a plausible criterion for distinguishing appropriate and inappropriate anticipation. The Argument from Excessive Extendability shows this is impossible. Moreover, Martin’s central claim that it is appropriate to anticipate experiences of some people not identical to one is untenable, as the Argument from Experiential Projection shows. If anticipation always involves identification with the anticipatee, then either this identification is right or it is wrong. If it is right, then the

[318] I have hard time understanding why anybody would think so. Yet it seems a very popular motive among the Humeans.

anticipatee is identical to the anticipator. And if it is wrong, then anticipation involves an illusion and so is inappropriate. Therefore anticipation is appropriate only if the anticipator is identical to the anticipatee. Martin's central claim is false. There seems to be no plausible way of defending a Parfitian Reductionist account of anticipation.

The remaining task is to show that the impossibility of genuine anticipation entails that there are no persons. I propose the following principle:

(SI) My self must be something with which I *can* identify.[319]

But we cannot identify with a future self whose experiences we cannot anticipate. The possibility of identification is limited by the possibility of anticipation. And by the principle "no self without the possibility of identification", the extent of the life of our self is limited by the possibility of identification. Thus the extent of the life of a person is limited by the possibility of anticipation.[320] And conversely. I have argued that anticipation entails identity. Thus we have the following biconditional:

(SA) I will exist at *t* iff there are some possible mental states at *t* that I can anticipate having.

(SA) asserts that to exist in the future I must have potential for having mental states then and that anticipating having such mental states would be appropriate for me now.

With these principles in hand, we can turn to Parfitian Reductionism. Let us start with the micro-scale. If no anticipation is appropriate, then in particular my anticipation of

[319] By "identify" I mean here "recognize as oneself". I think (SI) belongs to theses directly bound with the concept "self"; cf. ch. 2, 21f.

[320] This claim is roughly equivalent to the conditional used by Stone: "If at t1 I recognize that the intelligent being at t2 is identical to me, then if I know he will have an experience E, then it is rational for me to anticipate experiencing E." We are now in a position to see *why* this conditional is true, and how to properly understand the predicate "rational".

what will happen *three seconds from now* is inappropriate. By itself, this is enough of a refutation. But the conclusion that there are no people follows as well. If I could not anticipate experiences occurring in three seconds from now, my life would have to finish before that time (by (SA)). I would have to be a very short-lived entity indeed. But it is impossible to identify with a strictly instantaneous entity, if only because the act of identification itself takes some time. And it is likewise impossible to identify with a too short-lived entity. For, first, the "now" of ordinary speech is always relative to a certain activity. "Now" is the time of something what I am *doing*; and this always takes some amount of time ("now" can be therefore quite extended). Conversely, to identify with something I must identify it as an object having some properties. In my own case, these will necessarily be some mental properties. But, again, mental acts take some time. Finally, anticipation seems implicitly involved in great many of our actions: walking, driving a car, and so on. For all these reasons, the entity with which I can identify must be something which lasts for some time and which can appropriately anticipate at least her close future. Such entity must be endowed with a unity different in kind from any relation which can be had to another person, and which would thereby make anticipation appropriate at least over a short range of time. Actually, the Parfitian Reductionist may have a way to address this issue. She may subscribe to Michael Tye's view according to which the unity of consciousness entails that there is numerically one (complex) experience.[321] And she should reject Galen Strawson's picture of intermittently "restarting" consciousness. Then, as long as unity of consciousness lasts, the person would not, strictly speaking, be a bundle of experiences. There would be only one experience. The person would consist of this one experience. The oneness of experience would make anticipation over short ranges appropriate. Unfortunately, there seem to be breaks in continuity of consciousness. Our life does not contain just one long

[321] Tye (2003), 25-35.

experience, but many successive experiences. And it would be just inappropriate for me to anticipate an experience I will have tomorrow, as it is inappropriate for me to anticipate *your* experience tomorrow. Thus, at most, I could be something like a one-day self.

It is not quite easy to show that one-day selves would not be persons. Of course, we could say that Parfitian Reductionism collapses into something akin to Strawson's Transience View. Yet I want to advance a more ambitious claim: that it collapses completely into the No Self Theory. First, we may observe that denying that there is a relevant difference between my anticipating having what we ordinarily call "my experiences tomorrow" and my anticipating having *your* experiences tomorrow is still absurd. We deeply believe that our lives are separate and go on for a long time. Yet denying this is not as absurd as denying appropriateness of anticipation over a range of seconds. For when unity of consciousness gives way, it is harder to give good reasons for thinking that there is some principle which unifies the experiences of today and of tomorrow. This is what we deeply believe, but we have no watertight reasons to refute the sceptic. The reason not to call one-day selves persons will have to do more with consistency of Parfitian Reductionism than with metaphysics. On Parfitian Reductionism, it is enough that a bundle of mental states *could* be regarded as a relevant unity for a person to exist. In this sense, the existence of persons is "cheap"; and for this reason hard to disprove. To achieve it, we need to inquire into what is involved in regarding a mental bundle as a unity. The Parfitian Reductionist cannot explain this act in terms of any robust concept of the mental subject. For then either her own account would invoke entities the existence of which she denies; or at least our acts of "bundling" would involve the use empty concepts and so would involve an illusion. This would be a perfect reason to say that persons do not exist. Nor can the Parfitian Reductionist resort to co-consciousness relation in explaining bundling, since that would re-introduce the abominable principle "one person - one consciousness". So, it seems, our acts of bundling can have nothing to do with *metaphysical* reasons and concepts.[322] So what other reasons there are? There are two classes of reasons left. The first class has to do with *explanatory strategies*. By treating a certain bundle of mental states as a system, we avail ourselves of explanations of human behaviour in terms of beliefs, desires, reasons, narratives and so on. The self is thus an *explanatory unit*. The *explananda* are human actions conceived, precisely, *as* actions. What Dennett calls "intentional stance" is not just an effective method of making predictions. When applied to humans, this stance is necessarily bound with making the ethical distinction between actions and mere happenings. Thus we come to the second class of reasons for regarding a bundle of mental states as a unit. These are the *ethical* reasons. The person is for us a *moral unit*. Now, our explanatory and ethical interests are *reasons* to coin and use the concept of person, regarding mental bundles as making up relevant units (i.e. persons). But they are more than just a reason: the act of regarding a mental bundle as a unity *is* the act of regarding this bundle as an explanatory and moral unit. This is the relevant sense of unity. Thus, *for the Parfitian Reductionist*, persons are indeed *essentially* explanatorily and morally interesting. But, obviously, we do not regard one-day bundles as relevant explanatory and moral units. And larger bundles cannot be regarded as such units because of the inappropriateness of anticipation. Therefore, on Parfitian Reductionist principles, there are no persons.

8.7. Anticipation and Reductionism

In the foregoing discussion I have focused on Parfitian Reductionism; but, except for the previous section, nothing really hinged on the characteristic features of that view. In fact, I appealed to premisses which seem common to all Reductionists:

[322] In keeping with Parfit's insistence on the idea that the separateness of persons is non-absolute and metaphysically shallow.

(i) there are possible cases where personal identity is *indeterminate*

(ii) there is only a difference of *degree* between our relation to our future selves and to other future selves

(iii) there is no metaphysical "deep further fact" underlying distinctness of persons.

I have also argued that:

(iv) anticipation is a *reason-responsive* concept

I claim that, taken together, these propositions lead to the conclusion that, first, anticipation is excessively extendable and, second, that no anticipation is genuine. If we add two rather plausible claims:

(v) for persons to be morally interesting entities they must be able to appropriately anticipate their future

(vi) persons are essentially morally interesting entities

we get the conclusion that *there are no persons if the Reductionist premisses (i)-(iii) are true*. Thus my argument can be extended to all forms of Reductionism. But, unlike Parfit, non-Parfitian Reductionists may wish to reject one of the premisses (iv)-(vi). Unsurprisingly, the effort of Reductionist philosophers focused on rejecting premiss (iv) (or its equivalents for other morally relevant attitudes). There are various projects of showing that our morally relevant attitudes are insulated from judgments about the realities underlying the facts of personal identity. I have indicated some reasons why I think this is a wrong idea. But doing full justice to such projects would require systematic investigation of the *meta*-ethical aspects of the great debate on “what matters in survival”. This, I am afraid, is beyond the scope of the present work. Thus, let me state the conclusions of this chapter in a conditional form: *if* Parfit is right in claiming that our central ethical concepts are reason-responsive, *then* Reductionism entails that no anticipation is genuine and therefore is radically implausible. And *if* premisses (v) and (vi) are true, then Reductionism collapses into No Self theory. Even if premisses (v) and (vi) are wrong, then Reductionism enjoys no theoretical advantages over Transience and No Self views. Just like these views, Reductionism undermines our main interests in operating with the concept “person”. Given their simplicity and elegance, Transience and No Self views should commend themselves to a Naturalist as the more plausible option.

Chapter 9. Transience and No Self Views

9.1. Galen Strawson's Transience View

9.1.1. Anticipation

The Transience View advocated by Galen Strawson denies that, at least in the human case, there are *persisting* selves. There are only very short-lived, momentary, selves.

While in respect of the ethical thrust of the theory, Strawson's view may be kindred to Parfit's, in other respects Transience View has a very different flavour from Parfitian Reductionism. Strawson's approach is from the start robustly metaphysical. Moreover, his momentary selves are genuine metaphysical subjects of experience.[323] There is no reason for such theory to be committed to the view that selves are essentially morally interesting. Even though the arguments concerning anticipation which I employed in the previous chapter can be applied to the Transience View, this theory need not end up denying the existence of selves.

It is easy to see that the admission of momentary selves changes nothing as far as providing a criterion for appropriateness of long-term anticipation is concerned.[324] There is no long-term persisting self (in human case); so identity of the self cannot restrict anticipation. As for psychological continuity, the idea of Cosmic Consciousness cannot be exploited. Such object would clearly lack the unity of consciousness which is characteristic of and essential for the self as conceived by Strawson.[325] But considerations of arbitrariness and conventionality of the definition of psychological continuity remain in place. Moreover, it

[323] Strawson (1997), 338f, 344f ; Strawson (1999), 106f, 112-114. This is true despite Strawson's regarding the distinction between substances and processes and properties as metaphysically superficial, see Strawson (1999), 124-129 and Strawson (2008*b*), 273ff, 279ff.

[324] By "long-term" I mean "longer than the usual span of uninterrupted consciousness in humans". As in the preceding section, we can think that long term is anything longer than one day.

[325] Instead of saying 'the self as conceived by Strawson' or 'momentary self' I will use Strawson's term *SESMET* - Subject of Experience that is a Single MEntal Thing - with the tacit understanding that SESMETS are transient.

does not seem possible for Strawson to go for psychological continuity as the criterion of appropriateness. For Strawson observes that the *self* may be *experienced* as something persisting for a long time.[326] And what else can this experience be other than taking some past and future imagined situations to happen to this present self? Since Strawson admits that persistence of the self is something on the level of experience (even if this experience may be illusory), he must implicitly subscribe to something like Martin's "experiential projection" theory. This is a quite plausible theory. So it seems that what Strawson should say is this. Anticipation indeed presupposes identity of the self. In the end, it is identity which distinguishes appropriate from inappropriate anticipation. But, in fact, there are no long-term persisting human selves. Therefore no long-term anticipation is in fact appropriate.[327]

This conclusion is very hard to believe. It could be thought that an appeal to the idea of the double reference of "I" could blunt this conclusion. According to Strawson, "I" refers *both* to the (momentary) self and to the human being. Like "here", "I"can refer to more or less. If so, then perhaps one could say that it is the identity of the human being which is tracked by the acts of genuine long-term anticipation. Now, I think there is a lot to be said against the idea of double reference of "I".[328] But for our purposes it suffices to say that it is irrelevant to anticipation. It is the very mental self, as Strawson admits, which is experienced as persisting. It is the self which enters anticipation, not the human being. This is easy to see in the case of people anticipating having post-mortem experiences as ghosts or disembodied souls. Their anticipation is as good as any, while it does not presuppose the identity of a

[326] Though, according to Strawson, the self *need* not be always so experienced. Strawson (1997), 347, 353;, Strawson (1999), 106.

[327] I take Strawson to be saying as much, when he asserts that when he identifies with his present self, he* has no sense that events further in the past or in the future involve him* in an way (personal pronouns with asterisk refer to "that which is now experienced as the self"), Strawson (1997), 354f; Strawson (1999), 109; Strawson (2005), 68f. So, I think, Strawson does not really anticipate having these experiences (although he imagines them and reacts in ways which, as he says, are biologically hard-wired). So on his own terms, Strawson suffers no illusion.

[328] At least if it is taken seriously. Many philosophers make observations on the flexibility of "I"; for example, that a truck driver can truthfully say "I weigh two tons"; Glover (1988), 64-68; Unger (2000), 286ff. If the extension of "I" taking the human being in its scope is of this sort, then it is quite harmless, but it cannot do any heavy work.

human being. It is just the identity of the self which is presupposed. We can also note that when Strawson himself gets serious - when he addresses the question why he feels he* is involved in the far-away death of Strawson the human being - he does not appeal to the idea of identification with the human being. He admits that he really anticipates death and that this act concerns his very self.[329]

So, to repeat, Transience View entails a claim which is extremely hard to believe: that no long-term anticipation is appropriate. Still, it does not follow that there are no selves. SESMETS exist at least long enough to be capable of self-consciousness. Their existence and unity has nothing to do with ethical concerns. So, on a morally neutral definition of selves, selves do exist.

9.1.2. Associated doctrines

In keeping with my general method, I will now address ideas which lend intuitive support to the Transience View. In Strawson's philosophy there are two doctrines which, though not essential to his metaphysical view of the self, are closely associated with it.

The first of these doctrines concerns continuity of consciousness. The phenomenological description which Strawson gives of consciousness is markedly different from those which emphasize its seamless continuity. In describing his own consciousness, Strawson asserts that it is intermittently "restarting" from moments of unconsciousness.[330] Now, I have no reason to deny the accuracy of Strawson's description of *his* consciousness. I also think I know what he means by "restarting". Sometimes, on the border of sleep and wakefulness, or in states of intoxication, my consciousness seems (in retrospect) to be dis-continuous in some such way. But this is not the case with me during my normal waking hours. And, I suspect, many people are like me and unlike Strawson in this respect. For all I

[329] Strawson (1997), 355; Strawson (2005), 65.
[330] Strawson (1997), 356ff.

know, this may well be due to some differences in brain-activity. Thus the difference in phenomenology should not appear to us as necessarily mysterious. In any case, the main point I want to make is that Strawson's view of consciousness, based as it is on his experiences of his own consciousness, has no claim to being universally valid. Some people experience unbroken flow of consciousness throughout a day at least. Now, Strawson would agree with the principle:

(CS) If a self's consciousness flows on continuously, the self persists.

This, coupled with the observation that some people enjoy unbroken consciousness throughout a day, entails that at least *some* selves are not momentary. There are some selves of at least one-day duration. By itself, this is not a problem for Strawson's view.[331] Nevertheless, it is the "intermittent restarting" view which gives a measure of intuitive support to Transience view. For if consciousness comes in very short, strongly unified episodes, it is easy to think of the momentary subject of a given episode as not being usefully distinguishable from the episode. However, when the subject persists through a long chain of changes in consciousness, it seems natural to ask: what then is this subject? Its being no longer seems to be wholly exhausted by its being something unified with experience or constituted thereof; it becomes natural to think of it as something somehow *underlying* the stream of experiences. And, of course, it is much easier to think of even a one-day self as of a free and responsible *agent*, while it seems virtually impossible in the case of a momentary self.[332] This may lead us to question whether the self, which is something more than just consciousness, could not, after all, survive a break in consciousness. This question will be considered in the next section.

[331] See Strawson (1999), 130.
[332] Strawson actually rejects ideas of free will and responsibility on other grounds.

For now, let us look at the second doctrine associated with the Transience view. Strawson distinguishes between Diachronic and Episodic ways of experiencing the self. While people with strong Diachronic tendencies have a “strong sense that *the I that is a mental presence now* was there in the past and will be there in the future”, the Episodics lack this feeling.[333] More precisely, Episodics lack the sense that their self is involved in events occurring in a *further* past and future. What counts as “further” is not universally fixed. Strawson rightly observes that this depends both on the temperament of the person and on what she thinks about.[334] But in almost all cases the extent of the life of the “present” self as experienced by an Episodic will significantly exceed the average time of real existence of a human self (some three seconds according to Strawson). Perhaps people engaged in a meditation practice could cut down the experienced life-time of their self to a moment. But in common cases, the lifespan of the “present” self will be considered to be rather something like half an hour, as Strawson suggests in a footnote.[335] This shows, however, that even people with strong Episodic attitudes commonly presuppose that selves are objects persisting for a much longer time than Strawson would have us believe. This means that, whether we are Diachronics or Episodics, most of us deeply believe that we are persisting and *non*-momentary selves. Strawson is not one to be moved by such conclusion. Yet on my methodology this conclusion counts for something.

9.1.3. Unity of the self

The crux of Strawson’s theory lies in aligning the unity of the self with the unity of consciousness.[336] As long as there is unity of consciousness, there is one self. With this part of Strawson’s theory I am in full agreement. But Strawson adds the second part: when there

[333] Strawson (1999), 109; cf. Strawson (1997) 353f. Diachronicity and Episodicity are discussed at length in connection with Narrativity in Strawson (2005).

[334] Strawson (1999), 111.

[335] Strawson (1999), 111.

[336] Strawson (1997), 345f, 360; Strawson (1999), 119f, 129f, 132.

is a break in the unity of consciousness, there is necessarily a break in the unity of the self. One self goes out of existence and a new one steps in. The selves do not outlive the episodes of unified consciousness of which they are subjects. Has Strawson given us a good reason to believe this claims? Has he given a good argument against Persistence View and for Transience View? It is possible to extract two arguments from remarks made by Strawson.

The first argument relies on the refusal to take the distinction between substances and properties or processes seriously. The argument may run as follows:

(1) There are separate episodes of unified consciousness.
(2) There are separate unified conscious processes.
(3) There is no metaphysically significant difference between processes and substances.
(4) Therefore there are separate unified conscious substances.

For now, let us put aside doubts about the soundness of premises. Even if the truth of the premises were granted, two questions would remain. The first is this. Granting that there are momentary selves, which are in some sense separate, why could they not constitute a bigger, persistent self? In explaining why he thinks that there could not be a persistent human self, Strawson writes:

> "Many [...] insist that this inner subject is or can be something that has long-term diachronic continuity. On my view, though, this amounts to claiming that a many-membered set or series of SESMETS in a certain relation can be a single subject of experience. But a many-membered set of SESMETS in a certain relation is simply not the kind of thing that can itself be a subject of experience".[337]

I think I agree with Strawson on the last point, though it is hard to say what are Strawson's reasons for this claim. Strawson himself only says that it is *not* based on "philosophico-grammatical point about the word 'set'".[338] So what other reasons there are? I think that

[337] Strawson (1999), 132.
[338] *Loc. cit.*

considerations of *unity* are crucial at this point. According to Strawson, the unity of momentary selves is the strongest form of unity known to us.[339] The unity of the "set" constituted by separate selves is weaker and of a different sort than that of the particular selves. This thought can give rise to three lines of reasoning. First, it could be said that the unity between selves is too weak to generate a genuine substance; and hence a subject. Only the paradigm instances of unity should be seen as generating substances. Secondly, the unity of the subject of consciousness should not be something unrelated to the unity of consciousness. The unity of the subject should underlie and partly explain the unity of consciousness. This demand seems to rule out all sorts of unity arising from a "bottom-up" construction. Thirdly, augmenting on the second line, the unity of the subject of consciousness should be *conceptually* bound up with the unity of consciousness (if only because no deep distinction can be made between the two). The unity of momentary selves conforms to this demand; the unity of the set - where separate momentary subjects are externally and extra-consciously linked - does not. My guess is that the third line is closest to Strawson's thoughts on the matter. But now we need to ask: why the *set* (no matter how non-technically is the word used) is the *only* alternative to momentary selves? Why there cannot be a persisting thing which is *more* unified than a set, in such a way that its unity conceptually involves some measure of unity of consciousness, but which is nonetheless not totally bound with this unity, so that it can survive the breaks of consciousness? Or, to put it otherwise: is there no way that momentary selves could "coalesce", rather than make up a "set"? And this question leads to the second doubt about the first argument: in what sense are the selves "separate"? Is the move from "separate episodes of consciousness" (separateness equals lack of the unity of consciousness) to "separate substances" valid? Now, in premisses (1) and (2), separateness equals lack of the unity of consciousness. But in conclusion (4),

[339] Strawson (1999), 124.

separateness means that for two separate episodes of consciousness there are two distinct substances and no unified substance underlying both episodes. So, on the face of it, the argument is invalid. Even if substances were not distinct from processes, they could be dis-unified in the sense of lack of unity of consciousness (*qua* processes, so to say), but be unified metaphysically (*qua* substances). However, the argument is enthymematic, rather than invalid. As we should expect by now, there is a further premiss concerning the *unity* of the substance/subject involved. The crucial premiss in Strawson's argument is the following:

(A) The principle of unity of the substance/subject is necessarily the same as the principle of unity of the process by which this substance is "constituted".[340]

This claim is a consequence of Strawson's view of substance-process distinction, shortly stated in premiss (3). In the case of the self, it could also be supported by additional considerations, of the sort we discussed above. This premiss renders the argument valid. It also provides material for the second argument for Transience view:

(1) An object persisting over breaks in consciousness would have a different sort of unity than the unity of consciousness.
(2) Therefore it would not be "constituted" by consciousness.
(3) Therefore it would be a different sort of thing from the experienced (momentary) subject.
(4) Therefore it would not be a subject of experience.

Having satisfactorily stated the arguments, we can now take up doubts concerning their soundness. The most pressing question is: why is the subject so closely aligned with consciousness? Why is it "constituted" by consciousness? An answer emerging from Strawson's remarks is disappointing: the subject is *defined* in this way.

[340] See Strawson (1999), 128 for this usage of "constituted".

> Most philosophers use the term ‘subject of experience’, which forms the part of the term ‘SESMET’, in such a way that the subject of experience can be said to exist in the absence of any experience, and many have grown so accustomed to this use [...] that they no longer hear the extreme naturalness of the other use, according to which there is no *subject of experience* if there is no *experience*.[341]

This answer is disappointing, because it leaves open the question whether “subject of experience” is a substance or a phase sortal. Someone could accept Strawson’s use, but hold that in this case “subject” is a phase sortal and that the same *thing* is first a subject of an episode of consciousness, then ceases to be a subject and then is a subject again. If that is to be ruled out, there must be a *metaphysical* link between the thing which is a subject and consciousness. The question: “why the two are so closely aligned?” recurs.

The idea of definitional link is not, however, wholly useless. Let us try to picture a Strawsonian world. In this world, there is no useful distinction between substances and processes - at least, I take it, when they are correctly specified. A relevantly unified process will be a substance. An episode of consciousness is relevantly unified. So it “constitutes” a substance - a SESMET. Now note that nothing in Strawson’s account commits us to *simplicity* of substances/processes. They may be quite complex; as any episode of consciousness is, despite being unified. This thought opens up the following possibility. The episodes of consciousness can be *parts* of a larger process. Say, a life of a human being. This larger process will “constitute” a substance - a human being (and SESMETS will be its parts). One could claim that this is a thing which is more unified than a set; perhaps in a way that conceptually involves some measure of unity of consciousness, but which nonetheless can survive breaks in consciousness. And, since consciousness is a sub-process of the process which “constitutes” the human being, it is certainly right to say that the human being is conscious. So, it seems, it *is* a subject of consciousness. Does this pose a problem for Strawson? Not in the least. For Strawson is anyway ready to say that “I” is ambiguous between the SESMET and the human being. So, of course, *in one sense,* a human being is a

[341] Strawson (1999), 130.

subject. But in *another* sense, it is not. In the sense of “experienced pure mental presence” and “the thing ‘constituted’ by consciousness”, the SESMET is a subject. So, it seems, the problem in the end *is* terminological. There are SESMETS and human beings. There are subjects in sense one, and subjects in sense two. There is no conflict.

This solution all of a sudden seems to make the matters trivial. But it would be wrong to think so. First, Strawson’s metaphysical position is anything but trivial. For Strawson’s metaphysics could allow for elimination of human beings (which are less strongly unified than SESMETS) but not SESMETS. This, certainly, makes a big difference. Secondly, the solution based on distinguishing various senses of “subject” glosses over two important problems. The first is this:

(Q1) Is the self-as-experienced really experienced as “constituted” by consciousness and having the same principle of unity as consciousness?

Recalling our discussion in the previous section, we can reply with a resounding “No”. The self is - usually - experienced as something transcending a given episode of consciousness. What follows from this answer? We have sketched Strawson’s view according to which a human being could be called a “subject”; and the SESMET could be so-called in another sense. But now we should ask *what* is this second sense and where does it come from? What makes a difference between the two senses? It could not be “being pure mental presence”. For, evidently, one can experience oneself as such and think that one transcends a given episode of consciousness. So “being pure mental presence” is evidently applicable to things which are experienced and conceived as transcending consciousness and therefore as different from true SESMETS. Secondly, we saw that the point Strawson makes about the use of the phrase ‘subject of experience’ is irrelevant. A human being could be a subject of experience which does not outlive the episode of consciousness, because ‘subject of consciousness’ in such use could be taken - quite naturally - as phase sortal. As I see it, the only thing which makes a difference is this: that the SESMET is as closely unified with consciousness as possible; that it is a substance “constituted” by it. *This* is the special content differentiating the sense of ‘subject’ as applied to SESMETS from other senses. But now, where does this content come from? The answer is that it comes from Strawson’s arbitrary definition. For experience of the self does *not* say that the experienced self is something of this sort; it usually says something contrary.[342] Having grasped that, we can ask the crucial question which was obscured by Strawson’s treatment of the matter:

(Q2) Does experience provide us with some principles of *subjectual* unity which are different from the principle of unity of consciousness?

It is worth noting that Strawson in a way poses a false dilemma: either we go for SESMETS and unity of consciousness *or* we go for human beings and for links between subjects of individual episodes which are external and unrelated to the structure of consciousness. Both SESMETS and human beings can be called “subjects”, but there is a world of difference between their unity, the sort of thing they are, and consequently between the sense in which they are subjects. But I will argue that there is a viable *tertium quid*. In my argumentation, I will be as accommodating to the spirit of Strawson’s metaphysics as possible. In particular, I will not dispute the validity of Strawson’s view of the distinction between substances and processes and properties.[343] Even on such assumptions the Persistence View will emerge as

[342] This does not mean that SESMETS are subjects only in a deviant sense, and are not real subjects. That would be too quick. SESMETS could meet ordinary criteria of subjecthood. All that follows is that persistent things - like human beings, if they exist - would not be any *less* subject or selves than SESMETS. For the additional condition that SESMETS meet turns out to have no basis in our experience of the self.

[343] One reason is that I look favourably on the Cartesian understanding of substance as constituted by its principal attribute. Strawson’s view, as far as I understand it, is not very different from this Cartesian view.

preferable to the Transience View.

To better grasp the problematic of subjectual unity, let us start with a view more radical than Strawson's. Consider a *point* in time such that an experience occurs at this point. Corresponding to the points in time we could define *strictly momentary* selves. Such selves would be durationless. Time-wise they would be altogether *simple*. Therefore, they would be more strongly unified than ordinary momentary selves, which are not simple. The unity of strictly momentary selves is the paramount unity in nature. Someone could claim that only strictly momentary selves exist. Call this Radical Transience View.[344] The proponent of this view could employ an argument against Strawson which would parallel Strawson's arguments against Persistence View:

Argument from Paradigmatic Unity

(1) Only the strictly present subject is ever experienced (since the past and the future are not experienced).
(2) An object persisting over any duration of time could only be a "set" made up of strictly momentary subjects.
(3) Such object would have a very different and weaker sort of unity than strictly momentary subjects.
(4) Therefore it would be a different sort of thing from the experienced (strictly momentary) subject.
(5) Therefore it would not be a subject of experience.

For a moment, let us put aside the plausibility, and focus just on the idea of the argument. The idea is this. There are paradigmatically unified subjects. Anything less unified is just a "set" of these low-level subjects. Therefore anything less unified either is not a subject, or is a subject in some weaker sense. There are two lines of response to this argument. I will call the first the Proliferation strategy and the second the Coalescence strategy. We shall see that each of these strategies leads to abandonment of Transience View in its original form.

[344] Theodore Sider's four-dimensionalist Stage View is an actual example of the Radical Transience View; see Sider (2001), 188-208. Sider's reasons for embracing the view have nothing to do with the argument I present in this section.

The Proliferation strategy admits that there are low-level unified substances; and that higher-level substances are less unified. Still, these higher-level objects are genuine substances. Moreover, and this is the gist of the strategy, these higher-level substances are subjects in as strong a sense as low-level substances. This strategy entails multiplication of mental substances; and, on some variants, multiplication of subjects. Hence the name "Proliferation strategy".

This strategy comes in at least three variants. The first variant is to say that as there are many different substances on various levels, so there are *as many genuine subjects.* This variant seems to lead to an *excessive* proliferation of subjects (which would all *share* experiences). For any degree or form of unity we could define a corresponding substance and consequently a corresponding sort of subject/self. There will be strictly momentary selves, selves existing for one specious present (momentary selves), selves existing as long as a continuous episode of consciousness lasts (call them "continuity-related selves"), Episodic selves presumably characterized by some strong psychological continuity, and so on, down to human beings. But then, the talk of "subjects of experience" just cannot be taken seriously anymore. I can understand to some extent Strawson's original position: that one can identify both with an Episodic self and with a human being.[345] The difference between the two is clear and we can understand the reasons and mechanisms of such double identification. But what shall we do when we are confronted with a spectrum of subjects each of which can be experienced as a self? The experience - even the experience of the "pure mental presence" - does not indicate which of these candidate-selves is experienced (especially when we consider the upper part of the spectrum consisting of various short-lived selves). All these selves would have a good claim to being subjects. Can we identify with *all* of them? I find it impossible. I lose a grasp on the idea of identification. What does this act involve in the end

[345] I mention the Episodic self rather than the momentary self because, as we said, even among Episodics the identification with a momentary self is uncommon.

and *who* identifies with whom? And if we cannot clearly say which of the selves do we identify with does it not mean that we do not identify with *any* of them? If so, then these would-be selves are not selves at all. For these reasons, this variant of Proliferation strategy seems to me unattractive.

According to the second variant, there are many different substances on various levels, but only *one* higher-level substance should be *counted* as subject. The reason to make this move is that the low-level substances, despite being in themselves quite fit to be subjects, are *unified* as *parts* of a bigger subject. And when there is a bigger subject encompassing smaller would-be subjects, it is inappropriate to count the parts as subjects. This is a move familiar from debates about body-oriented theories. It is possible to deny that my head is the subject of my thoughts, just because it is a part of the organism, even though if the head existed by itself (and was artificially supported), it would count as a subject. I will not comment on the plausibility of this move. For present purposes it should be noted that in this variant the bigger subject is *not* just a "set" whose parts are externally linked. The bigger subject is a genuine subject. It has to be unified by a genuine *subjectual* unity (recall our guiding question (Q2)). But then, if this move would work for moving from Radical Transience View to Transience View, it would also work for the move from Transience View to Persistence View. All that would be necessary for this move to work would be to find a positive answer to (Q2); i.e. to find a principle of subjectual unity different from unity of consciousness.

The third variant of Proliferation strategy has it that there are many different substances on various levels, but there is *only one true subject*. The low-level substances such as strictly momentary "selves", momentary "selves", and continuity-related "selves" - all substances defined by some form of unity of *consciousness* - are not to be counted as subjects at all. Rather, these substances can be thought of as something like complex *sense data*. The subject would stand in a particular relation to them - it would be, let's say, aware *of* them. The nature and unity of the subject would not be explained in terms of consciousness only. The proponent of this variant would therefore have to come up with an answer to our question (Q2). Since on this variant too the unity of the subject is not tied to the unity of consciousness, this variant leads to abandonment of Strawson's Transience View.

Consider now the second strategy: Coalescence. It consists in rejecting the move from "there is a paradigmatic form of unity" to "there are paradigmatically unified separate substances". Why would there not be separate substances with the strongest degree of unity? Because they would always "coalesce" into a substance characterized by a weaker sort of unity, but sufficiently strong to produce a truly unified substance. The given paradigm unity would have to be regarded as unity not of substance but of something *in* the substance. This could be called an abstract aspect of the substance or a virtual part thereof. The reasons to regard the initial candidate for substancehood as just a virtual part of a bigger complex substance would be that this candidate is unified with other similar candidates in a way which is not admissible for substances. The mutual dependence of the virtual parts is too strong; they could not be regarded as complete independent substances.

This abstract reasoning will become more tangible when we consider the actual soundness of the argument for Radical Transience View. The question which first comes to mind is: what is actually meant by "present" when one talks about the present self? If it is a point in objective time, then one can argue that no experience can occur within a point. Moreover, points can be taken to be mere mathematical abstractions and not real units. On the other hand, if one takes the *phenomenal* present - i.e. the specious present - then this is something unified as strongly as possible. Even though specious present is not simple, it has a true unity. This unity is sufficiently strong to generate a real substance: a momentary self. The strictly momentary selves have to be regarded as "coalescing" into a momentary self, and as abstractions thereof.

But the same move can be repeated with regard to momentary selves. For *continuity* of consciousness transcends specious presents. It links experiences which are not *directly* continuous, as experiences within a single specious present are. It is a different form of unity than the unity of a single specious present. And yet, continuity of consciousness entails the unity of the subject of consciousness. Strawson's remarks suggest that he would agree. But then, why is it legitimate to rely on the unity of continuity transcending specious presents and not on the stronger unity of a single specious present? For Strawson, this problem actually does not arise. For his vision of "restarting consciousness" seems to imply that individual episodes of unified (continuous) consciousness are coextensive with individual specious presents and are marked off on both sides by breaks in continuity of consciousness. In effect, it is not clear what Strawson means by "unity" of experience; and it is not clear whether, in speaking of possible selves whose consciousness would not be gappy, he imagines their consciousness to be just continuous, or as being one very extended specious present.[346] Be it as it may, the "restarting consciousness" account is not universally valid. So it won't solve the problem of choosing between the unity of specious present and continuity. Even if Strawson insisted that the self is not to be distinguished from consciousness, the problem would remain. For consciousness itself is characterized by two different sorts of unity. Which of them generates the mental substance? It is, I contended, continuity. But this means that momentary selves (selves coextensive with a specious present) have to be regarded as "coalescing" into a continuity-related self.

The Coalescence strategy is sound. It turns out that the following situation is metaphysically possible. There is a paradigm form of unity (unity of specious present). However, there are no substances which are characterized by this unity. That is because these

[346] Strawson (1999), 130.

would-be substances "coalesce" into a larger substance characterized by a *different*, and in a sense *weaker* form of unity. Once the soundness of this general strategy is admitted, there is no reason why it could not be applied in turn to continuity-related selves to obtain a larger self not so closely related to continuity. This does not mean, however, that any substance-generating unity would do. Recall the question from which we started the present reasoning:

(Q2) Does experience provide us with some principles of *subjectual* unity which are different from the principle of unity of consciousness?

The account of the subject and of the subjectual unity has to meet several demands. First of all, it should be grounded in experience of the self. Secondly, the account of the unity of the subject has to explain why the persisting substance is a truly unified substance, rather than just a "set". In other words, we need to see why momentary selves - the would-be substances "constituted" by consciousness - have to be regarded as virtual parts of the bigger substance. For this order to be met, consciousness itself - which can be regarded as a process or a property "constituting" a would-be substance - has to be seen as something which is not conceivable except as a sub-part of a more comprehensive process. Or, in Cartesian terms: consciousness will have to be understood as a *mode* of a more comprehensive attribute which would "constitute" a proper substance.[347]

This is a tall order. But I think it can be met. I will first present my own proposal.

[347] One of the aims of considering Radical Transience View is to show that Strawson's point about there being no difference between substance and property is no answer to the question: "why SESMETS should be conceived as substances rather than properties or processes or the like of an enduring substance?". Even if substance has to be, in the end, understood as "constituted" by *a* property or process, we should make a Cartesian distinction between *principal attributes* (which "constitute" substances) and *modes*. In other words, not every specifiable property should be regarded as "constituting" a separate substance. It would be crazy to think otherwise. Radical Transience View is an example of what happens when the attribute-mode distinction is not used. It amply shows, I think, that the distinction has to be made. I also suppose that Strawson would not oppose it. But then, Strawson gives no specific reason why consciousness should be regarded as a substance-constituting principal attribute, rather than a mode. In other words, he offers no reason why SESMETS are substances rather than abstractions or phases or modes of a persisting substance.

Then, to be on the safe side, I will sketch some fall-back positions. My proposal is Dualist and is, in a sense, trivial. The essence of the substance which is a subject is just this: to be a subject of centered and potentially continuous consciousness. The unity of such subject cannot be specified otherwise as, well, the unity of such subject. This, after all, is what Simple View is all about. The unity of the self cannot be reduced to any other form of unity or continuity. There are, however, two nagging questions. Do we really have a grasp on *what* this subject is? And how exactly are we to conceive the relation of this subject to consciousness?

First, let me repeat that the subject is experienced. It is one thing to be aware of one's consciousness. It is a different thing to experience *oneself* (people practicing Buddhists meditation presumably have the former experience but not the latter). Moreover, since the self is experienced as persisting, it is thereby experienced as something more than just consciousness. It is experienced as the subject underlying consciousness. But what after all do we grasp in experiencing the self? What is the content of this experience? As I said in the previous chapter, experiencing oneself is more a matter of experiencing one's consciousness as being structured in a particular way - of seeing it as a state of a substance - then of there being a particular new content *in* experience.[348] It is not like a new sense-datum-like thing pops out and waves a flag "subject". Such appearance would be useless.[349] Still, this does not mean that the experience of the self is wholly contentless. It is just that we should not expect there to be some sort of easily graspable quality such as qualities ascribed to material things. The self is first of all experienced as mental *presence*.[350] Secondly, the self is experienced as

[348] See remarks to the similar effect in Strawson (1999), 112-115.

[349] If such sense-datum were taken to be the subject, then regarding such self-consciousness as an illusion would seem sensible.

[350] Note Strawson's definition of 'sense of the self': "the sense that people have of themselves as being, specifically, a mental presence; a mental someone; a single mental thing that is a conscious subject of experience" in Strawson (1997), 338.

active: experiencing, attending, willing etc.[351] That is enough to give content to the notion of the self.

When it comes to the relation of the subject to consciousness, there are several options. First, the subject can be conceived as a *substratum* or a "field" which can be filled with actual contents. The actual consciousness would be the state of this field being filled. Secondly, one can take Unger's line and ascribe *powers* to the self. The self will have a power to be conscious in a unified way. Consciousness will be a state of activation of this power. Now, powers, according to Unger, are categorical properties.[352] So, in principle, there is no problem with a power being a principal attribute "constituting" the self. The third line is quite similar. The self can be conceived as a pure mental *activity*. Consciousness is a particular state of this activity. To understand the situation of the self during a period of unconsciousness we can use an analogy. Imagine staying in a completely dark room with your eyes open. You look, but you do not see. Although you are active (you look), the external conditions make it impossible for you to attain the completion of this activity (you do not see). As seeing is to looking, so consciousness is to mental activity of the self (or, more precisely, to a part of mental activity of the self, which also involves attending, willing etc.).

Finally let us consider the relation of unity of consciousness (in its different forms) to the unity of the subject. The subject is conceived as the unifying principle, as the centre and the point of origin of consciousness when consciousness is present. As I argued, the unity of the subject is a necessary condition of continuity of consciousness. And it is a

[351] Activity as an ontological category does not entail anything about moral agency or control over one's life and so on.

[352] Unger (2005), 232-239. Strawson argues for an even stronger position: that all categorical properties are *identical* to power properties, Strawson (2008*b*). Thus Strawson would have no reason to complain about the characterization of the self in terms of powers or activity; and indeed he notes that Berkeley's and Fichte's characterizations of the self are valid characterizations of a substance, Strawson (2008*a*), 546.

necessary condition of co-consciousness of experiences within a given specious present.[353] This connection between unity of the subject and unity of consciousness is best expressed by saying that, in reality, the relations unifying consciousness are 3-place relations: "Experience A and experience B are had co-consciously (or in dynamic succession, or continuously) by subject X".[354] If so, then any 2-place relation between experiences, as unity of consciousness is normally conceived, is an abstraction from a more complex reality.

I have argued that consciousness and unity of consciousness *can* be conceived in a way which makes clear why would-be momentary substances should be regarded as merely virtual parts of a persisting subject. The question now is whether we *have* to conceive consciousness in some such way. I suggest that the answer is "Yes" *if* we want to bring in subjects of consciousness at all. To experience the subject, as we said, is to experience one's consciousness (or, more broadly, one's mental life) as being structured in a certain way; namely, as being unified by a centre of mental activity. The account I presented remains faithful to this experience. On the other hand, if the subject were wholly "constituted" by consciousness and had no specific properties of its own , then one could doubt whether the idea of the subject does any real job at all. In fact, the difference between Transience View and a No-Self View which would recognize the facts about unity of consciousness seems merely verbal.[355] Of course, one could also state a Transience View about the selves constituted by something more than just consciousness. After all, from the fact that a self

[353] Is it also a sufficient condition? Yes, but in a trivial sense. If you take two experiences within a specious present, they are co-conscious just in virtue of being within a single specious present. On my account, co-consciousness and being within specious present are identical, see ch. 7, 151.

[354] I do *not* claim that all there is to unity of consciousness is just unity of the subject. The adherents of this position (see e.g. van Inwagen (1990), 207) seem not to take notice of the observable relation of co-consciousness. In effect, they regard as possible situations which, from my point of view, involve one subject having two separate centres of consciousness. Now, I take co-consciousness between experiences to be observable; and I define the subject and its unity in such way that they are necessarily linked to co-consciousness. But I say that what is observed - namely co-consciousness - can be conceptualized either as a two-place or a three-place relation. And I claim that the latter conceptualization is more ontologically perspicuous.

[355] Strawson's remarks about Buddhism indicate that he would be satisfied with such conclusion; Strawson (1999), 101.

could survive a break in unity of consciousness (in the sense that it would not logically follow from the fact that a break occurred that a self ceased to exist) it does not follow that it *would* survive. I admit that I do not have a proof to show that the self must persist for a long time. But neither do I see any reason to be especially sceptical about this. In my view, persistence of the self is a fundamental assumption of the same kind as existence of other minds and existence of an external world.[356]

There may be other accounts of persisting subject meeting the demands I set for such accounts beside the Dualist position I presented. But one position is enough to make the point that a theory with the required features is possible. Now, to make my argumentation complete, I will present two fall-back positions available to any Dualist who prefers to regard the self as constituted by consciousness and who wants to preserve Persistence View.

The first position may be attributed to Descartes. The self, according to Descartes, is identical to “thinking”. That is, thinking is the principal attribute “constituting” the self. Descartes “thinking” is often understood as “consciousness”. For our present purposes I will adopt this interpretation. Since Descartes thought that consciousness “constituted” the self and believed that the self persists for a long time (indeed, that it is immortal) and since he did not allow for intermittent existence of substances, he held that we are conscious at every moment of our existence. Even when one has fainted or is in deep sleep, one has some consciousness. Presumably, experiences occurring during such periods are not very interesting. They could, for instance, consist in some very vague body-feeling; or perhaps of a single visual experience of darkness. Such experiences would succeed richer experiences of the period of falling into sleep; and would be succeed by richer experiences preceding wakefulness (or, as the case may be, by vivid dreams). Since the hypothetical experiences would be so poor and monotonous, it would not be surprising that, first, they

[356] See ch. 2,

would not give rise to consciousness of passing time and, secondly, that they would not be remembered.

This position is often ridiculed. But it is interesting to note how difficult, or rather impossible, it is to empirically establish intermittence of consciousness. First of all, I do not find the idea of someone *reporting* that he is unconscious intelligible. But, it may be said, we have a good reason to believe in intermittence of consciousness if we observe that a deeply sleeping person does not have the right kind of brain-states to be conscious. But how do we know that the brain states of deep sleep phase do *not* support consciousness? We cannot rely on reports after all. We can only rely on it *seeming* to us that we were unconscious in deep sleep. But this is a very interesting kind of appearance. It is about the past, but it is not memory. For, *per impossibile,* we would have to experience the episode of unconsciousness to remember it. I suggest that this appearance consists, roughly, in awareness of the lack of connection between the last contents we remember from before sleep (or from dreams) and the present content.[357] So the underlying fact is precisely the *absence* of memory. Perhaps we could say that if we were conscious we should remember we were. But the empirical evidence suggests just the opposite. The functioning of brain regions responsible for memory is crucially affected during sleep. And it is by now well-established by science that we do not remember even the best part of our conscious vivid dreams. So there are excellent empirical grounds to doubt the reliability of our memory with regard to the periods of sleep. The only way to make a valid argument for intermittence of consciousness is to presuppose that there are some necessary links between consciousness and overt bodily behaviour. But this is to presuppose one or another aprioristic philosophical theory and not to make an argument based on empirical findings.

What may displease one about the Cartesian position is the fact that it requires

[357] The connectedness of *contents* is something altogether different from unity of consciousness; cf. ch. 1, 13. Strawson agrees; Strawson (1997), 358.

introducing conscious experiences of purely hypothetical kind. The alternative position which I now want to propose does not have such drawback. On the downside, it may seem much more metaphysically extravagant. The self, in my view, is not located in physical time at all. One of the main reason to think so is that according to standard Dualist accounts, the immaterial self is not located in space. And within the framework of relativity theory - our best physical theory concerning the matter - it simply makes no sense to allocate location in physical time without location in physical space.[358] On the other hand, the self undeniably exists in time. But this, I contend, is phenomenal or subjective time. This time-order is the order of states (or acts) of consciousness. If so, then by the very nature of phenomenal time, there are no periods of unconsciousness in phenomenal time. If the self exists in phenomenal time, it is always conscious.[359] The belief that there is a time when the self is unconscious originates with mapping the events in phenomenal time onto the physical time. This mapping indeed leaves gaps. But this has no importance for metaphysics. For the mapping gives at most the *apparent* location of conscious events in physical time, since in reality these events are not located in physical time at all.[360] The problem which remains is what to do with the apparent lack of continuity between conscious states before and after deep sleep. If the self is

[358] There are yet other reasons to think that the self is not located in the physical time. This idea can provide a non-eliminativist solution to the puzzle-cases presented by Dennett (1991), 114-125, 140-143, 162-164, 167. But the most important reason is that experience manifestly has A-time features, but there are compelling reasons to think that the physical time has only B-time features. This theme is developed in Robinson (2007). I would like to note that while my thoughts on the matter are surprisingly close to Robinson's (we developed similar ideas at roughly the same time without ever discussing them), there are some crucial points of difference. According to Robinson, the self is fundamentally a-temporal; its experiences intrinsically possess only proto-temporal features; and subjective time is "constructed" in favourable circumstances. In my view, the self manifestly changes in its intrinsic properties, to wit, it ceases to experience one thing and starts to experience another. The order of changing experiencings constitutes the subjective time-series. I submit that it is unintelligible how changes occurring in one substance could give rise to anything but a linear time-order (Robinson notes that non-linear ordering or lack of any temporal ordering of experiencings are possible on his view, *op. cit.*, 76).

[359] This solution has been proposed in Evans (1970), 184ff. Evans does not consider, however, the difference between the self being always conscious in subjective time and its having structurally continuous consciousness.

[360] This position is problematic if one accepts causal theory of time. There are two solutions. The straightforward occasionalist solution is to deny interaction between the physical and the mental realm. The second solution consist in saying that there are two complete time-orders which are partially coordinated, but there is no overarching complete ordering. Thus mental events would have an approximate location in the physical space-time, while having a fully determinate location in the phenomenal time.

constituted by consciousness, there can be no break in continuity. This is the principal issue. I will now ask the reader to think about the situation which probably occurred to him. It happens, most often when one is extremely tired, that one falls asleep and then wakes up without any sense of time-lapse between falling asleep and waking. One wakes up astonished that one was sleeping. It is as if one went directly from the last pre-sleep experience to the first waking experience.[361] There is, certainly, a disruption in the meaningful continuity of *content* of consciousness. But as regards the *structure* of consciousness (and unity of consciousness concerns only the structure of consciousness), it is almost as if the last remembered pre-sleep experience and the first post-sleep experience were directly successive. I say “almost” because I think that this is not exactly what happens. Rather, I suppose that in all cases of falling asleep (or fainting), the last clear waking experience yields to chaotic and poor experiences occurring in the transitory phase between waking and sleep. Sometimes we are able to observe or remember experiences of this kind. Similar indistinct experiences occur in the transition from sleep to wakefulness. So I think that what happens is this: pre-sleep waking experiences are continuous with experiences of the transitory falling-asleep phase; and *these* experiences are in turn continuous with experiences of the transitory phase preceding waking up, which, in turn, are continuous with post-sleep waking experiences. Now, this is admittedly as unverifiable (and unfalsifiable) a hypothesis as the Cartesian hypothesis. However, it has three advantages over the traditional Cartesian position. First, while one may be reluctant to admit regular failures of memory, there is nothing at all objectionable with saying that we cannot perform such a sophisticated operation as observing continuity of consciousness during sleep. Secondly, my hypothesis does not introduce any extra kind of experiences. It makes do with perfectly familiar experiences of the transitory phases. Thirdly, it allows one to agree that brain-states in deep

[361] I have been told one has similar impression after coming out from full anesthesia.

sleep phase do not sustain consciousness. For even if there is a temporal gap between events sustaining consciousness, there is no gap in consciousness itself.

9.1.4. Conclusions

The arguments in favour of Transience View and against Persistence View extracted from Strawson work fail to achieve their purpose. Persistence View can be defended even when we grant Strawson's controversial crucial claim that there is no deep metaphysical difference between substances and their properties or processes. In my view, Strawson neglected another distinction: that between principal attributes and modes. I have shown that it is possible to conceive consciousness as a mode of a substance which can survive breaks in unity of consciousness. I have also suggested some reasons for thinking that this is how consciousness should be conceived in any ontology that is realist about selves. Finally, even if one conceived consciousness as the principal attribute of the self, it does not follow that the self would be transient. For there are at least two ways to defend the view that human consciousness in fact does not contain any discontinuities. From the possibility that the self may survive breaks in consciousness and the possibility that human consciousness is in fact unbroken it does not follow that the self in fact persists for a long time. But neither do we have good reasons to think it does not. The Persistence View remains a tenable option. The Transience View has not been disproved either. Perhaps it is possible to conceive consciousness as a principal attribute "constituting" a genuine self. And it is possible to believe that selves (whether "constituted" by consciousness or not) are transient. I have found no conclusive metaphysical argument to settle the matter. But in the absence of such arguments, we can make a choice between theories based on other considerations. There are several good reasons to prefer Persistence View over Transience View. First, the former avoids multiplication of selves and complications with ambiguous reference of "I". Secondly, Persistence View accords with the usual experience of the self. Thirdly, it allows for us having genuine anticipation - what is denied by the Transience View. This denial is the most unpalatable consequence of the Transience View. Finally, being consistent with our experience and allowing for genuine anticipation, the Persistence View has the status of a deep belief underlying innumerable attitudes concerning our futures. In contrast, Transience View lacks, as I argued, any distinctive intuitive support.

9.2. No Self Theory

The greatest advantage of the No Self Theory is its elegant simplicity. Its dialectical position *vis a vis* non-eliminative theories of persons is, however, not at all favourable. In setting out the methodological groundwork for my work, I observed that mentally healthy adult human beings provide the paradigm example of persons, and that the force of our interest in having a relevant concept applying to human being is such that, even if we were faced with a devastating criticism of our actual concept "person", we should rather try to revise our conceptual system than consent to saying that human beings are not persons. This means that No Self Theory has to make a terribly ambitious claim: *that no possible theory of person-like objects is tenable.* It is not enough for a No Self theorist to argue that all actual theories are implausible and that No Self theory is the best theory. The best theory is not necessarily the true theory; nor an implausible theory a false one. Moreover, even if all current positive theories of the self were implausible, Agnosticism about the nature and identity of persons would remain as a preferable alternative to eliminativism.[362] The position of No Self Theory *vis a vis* non-eliminative views is strictly analogous to that of Scepticism *vis a vis* positive views about knowledge. Perhaps no actual theory of knowledge is quite plausible. Perhaps, on balance, arguments for Scepticism are much better than arguments for any positive theory. Yet even if that were the case, Scepticism is our last choice. To turn Sceptic is to give up on

[362] Perhaps the vast majority of people can be regarded as largely Agnostic about the nature and identity conditions of themselves. Despite such agnosticism, most people do not have doubts about existence of persons.

satisfying interests driving the search for a positive view of knowledge. This is to be done only when all hope of salvaging the idea of knowledge is gone.

One remark needs to be added now about the idea of revising our conceptual system. It may be possible to define persons into existence too cheaply. There is a danger that a seemingly realist view of persons could be merely verbally different from No Self Theory. How to avoid this danger? The test which realist theories should pass is this: they should preserve the interests which motivate our attachment to having an applicable concept "person" (or "person-like thing"). No Self is a theory of last resort precisely because it gives up on those interests. If a seemingly realist theory likewise gives them up, then either it is merely verbally different from No Self Theory, or at least it does not enjoy any advantage over No Self Theory. Moreover, its analysis or revision of the concept "person" would be unmotivated, since it does not keep faith to the interests we had in framing the concept. The clearest example of such central interest is preservation of genuine anticipation. I have argued that some theories fail to preserve genuine anticipation. And I suggested some reasons to think that they differ only verbally from the No Self Theory. We may now see these suggestions in the light of a broader methodological framework.

In practice, No Self theorists do not try to establish the terribly ambitious claim that no possible realist theory of the self can be true. The usual strategy seems rather this. First, a general metaphysical framework is established, or, more often, assumed. This, nowadays, is the truth of a modern Naturalist view of reality.[363] More precisely, what is presupposed is the general correctness of the Naturalist Reductionist *picture* of the world (but not the realist Reductionist metaphysics).[364] Then it is argued that there is no place for

[363] Traditional Buddhist theories started with establishing quite different metaphysical frameworks.

[364] I use "Naturalist" rather than "Reductionist" here to avoid confusion. No Self Theory is, from one point of view, an anti-Reductionist theory. Yet all the modern No Self theorists I know of (Unger, Stone, Giles, Metzinger and Dennett *if* he is a No-Selfer), think that the vision of reality associated with Reductionism is correct - i.e. that there are no metaphysical " further facts" about higher-level entities. This is what I mean by "Naturalism" in this section. Now, unlike the Reductionists, No-Self theorists may think that as there are no

persons in such scheme of things. So, for instance, early Unger starts with the assumption that the most likely candidates for persons would be materialistically conceived human beings. But, he argues, given the nature of physical reality and given the nature of our sortal concepts (which are "vague discriminatory concepts"), there can be no human beings, nor indeed anything much like human beings[365]. Similarly, Stone assumes that a Naturalist vision of reality is true and then argues that any constructivist Reductionist theory - like that of Parfit - cannot produce anything which would satisfy well enough the concept "person".[366] Unger's and Stone's work nicely complement one another. Early Unger attacks the Physically Based approach to persons. Stone's arguments concern primarily the Psychological approach. We can see another complementarity. Unger can be seen as arguing that no true unity can be found in complex physical objects. Thus, his arguments would be addressed to people with robustly metaphysical tendencies. Stone, on the other hand, considers more conventionalist or constructivist approaches. So even if no single No Self theorist refutes all alternative theories, collectively they may cover a good portion of the field (even when the field is cut up in different ways). However, assuming the general correctness of the Reductionist picture of reality leaves out a big gap in the No Self theorist's argumentation. This gap is inhabited by the Simple View. Even if all the arguments by No Self theorists against Reductionist theories were sound, this would not suffice for establishing No Self theory. For the same arguments would work for the benefit of the Simple View. If Reductionism were shown false, then one should choose *either* No Self Theory *or* Simple View. In fact, the Simple View would benefit *more* from anti-Reductionist arguments than No Self theory. For if Reductionism collapsed into Eliminativism, this would be an excellent reason to rethink the assumptions leading to an unwelcome conclusion. This is not an example of wishful thinking. The issue turns on how well the fundamental metaphysical framework is established. Purely logical arguments very

"further facts", there are no higher-level entities either. I strongly sympathize with this ontological approach.

[365] Unger (1979*a*); Unger (1979*b*), Unger (1979*c*).

[366] Stone (1988); Stone (2005).

rarely settle anything in metaphysics. Theories are chosen on the basis of their seeming *plausibility*. But plausibility is a foggy idea indeed. Most often it comes down to the notion of a view being more or less faithful to some favourite *intuitions*. And when it comes to that, intuitions about persons have as much right - or maybe more - to have a hearing as any other intuition.[367] As I argued throughout the book, Reductionism and, *a fortiori,* Eliminativism violate our intuitions, while the Simple View has the best overall intuitive support. Thus the better anti-Reductionist case the No Self theorist makes, the stronger the case for embracing the Simple View.

Thus it is imperative for the No Self theorist to prevent the move to the Simple View. But this is an exceedingly difficult task. First, in the fold of the Simple View one can find radically different theories. It will suffice to mention just a few. There are materialist Constitution views of Lowe and Baker. There is Chisholm's "particle hypothesis". And, of course, there are multifarious Dualist theories. Given this variety, it would not suffice to reject just one general metaphysical framework (like Dualism, or the robustly metaphysical approach to unity of objects). For it seems that a Simple View can be stated within any such framework. What makes the task of refuting the Simple View even harder is the increasingly fashionable agnosticism or open-mindedness of philosophers about the ultimate nature of physical reality. At this point the task of the No Self theorist ceases to be merely Herculean and becomes Sisyphean. For rejection of any particular Simple View account would give us a reason to re-think its particular metaphysical assumptions. In effect, a methodologically motivated Agnostic Simple View would be permanently available to a friend of self. In addition, there is always available the dogmatic Agnostic Simple View position: taking the belief in one's existence and identity as epistemologically basic beliefs, impervious to

[367] Unger seems to think along similar lines when he speaks of the "humanly realistic philosophy"; Unger (2006), 36-40.

considerations about the nature of the self.[368]

At this point the No Self theorist may lose patience and protest that Simple View just does not make sense. We will hear a Humean appeal to experience and an allegation that Simple View is a confused metaphor either based on an illusion or not based on experience at all.[369] Yet the accusation of not keeping to phenomenology can be easily turned around. To say that there is no clear experience of the self is simply to ignore the work on experience of the self done by so many psychologists and philosophers.[370] Doubts about self-experience seem always to rest on misconceptions. With regard to experience of self-presence, it may not be noted that this experience concerns the structure and ontological categorization of consciousness and not a specific content in consciousness. The claims that self-perception is illusionary seem to ignore the possibility of having a naive realist view of self-perception. As for experience or thinking of the self as existing at other times than present, the prevailing habit of thinking in four-dimensionalist terms tends to obscure the fact that nothing more is needed for thinking about ourselves at other times than ascription of tensed properties to the present self. Admittedly, how ascription of tensed properties is to be understood may be a matter of further debate.[371] I favour the A-theory interpretation and the view of the self as a simple enduring substance. But then, B-theory can muster arguments too. And, possibly, B-theory may require temporal parts theory. And, possibly again, temporal parts theory cannot be squared with any sensible idea of the self as a truly unified simple substance. If we were constrained to go down this path, then the Simple View would seem to be in some trouble. However, such a dark day could never arrive. The debate between A- and B-theory is one of the perennial metaphysical disputes. In this case, my earlier point about plausibility and

[368] To use the analogy to scepticism again. Having an agnostic view about the self is strictly analogous to holding to the belief that there is a physical world while being ready to discard any particular theory about it. Somehow, the latter position does not strike most philosophers as awkward.

[369] Giles (1997), 27, 31, 38-40, 44-47.

[370] Neisser (1988); Strawson (1997); Martin (1998) - to mention just a few authors who can hardly be suspected of being corrupted by the Simple View.

[371] See e.g. Sider’s temporal counterpart semantics of tensed predication in Sider (2001), 190-199.

intuitiveness of theories can be brought in. Finally, even if it was shown that the very notion of continuing substance is defective, this would not exorcise the spirit of the Simple View. For in such case, indeed, a deep revision of our metaphysical thinking would be required. The most far-reaching suggestion along this line is that the “I” should be accorded a unique ontological status.[372] Or, we could settle for some non-substance notion like “process” as the fundamental ontological category. But then we could hold the view that there is a mental process which is unified, has determinate limits, and is necessarily separate from other such processes. For all intents and purposes, such a view should be quite satisfying to anyone with inclinations for the Simple View. The indefinability of personal identity, which is the defining feature of the Simple View, is not really what this view is about. The motivation for this view, as I see it, is to have a self which is truly unified, well-delimited (with no indeterminacy), separate from others, which can genuinely anticipate its future and so on. Since Naturalist Reductionism fails to preserve such features of the self, it has to be rejected. This means that giving a reductive definition of identity has to be opposed. But this is just a means for preserving the important features of the self, and not the aim. If, for example, it would be possible to produce a definition of personal identity in terms of consciousness such that it would preserve the most important features of the self, then such account would provide almost all that I care about. There would remain, however, the issue of *agency*.

At this point we may ask what is the difference between the Cartesian view of the self as “constituted” by an unbroken consciousness and a No-Self view according to which there is, likewise, long-term continuous consciousness and the streams of consciousness are as separate and well-delimited as Cartesian Selves? I suggested that Strawson’s Transience View is only verbally different from No Self theory. Does the same hold for the Cartesian view? I think not. First, even if one, like Strawson, denied deep difference between processes

[372] This is by no means an unpopular view; in post-Kantian continental tradition it may be the dominant view. In analytical philosophy it is explored for example by Maddell (1981) and van Frassen (2005).

and substances, one could hold that the concept of substance is analytically related to *persistence*. We say there is a substance when we see diachronic unity-in-diversity. Thus, long-lived Cartesian selves, but not Strawson's transient selves, would deserve to be called substances. Secondly, categorization in terms of substance allows one to ascribe mental powers - among them the power to act - to the subject. Ascription of extensive mental powers is essential to our understanding of the self. I find it hard to see, however, how such powers could be ascribed to a self identical to a specific conscious event, or even to a continuous stream of such events. The difference between the Cartesian self and the No-Self stream is that although in both cases consciousness is the only *non-dispositional* property, the Cartesian self may have many dispositional properties beside. Of course, the stream of consciousness could develop one way or another. But this does not amount to ascription of dispositions to the stream. For what makes it true about the stream are dispositions of *other* things (like brains, environment etc) which determine the way the stream of consciousness develops. This also makes it hard to ascribe agency to the stream. The Cartesian self, on the other hand, satisfies requirements of agency and other requirements imposed by the concept of person.

Chapter 10. The Simple View

10.1. Results so far

The previous chapters were devoted mainly to the criticism of Complex Views, Transience and No-Self views. But in the course of argument there also emerged a number of positive conclusions. Let me bring these together now.

Personal identity

(1) The shared concept "person" is unspecific. No theory can be established on the basis of straightforward conceptual analysis.[373]

(2) Practices of identifying persons do not provide argument in favour of any theory. Physical continuity may be treated as mere evidence of personal identity in usual cases.[374]

(3) Emotional identification with one's past/future/possible self depends on the perception of the psychological similarity of the present self and target-self and on the meaningful link between their lives. It does not depend on the judgment of numerical identity. Judging that one is identical to the target-self is at most a necessary, but not a sufficient condition of emotional identification.[375]

(4) Self-identification across time presupposes the grasp of the flow of personal time.[376]

(5) No use of third-personal descriptions is involved in (i) our understanding of the reference of "I", and (ii) our implicit understanding of ourselves and (iii) self-recognition in imagined and remembered situations. Imagination and memory involve a "projection" of the present self into the imagined/remembered situation, or in other words, seeing oneself as involved in the situation in the relevant way. This makes self-recognition automatic and immune to error

[373] Ch. 2, 63f.
[374] Ch. 2, 34-37.
[375] Ch. 4, 93-95; 104.
[376] Ch. 6, 141.

through misidentification.[377]

(6) We can track our identity in possible worlds by starting from a first-personally given experience and tracking potential continuity of our consciousness. Without such anchoring in an actual experience, we do not know how to answer questions about TW-identity of persons.[378]

(7) Our self-experience and self-concept do not essentially contain any individuating features which would allow us to answer questions about our TW-identity.[379]

(8) The only shared intuition about the numerical identity of persons across time is that a person is a locus of possibly continuous consciousness. This intuition is expressed by the following necessary condition of personal identity:

(C) For any persons x, y s.t. x exists at t_1 and y exists at t_2 ($x = y$ only if there is a *possible* (or actual) series of conscious states (A,..., B) such that:
(i) x has A at t_1 & y has B at t_2 & the states between A and B occur at times between t_1 and t_2
(ii) every two successive states C and D in the series (A,..., B) *either* are (weakly) continuous *or* there is a possible series of (weakly) continuous experiences (C*,..., D*) s.t. C* is had by the person who has C and occurs at t(C) and D* is had by the person who has D and occurs at t(D).)[380]

(9) Anticipation consists in "projection" of the present self into a future situation, or, equivalently, in ascription of tensed properties to the present self.[381]

(10) Identity is the necessary and sufficient condition of appropriate anticipation:

(SA) I will exist at t iff there are some possible mental states at t that I can anticipate having.[382]

[377] Ch. 6, 141-144.
[378] Ch. 4, 102 ; ch. 6, 149f
[379] *Ib.*
[380] Ch. 7, 159.
[381] Ch. 8, 197f.
[382] Ch. 8, 203.

The subject (the thinking substance)

(1) Persistence of the substance is necessary for personal identity. [383]

(2) The subject is not something which thinks only indirectly or derivatively.[384]

(3) The subject can be directly experienced. There is no good reason to regard this experience as an illusion.[385]

(4) The subject can be appropriately conceived on the basis of experience as a thing able to survive breaks in consciousness. It can be conceived as a substance which (i) is characterized by a peculiar and irreducible unity and (ii) whose nature is characterized by (or consists in) potential for continuous consciousness, various mental powers and activities.[386]

(5) The subject can also be appropriately conceived as a substance "constituted" by consciousness. This view is consistent with Persistence View of persons on the assumption that our consciousness does not have gaps. One can choose between two versions of this view: the traditional Cartesian view and the phenomenal-time-based view.[387]

(6) The embodiment of the subject can be conceived as lawful. A rough but plausible criterion of embodiment is:

(E) One is embodied as long as one's brain continues to exist and to have the capacity to support unified consciousness.[388]

Consciousness

(1) Absoluteness of consciousness is consistent with the data of science. Spectrum of Decomposition provides no valid objection to absoluteness of consciousness.[389]

[383] Ch. 3, 68f; ch. 4, 94; ch. 5, 114-117.
[384] Ch. 5, 127f.
[385] Ch. 8, 199-202; ch. 9, 238.
[386] Ch. 9, 225-228.
[387] Ch. 9, 229-233.
[388] Ch. 7, 171.
[389] Ch. 2, 44ff.

(2) Streams of consciousness are necessarily indivisible (no Fission) and separate (no Fusion).[390]

(3) Unity of consciousness requires sameness of subject.[391]

The overall picture which emerges from these conclusions favours the Simple View. Realism about the subject and the substantivalist view of the subject seem to be correct. Our experience and concept of the subject tell us no more than that it is a locus of potentially continuous unified consciousness. The features of consciousness - determinateness, indivisibility and separateness - are uncongenial to Reductionism.

10.2. Intuitive motivation for the Simple View

The common tenet of the Simple View or Primitivism is the un-analyzability of personal identity due to the selves' position among the ultimate components of reality. There are three basic pro-SV intuitions.

(A) Questions about personal identity necessarily have determinate and convention-independent answers. This claim has various sources: considerations of anticipation, metaphysical arguments, conception of the relation between the subject and consciousness. Complex View Reductionism seems to lead inevitably to indeterminacy of personal identity. So it should be rejected in favour of the Simple View.

(B) Consciousness is all-or-nothing; streams of consciousness have determinate limits, are indivisible and separate. Arguably, subjects share these features. Again, only the Simple View seems consistent with this fact.

[390] Ch. 7, 161f.

[391] Ch. 7, 163; see also section 10.5. below. (3) does not exclude the possibility that unified consciousness could be shared by many selves each of which would be a subject of all unified experiences. What is excluded is that one subject has only a part of the unified consciousness and another subject another part.

(C) Imaginability of Body-Exchange, Reincarnation and similar cases. There are two ways to argue for Simple View on this basis.

> (a) The standard route is to argue directly from imaginability to possibility. Since a given self can have different bodies and lives, its identity cannot depend on any physical or psychological facts, but must be something *sui generis.*
>
> (b) The cautious route, which I have taken, is to take imaginability of such cases to show the nature and extent of our grasp on the nature, persistence conditions, and TW-identity of the self. In this way, cases contribute to the argument for a justified *conception* of the self.

That much is common to Simple View theorists. Particular versions of SV will be motivated by diverse intuitions.[392] Special Ontology theorists rely on the appeal to the absolute difference between third-person and first-person perspectives and on the claim that identity of the substance is not necessary for personal identity. Materialists couple materialist intuitions with anti-Reductionism. Dualists, finally, can appeal to a particularly rich assortment of intuitions and arguments. The following are noteworthy:

(D) Appeal to self-experience. The Dualist conception of the self is phenomenology-based. According to this conception, the self is:

(i) embodied but distinct from the body,
(ii) essentially mental,
(iii) a directly involved, non-derivative thinker,
(iv) immediately accessible
(v) strongly unified with consciousness and not something which merely sustains or

[392] Cf. 10.3. below

"realizes" consciousness.

The Dualist may then argue that the self so conceived is a substance with a complete essence.[393] *Adding* further attributes to the self will result in a flawed notion of substance and in the impossibility of explaining the relation between the subject and thinking.[394]

(E) Anti-Reductionist arguments concerning the unity of the subject. Two most promising arguments of this sort are the Homunculism Problem and Unger's Problem of the Many.[395]

(F) Considerations of the relation between self and time. Arguably, selves do not exist in the physical time, but have their own subjective time.

To sum up, the Simple View is motivated by investigation of self-experience, relation between the self and consciousness and by anti-Reductionist metaphysical considerations. Imaginary cases may play an auxiliary role.

10.3. Varieties of the Simple View

Many different theories count as Simple Views. Here, along with a tentative classification, I offer some suggestions about the main problems facing these theories. Although the primary aim of my work is to argue for the truth of the Simple View as such, I want to offer some arguments for preferring the Dualist version.

[393] See Hawthorne (2007).
[394] Plantinga (2007), 106-188.
[395] Unger (1980), Unger (2004).

10.3.1. Special Ontology Approach (SOA)

Proponents of this approach deny that persons are substances.[396] The principal problem is to say *what* they are. SOA theories run two risks. The first is lack of intelligibility. The second is the lack of clear distinction from Psychological Theory on the one hand and from substantival Simple View on the other. If the self is equated with, say, a structural feature of consciousness, then this view can plausibly be stated as a version of PT. Or suppose that the self is regarded as a special primordial activity, and selves are ultimate, individual somethings capable of independent existence. This view can plausibly be construed as affirming substantiality of selves, while advancing radical claims about the nature of our substance and about relations between ontological categories. The second problematic aspect of SOA is the rejection of the thesis that sameness of substance is a necessary condition of personal identity. It is this thesis that provides the strongest argument against PT. In the absence of this argument, a consciousness-based version of PT could well turn out to be more plausible than SOA Simple View. Special Ontology should be the last choice of a SV theorist.

10.3.2. Dual Aspect SV: selves as relatively simple material substances

According to this theory, selves are substances which have both mental and physical properties (hence “Dual Aspect”), but they are not to be identified with bodies or similar gross physical objects. They are mereologically simple material substances, or at least they are mereologically *rigid* i.e. they cannot change parts.[397] Chisholm suggests that a self may

[396] Maddell (1981), 134-139; van Frassen (2005), 87f (but see *ib.* 110).

[397] Chisholm (1989), 125f. E. J. Lowe’s theory could be classified as Dual Aspect theory, given Lowe’s central claims: selves are merologically simple and they can have both mental and physical properties. We may also be tempted to classify Lowe’s theory as Emergentist Naturalism, given his description of the emergence of selves in the natural order, Lowe (1996), 47-51. There are, however, crucial difference between Lowe’s position and these theories. First, Lowe’s selves are *psychological* substances and do not possess any physical properties *essentially*, *op. cit.* 32f. Thus they are not ordinary physical substances, even if they are not essentially wholly immaterial. Any physical properties they actually possess, they possess only in virtue of having a body. Selves have physical properties by courtesy so to say. And the relation of the self to the body is not that of constitution or a similar relation, but is depicted on broadly Cartesian lines of special

be a microscopic particle or sub-particle located inside the brain. One problem with this suggestion is empirical: *all* matter composing my brain is apparently exchanged every few years or so. If I am a particle inside the brain, then I should inhabit my body no more than a few years; but this does not seem to be the case. The obvious conceptual problem with Dual Aspect is that the relation between the material nature and the mental nature of the person seems inexplicable. This problem may be common to all Materialists, but Chisholm's theory faces it in a particularly stark form. The connection between the actual material and mental properties of the person seems absolutely contingent, given that mental properties cannot be identical to or supervene on brain-states; rather, they must be conceived along the Dualist lines but located in the person-particle. More importantly, Dual Aspect theory cannot provide a plausible conception of the substance. If the substance is nothing over and above the essence, and Dual Aspect theory presents us with two wholly different essences, then the Dual Aspect substance has no internal unity.[398] It is impossible to say what makes it the case that this is *one* thing and not two. If, on the other hand, the substance is a unifying factor over and above the essences, then it seems to be just a "featureless" peg for essences. This is not the most attractive conception of the substance, to say the least.

10.3.3. Emergentist Naturalist SV

On this view, selves are objects related to material objects (bodies, animals, masses of matter etc.) by constitution or a similar metaphysical relation. For various metaphysical reasons, proponents of such theories reject the possibility of giving necessary and sufficient conditions of persistence of selves in terms of lower-level entities. This view is amply represented in the Aristotelian tradition. Recent theories of this kind have been developed by Baker and Lizza

relationship of the self and its body in action and perception, *op. cit.* 37f. In view of these factors, it seems best to classify Lowe's theory as non-standard Substance Dualism.

[398] If causation between the purely mental and the purely physical is thought problematic because of the intuition "But they are so different!", then substantial unity of the mental and the physical ought to be felt as at least as problematic.

among others.[399]

Emergentist SV faces the same problems as Chisholm's theory, yet the proponents of this approach can draw on the mainstream Materialist work in philosophy of mind. Thus the problem of the contingent relation between the physical and the mental can be alleviated, if not solved. On the other hand, Emergentist SV faces a problem avoided by Chisholm: indeterminacy of identity engendered by mereological complexity. Let me illustrate this on the example of Baker's theory. According to Baker, capacity for first-person perspective is an essential property of persons. One has the capacity for the first person-perspective at *t* if either (i) one has manifested the first-person perspective some time before *t* or (ii) one is at *t* in an environment conducive to the development and maintenance of the first-person perspective.[400] First-person perspective itself is an ability. A person with first-person perspective survives as long as her perspective survives. There is no indeterminacy about it. So Baker claims that her account escapes indeterminacy. Unfortunately, *having* a first-person perspective is *not* necessary for being a person. What is necessary is the *capacity*. Consider now the spectrum:

Spectrum of birth-circumstances

Being in the womb is not to be in the environment conducive to development of first-person perspective. But one can imagine a range of environments into which one can be born which will be *more* or *less* conducive to development of first-person perspective. For simplicity, consider who takes care of the child. On the positive side of the spectrum, we have healthy and loving human parents. Less fortunately, one's parents can be humans with more or less severe mental disabilities. Then, one can be adopted by chimpanzees or wolves. Finally, all one's physiological needs are attended to by machines, with no linguistic or emotional communication going on. It is clear that the cases on the far positive side are conducive to the development of first-person perspective and the cases on the far negative side are not. But there will be cases in the middle when we will not know what to say, and where it might be a matter of pure chance whether the capacity would develop or not. So I conclude

[399] Baker (2000); Lizza (2006), 74-93.
[400] Baker (2000), 92.

that it is *indeterminate* when an environment is “conducive to the development of first person perspective”. But if so, there will be beings about which we could not say determinately whether or not they are in right circumstances to have the capacity for first-person perspective. Therefore, there will be cases when it is indeterminate whether an *x* is a person or not.

This argumentation shows a general point. If one makes the *capacity* for mental functions essential to personhood *and* one makes standard materialist assumptions about the dependence of mental capacities on the states of the body, then indeterminacy almost inevitably follows. It would not follow if “capacity” could be rid of vagueness and if we had precise laws governing the relations of brain- and mind- states. Such conditions seem unlikely to be fulfilled. If definition in terms of capacities leads to indeterminacy, perhaps one should focus on *potentiality*. Yet this too would not get the Materialist out of indeterminacy. For there to be a potentiality there has to be a material object having it. But the beginnings of complex material things are gradual and Sorites-type arguments apply to them. It is not fully determinate at which point a material being starts to exist. Secondly, consider resuscitation of corpses. In some cases we have no doubts that we have resuscitated *the same* person. But we can devise a spectrum of cases when my body is more and more damaged, more and more torn apart and scattered and more and more technical refinement is required to bring it back.[401] Now, there are two questions: at which point restoring the same body ends, and constructing a new body from old material starts? Secondly, is the person resuscitated *really* the same person as the one which previously lived? It would be ridiculous for a Materialist to insist that the first question has a determinate answer. But then it is not determinate (i) when my body ceases to exist and (ii) when I have the potentiality to be resuscitated and when I lose it. But then it is indeterminate when I lose the potentiality for having some select mental function, so it is indeterminate when I, the person, cease to exist. Secondly, given the indeterminacy of my body’s survival, I see no further *fact of the matter*

[401] Unger (1990) is a veritable mine of such spectra.

that could settle the second question.[402] In cases when much of my body and brain is reconstructed, it will be indeterminate whether the emerging person is me or not. I conclude that standard Materialism (almost) inevitably entails indeterminacy of questions of personal identity. The advantages accruing to Emergentist SV from drawing on mainstream Materialism come with too high a price. Accepting indeterminacy undercuts the main motivation for SV. Emergentist SV is not an attractive option.

10.3.4. Substance Dualism

Some variants of Substance Dualism have been discussed in ch. 9. On these views, the self is a mereologically simple immaterial substance. This has three advantages. Far-reaching revision of ontological categories is not required. Indeterminacy can be avoided. The problem of the unity of substance with dual nature does not arise. Thus Substance Dualism can avoid the problems facing other SV approaches.

Substance Dualism has been widely criticized. Most of anti-Dualist arguments have little to do with personal identity. The best of them concern specific issues in philosophy of mind.[403] There are, however, three issues related to personal identity that are worth discussing. The first is the relation between the subject and consciousness. The second is Subject-Transfer. The third is the doubt whether Substance Dualism provides a credible account of *persons*. Can the simple self carry the burden of our ethical and emotional attitudes? Since the answer to the last question has already been sketched in ch. 3, in the sections which follow I will address only the first two issues.

[402] Of course, there are facts about the psychology of the resulting person. But to make these determine the person's identity is to embrace PT, which is committed to indeterminacy anyway.

[403] Plantinga provides a useful overview and discussion of most popular anti-Dualist arguments. Plantinga (2007), 12-136.

10.4. Subjects and consciousness. Determinateness, indivisibility, separateness.

What are the reasons to think that selves, like consciousness, are indivisible, separate and do not suffer from indeterminacy? Take determinateness first. The most popular line against indeterminacy appeals to anticipation. The standard reasoning goes as follows:

(1) Anticipation traces identity.
(2) If indeterminacy of identity were possible, then in some cases it would be indeterminate whether I can appropriately anticipate some future mental states.
(3) Since indeterminacy is a linguistic or conceptual phenomenon (there is no indeterminacy in things), it can be removed by adopting an appropriate convention.
(4) Whether or not I can anticipate some mental states can be decided by a convention.

This conclusion is strongly un-intuitive. Using the ideas developed in ch. 8, we can offer an argument to a stronger conclusion:

(1) Indeterminacy of identity and of anticipation can be understood only within the Reductionist framework.
(2) Reductionist accounts of identity cannot explain why extension of anticipation to mental states of other people is absurd.
(3) Therefore the Reductionist accounts do not provide justification for the distinction between appropriate and inappropriate anticipation.
(4) Therefore these accounts do not preserve this distinction.
(5) Therefore, if Reductionism were true, all anticipation would be inappropriate.
(6) One cannot identify with X if one cannot anticipate X's experiences.
(7) Persons are essentially something one can identify with.
(8) One cannot identify with "persons" as conceived by Reductionism.
(9) "Persons" as conceived by Reductionism are not genuine persons. There are no persons in the Reductionist world.

Another way to argue against indeterminacy is to question the ontological assumptions which make it intelligible. Admission of indeterminacy can rely on 4D ontology.[404] Within this framework, we can say that an object (tenselessly) exists, but its temporal *borders* are indeterminate. This makes good sense. Indeterminacy of identity is reduced to the intelligible case of indeterminacy of parthood relation. Alternatively, as in Lewis' account, it is reduced to ambiguity of reference between many well-delimited objects. Now consider enduring, *wholly present* objects. Such objects exist at times. Take thing A and time t. From the principle of the excluded middle we get the proposition "Either A exists at t or A does not exist at t". Now, it makes no sense to say that A somewhat exists at t; or that it exists to some degree but to some degree it does not. Existence is not a property for which such moves make sense. To get indeterminacy in 3D world, one would have to reject the principle of the excluded middle for the case of existence. This is unintelligible. What may obscure the fact that 3D world excludes indeterminacy is the way that cases are usually put: "At t1, A exist. At t2, there is *a* thing which is a candidate for being A, call it B. It is not determinate whether B is identical to A". However, when we pause to think about *temporal* properties of the things, we will see that Evans-style arguments apply to this case. A has the property F of determinately existing at t1. If it is indeterminate whether B existed at t1 - and that must be meant by saying that it is indeterminate whether it is identical to A - then B does not have F, and so it is not identical to A. No doubt, it is possible to adopt a formal construction circumventing these arguments. But whether it would result in an intelligible account of existence and temporal properties of things is dubious. The only sensible option is to say that in the cases of the indeterminacy, there are lower-level things which are candidates for making up person A, but our concepts leave it indeterminate whether they do. But this way of looking at things is not just Reductionist; it belongs to Logical Construction approach. It means that there is no *fact of the matter* whether person A exists at a time. Its mode of

[404] Accepting 4D ontology means accepting temporal parts. Identification of persons with temporal sums is not required.

existence is concept-dependent; and that is why the principle of the excluded middle may fail to hold. To conclude, acceptance of indeterminacy requires either 4D ontology, or Logical Construction ontology. What is common to both is the idea that at a given time there are parts of persons, in particular mental states, which can be individuated (metaphysically and epistemically) independently of persons. This is a highly questionable assumption.[405] And there are independent reasons to reject these approaches which I have discussed in my work.

Indivisibility and separateness of selves means that fission and fusion of selves is impossible. This is a consequence of the thesis that streams of consciousness can neither divide nor fuse. Consider the following argument:

(1) Fission of a subject within a specious present is impossible.
(2) Every conscious moment is within at least one specious present.
(3) Therefore Fission at a time when the subject is conscious is impossible.
(4) If Fission of a subject of consciousness is possible at all, it should be possible at a time when the subject is conscious.
(5) Therefore Fission of a subject is not possible.[406]

To say that the subject could split, but only when unconscious, seems unbearably *ad hoc*. Reductionism cannot plausibly explain why fission of consciousness and fission of actually conscious subjects is impossible. Therefore it has to be rejected.

Given that there are good reasons to think that selves have determinate limits of existence and are indivisible and separate, we can demand of theories of personal identity to explain why selves have these features. Does Substance Dualism meet this demand? The case is obvious for the Cartesian accounts where selves are "constituted" by consciousness. On such accounts, 'subject' and 'consciousness' are but two names for one thing. Since consciousness has the required metaphysical features, so do selves. Accounts on which

[405] See Schechtman (1990), 79-86; Lowe (1996), 25-32; Slors (2001), 64-81.
[406] My thanks to Howard Robinson for suggesting this way of presenting my ideas.

subjects transcend consciousness can also provide straightforward explanation. Let us start with indeterminacy. First, selves are mereologically simple ultimate constituents of the world. Persistence of such things is a primitive feature of the world.[407] And on the fundamental level of reality there is no indeterminacy. Indeterminacy is intelligible only in the case of complex things. Secondly, according to Dualists, the relation of mental states to the self is that of properties (or modes) to the substance which is their primary bearer. Thus, like with all other properties, their identity is metaphysically dependent on the identity of the substance. So again, one of the necessary assumptions for intelligibility of indeterminacy is ruled out by the general metaphysical scheme. Fission and Fusion are ruled out for similar reasons. Such scenarios are conceivable only for complex things or at least for things which are made of some *stuff*. Try imagining a simple thing splitting and becoming two. How can one distinguish this scenario from the one where a thing is annihilated and two new things come into existence? There is absolutely no difference. One can distinguish between these scenarios when one imagines that the thing is made of some stuff and the post-Fission things are made from the same stuff. Then the thing does not survive, but something of it (the stuff) does. So we can say that it is a case of Fission, and not annihilation-cum-creation. Now, the Dualist maintains that the self is mereologically simple and is not made from any kind of stuff. Fission and Fusion of such things are not intelligible.

This discussion highlights the crucial point about the (sensible) Dualist conception of the subject of consciousness. The subject is neither a Kantian unknowable "principle"; nor is it made of an unknowable stuff "underlying" consciousness. The Dualist conceives the subject either as substance "constituted" solely by consciousness, or as substance "constituted" by purely mental attributes (or powers or activities or processes) such that consciousness is conceptually and metaphysically dependent on them (being a mode of

[407] Oderberg (1993), 143-146; Lowe (1998), 164-173.

an attribute, or being a state of activation of a mental power etc.). Thus the Dualist self is directly knowable and strongly unified with consciousness. Anti-Dualist arguments often rely on ignoring this conception. This is the case with Subject-Transfer.

10.5. Subject-Transfer

The idea that the person is identical to a substance which is a subject of consciousness underlies most of my work. I have briefly argued for it in ch. 5.[408] Here I want to meet head-on the most popular objection to this thesis:

> (ST) The substance is something which "underlies" or "sustains" consciousness. We can imagine our consciousness flowing continuously from one substance to another. Such change need not be registered in consciousness. But since our consciousness would survive, we would survive the change of substance. Therefore, we are not identical to substances "underlying" consciousness.[409]

One way to show that (ST) is wrong is to go through a reasoning already familiar from my discussion of Fission. Take two subjects S1 and S2 for supposed consciousness-transfer. Let S1 cease to exist at t_1, S2 start to exist at t_2. There should be continuity of consciousness between S1's and S2's consciousness. So let A be the last experience of S1 at t_1, B the first experience of S2 at t_2. A and B are directly continuous. Now, this means that A and B are *experienced together* (as co-conscious or as flowing one into the other). So in order to experience A as it is actually experienced, S1 would have to experience B. But B occurs when S1 no longer exists and is supposed not to be experienced by S1! The idea of subject-exchange is absurd. This vindicates my thesis:

[408] 114-117; 127f.

[409] Perry (1978), 9-17; Dainton (2008), xv, 18-20. Although Dainton envisages transfer between "underlying" substances - and takes the soul to be a possible substance of this sort - he does not envisage subject-transfer. Dainton too holds that unity of consciousness entails sameness of the subject, and that the subject is a substance (which he conceives along Constitution PT lines). Thus he is not involved in the absurdities of subject-transfer.

(CS) If a person's consciousness flows on continuously, the person persists.

Whence the mistake? Consciousness is apparently conceived in (ST) as *process*. In general, processes can be carried out by many subjects. Take "the carrying of the coffin". The coffin may be passed from one team of carriers to another. And yet it makes sense to say that there is just one carrying of the coffin. It is this sort of picture, I guess, which is presupposed by (ST). Why is it wrong in the case of consciousness? The root lies in a vague or mistaken conception of the relation between consciousness and the substance. Consider available options. First, identity. On this Cartesian-Strawsonian view, the stream of consciousness is identical to the subject (thinking substance). But then bodies and souls have to be things which just externally "sustain" consciousness and are *not* subjects. So we have no Subject-Transfer on this option. Moreover, the Dualist does *not* conceive the soul as something which externally sustains consciousness. The soul is just the substance-subject. If consciousness is a substance in its own right, then *this* is the soul; no further substance needs to be postulated. So the identity option presents no problem for Dualism. The second option is to construe processes either as properties of substances, like Aristotle, or as strings of events construed as instantiations of properties in substances at times.[410] Let us say for short that on such views consciousness consists of properties. Now, if *my* consciousness consists of properties of *this* substance, then it can no more transfer to another substance than the particular redness of *this* pen can transfer to another substance. Consciousness conceived as trope is not transferable. But perhaps consciousness could be a *complex* process, consisting of sub-processes in many substances? The singleness of this process would consist in continuity. This, however, would require that conscious events in different subjects be linked by co-consciousness. But, first, co-consciousness seems not so much an external relation between two discrete entities or events, as the phenomenon of *fusion* of experiences into a real unity.

[410] Swinburne (2007), 142.

It is sensible to think that this unity results in one thing or that there is a single experience whenever the unity of consciousness is present.[411] Yet it makes little sense to think that the tropes of two substances could fuse into one trope. So much for clear options. If processes and their relation to substances were construed in some yet other way, then from the Dualist perspective, such conception would dispose of genuine subjects. Substances could at most stand in some vague metaphysical relation of "sustaining" or "realizing" to consciousness. If they neither *are* conscious nor *have* consciousness in some robust ontological sense, then their relation to consciousness is merely external. (ST) based on such conception would be either question-begging or spurious. The reasons given for the non-Dualist conception of subjects would be the only thing that mattered. But actually the reason probably forthcoming would be... (ST) itself. The opponent of Dualism may say: "We can imagine body- and soul-transfer. So evidently we do not *conceive* ourselves as bodies or souls. Didn't you use the body-transfer argument against PBA? So why can't we use soul-transfer against Dualism?". Now, a lot can be said against the alleged analogy between body- and soul-transfer arguments. But let it suffice here to say that while it is relatively obvious what the body is, what is meant by 'soul' is not. If what is imagined in the soul-transfer scenario is just some substance "underlying" consciousness, then I happily acknowledge we can imagine the transfer and that we do not conceive ourselves as anything like that (be it a material or an immaterial thing). But the fact that we do not conceive ourselves as *some* kind of substance does not show that we do not conceive ourselves as *any* kind of substance. Both self-experience and common-sense tell us that we are *things*. What kind of things? Well, things which are *direct* subjects of thinking. This is what a subject, what a "soul" is. Whether it is material or not, and in what relation it stands to our bodies are further questions. Now, to coherently frame a transfer scenario featuring properly conceived subject would be no mean

[411] Strawson (1999), 124; Tye (2003), 25-35.

feat. If we experience and conceive our self as subject-substance, to say that we could transfer between subjects is equivalent to saying that we could transfer between selves. But this is absurd. This shows that the Dualist use of body-transfer against PBA does not come down to the assertion that one can imagine consciousness moving from one material substance to another. The argument will be that one can imagine the *subject* to occupy different bodies; and this will show that having a particular body is not a part of our conception of the subject, of ourselves. The very fact that we can think of subject as occupying rather than being a body already shows that being a body is not a part of our concept of the subject.[412] But with regard to the soul - the mental substance-subject - it is absurd to think that one "occupies" it. To be a self is to be a mental substance-subject, regardless of what other properties one may have.[413] To conclude, (ST) is not analogous to the Dualist body-transfer-based arguments against PBA. And when we get clear about the Dualist conception of subjects and consciousness we see that (ST) is either absurd or directed against a straw-man.

10.6. Potential continuity of consciousness - a necessary and sufficient condition of identity?

I have proposed (C) as a formulation of a *necessary* condition of personal identity:

(C) For any persons *x*, *y* s.t. *x* exists at t_1 and *y* exists at t_2 (*x* = *y* **only if** there is a *possible* (or actual) series of conscious states (A,..., B) such that:

(i) *x* has A at t_1 & *y* has B at t_2 & the states between A and B occur at times between t_1 and t_2

(ii) every two successive states C and D in the series (A,..., B) *either* are (weakly) continuous *or* there is a possible series of (weakly) continuous experiences (C*,..., D*) s.t. C* is had by the person who has C and occurs at *t*(C) and D* is had by the person who has D and occurs at *t*(D).)

[412] Or, more cautiously, that being an *ordinary* human body is not a part of our concept of ourselves.

[413] It makes sense to ask whether the soul thus conceived is material or immaterial; a common question in antiquity.

Is potential continuity of consciousness also *sufficient* for identity? It is, on three assumptions. The first is that consciousness cannot be transfered from one self to another. This assumption was vindicated in the previous section. The other assumptions are that states of consciousness depend on subjects for their individuation and that they cannot be shared by subjects. I will not defend these two, sensible though they are. I am now only interested in the question whether potential continuity of consciousness could provide a full analysis of personal identity. Now, if the condition imposed by (C) is met for some X and Y, then, given our assumptions, it follows that X and Y are identical.[414] This gives us:

(C*) For any persons *x*, *y* s.t. *x* exists at t_1 and *y* exists at t_2 ($x = y$ **if and only if** there is a *possible* (or actual) series of conscious states (A,..., B) such that:
(i) *x* has A at t_1 & *y* has B at t_2 & the states between A and B occur at times between t_1 and t_2
(ii) every two successive states C and D in the series (A,..., B) *either* are (weakly) continuous *or* there is a possible series of (weakly) continuous experiences (C*,..., D*) s.t. C* is had by the person who has C and occurs at *t*(C) and D* is had by the person who has D and occurs at *t*(D).)

Does this mean that Primitivism is wrong? No, for the analysis of personal identity provided by (C*) is circular and not fully informative. First, in many (or perhaps all) cases when *x* and *y* are not linked by actual continuity of consciousness, we have to appeal to trans-world identity of persons to establish that the conditions imposed by (C*) are met. Secondly, I think we have no way of determining which states could be continuous (or not) other than by postulating that they belong to the same person (or not). Thus, seeing the truth of (C*) presupposes a prior independent understanding of persons and their persistence. Thirdly, the assumptions of ontological dependency and privacy of consciousness - required to make (C*)

[414] Take the simplest case: X is in state A and Y in state B. A and B occur within 5 minutes from each other and are not in fact continuous, but there could be a series of states (A, X, Y, B) where successive states are continuous. Then there will be a possible world where A and B are continuous. By our assumptions they belong to one and only one subject. Since B could not belong to a different subject (by the assumption of ontological dependency), B's subject in the actual world is X. Therefore X = Y. For more complex cases the travel into possible worlds will be iterated (and A and B may be supplanted by their counterparts), but the same reasoning will apply.

a criterion of identity - are anti-Reductionist in character and, again, require a prior understanding of persons. In short, (C*) offers nothing like a reductive analysis of personal identity.

I do not deny that a neo-Lockean or Humean could adopt a *similar* criterion for her purposes. Perhaps the idea of potential continuity of consciousness could be cashed out e.g. in nomological terms, so that circularity could be ridden off. I want to say only two things. First, if the person is to be viewed as "built up" from mental states, then I see little rationale for the assumption that they cannot be shared. So any (C*)-like criterion could at most provide an analysis of persistence, but not identity of persons. Secondly, *if* such project were successful, then I would not pick quarrel with it just because it is not Primitivist. The real questions would be whether it is consistent with the indivisibility and separateness of persons, whether it accounts for their agency and whether it can answer to the Dualist challenges.

Chapter 11. Conclusions

Persons are real, irreducible, ultimate components of reality. Their identity is a primitive fact that cannot be fully analyzed. This is the Simple View of persons. Its Consciousness-Based version is the best theory of personal identity we have. And arguably the best theory of persons consistent with it is Cartesian Dualism. In terms of specific theses, I argued for the following:

(SV1) There are persons.

(SV1*) There are persons persisting for a reasonably long time.

(SV2) It is impossible to provide a non-circular criterion of diachronic identity of persons i.e. to specify a relation R s.t. it would be true that for any persons A, B "A = B iff R(A, B)" is true and the holding of R of A and B could be described without presupposing that A and B are a single person.

(SUB) Persons are substances.

(SUBJ) Persons are subjects of mental states ("thinking things").

(CS) If a person's consciousness flows on continuously, the person persists.

(PCN) Persons essentially possess the potential for consciousness.

(PCCN) Persons essentially possess the potential for continuous consciousness.

(C) For any persons x, y s.t. x exists at t_1 and y exists at t_2 ($x = y$ **only if** there is a *possible* (or actual) series of conscious states (A,..., B) such that:

(i) x has A at t_1 & y has B at t_2 & the states between A and B occur at times between t_1 and t_2

(ii) every two successive states C and D in the series (A,..., B) *either* are (weakly) continuous *or* there is a possible series of (weakly) continuous experiences (C*,..., D*) s.t. C* is had by the person who has C and occurs at *t*(C) and D* is had by the person who has D and occurs at *t*(D).)

(I&S) Persons are necessarily indivisible and separate.

(DT) Identity of persons is necessarily always determinate.

(MS) Persons are mereologically simple.

(NC) Persons are not constituted by other things or properties of other things.

(D) The only essential property of persons is being “thinking things”.

(DC) Persons are either “constituted” by consciousness or are “constituted” by essentially consciousness-related mental properties (or powers or activities etc.)

My argument for these conclusion has been complex and multi-faceted. It is now the time to bring all the strands together.

Eliminativism about persons has been rejected on methodological grounds. We have self-experience and there are no good reasons to regard it as illusory; and we have deep pragmatic reasons to keep to Realism about persons until it is shown totally hopeless. But it is not hopeless - we can frame consistent theories of persons - and even cannot be shown to be hopeless, given the Agnostic option.

We can also be reasonably confident that persons persist for a long time. If persons exists, they are either what Reductionism says they are, or what the Simple View says they are. I have argued that Reductionism is false and the Simple View true. But then, at least on some SV theories, persons *can* persist for a long time: the unity of a person “constituted” by mental properties (cf. (DC)) is tight enough to be a truly substantial unity - even by demanding Strawsonian standards. In the absence of fundamental ontological obstacles, we have no reasons to doubt our own persistence. To the extent that any theory of persons generates reasons to doubt their persistence, this is a reason to reject it in favour of any acceptable theory which does not generate such doubts.

The primitiveness of identity, enshrined in (SV2), has been defended in several ways. First, I aimed for logically exhaustive classification of (mainstream) Reductionist theories of persons and argued that they all face seemingly insuperable obstacles in ontological and intuitive dimensions. Primitivism does not seem to suffer from comparably grave problems. Thus, on the whole, it is the best theory. The other strand explored the Reductionist's commitment to indeterminacy of personal identity. Indeterminacy of personal identity leads to indeterminacy of anticipation. The latter seems to be un-intelligible and leads to absurdities like excessive extendability of anticipation. Reductionist accounts seem unable to explain why indeterminacy of anticipation is so absurd.

The theses that persons are substances (SUB) which are subjects of mental states (SUBJ) were taken as part of the mainstream ontology. This is how the self is figured in ordinary self-experience. I argued that these ideas underlie the intuitions of the central importance of physical continuity given background Materialist assumption. I also defended the idea of substance-subject against the most popular challenges: experiential vacuity and subject-transfer. Since no compelling reasons were found to reject the theses or doubt the veracity of self-experience, we are entitled to hold to them.

Like my case for Primitivism, my argument for the Consciousness-Based view of personal identity (characterized by (CS), (PCN), (PCCN) and (C)) has been cumulative. In chapters devoted to "intuitions" I argued that we neither experience nor conceive nor deeply believe ourselves to be the kind of things that the classical (non-Consciousness-Based) Psychological Theory or Physically-Based Approach propose we are. Taking pro-(classical)-PT intuitions at face value would be inconsistent with our deep beliefs about the way people can change and with our actual treatment of amnesia and senility cases. Our intuitions which seem to favour PT can plausibly be explained as expressions of emotional identification and feelings about the loss of familiar life. At the bottom, they are not intuitions about numerical identity at all. Nor is our grasp of our identity tied to any bodily criterion since we can intelligibly frame radical Body-Transfer scenarios. In sum, intuitions about personal identity are sparse. Sparse, but not wholly absent. First, in entertaining any imaginary scenario, we can retain a firm grasp on our identity as long (and only as long) as we imagine that the end-states could be linked by continuity of consciousness to the start-point states - in the same sense of “could” that we think various states in our life could be so linked. Secondly, from the definitional assumption that persons essentially possess the potential to be conscious we can derive the conclusion that any experiences they have at different times could be linked by continuous consciousness. (PCCN) and its technical formulation (C) are supposed to correctly express these two ideas. (CS) is grounded in considerations of phenomenology of co-consciousness (since experiences are experienced-together, a new self cannot “jump in the middle” of the stream) and non-transferability of token-properties between substances. Given (CS) and further assumptions (ontological dependency and privacy of consciousness), potential continuity of consciousness may even be a necessary and sufficient condition of personal identity. Yet a criterion formulated in these terms will be circular and uninformative. Consciousness-Based Approach imposes constraints on the possible histories of persons; but they remain irreducible.

In my discussion of Reductionist metaphysical theories of persons, I aimed at judgments of comparative plausibility of theories or at showing the collapse of one theory into another, in order to select the stronger rivals to the Consciousness-Based Simple View. Let me briefly summarize the results. Within the PT camp, the Lewisian approach is least plausible, as it violates the “Only past and now” principle of identity of persons and undermines the agency of persons. Within 4D metaphysics the Transience View seems more plausible than the Lewisian approach.[415] The Diachronic Form Constitution View does not have any advantage over the Lewisian approach, since to deal with the *Argument from Initial Parts to Double Identity* it has to accept counterpart theory and face much the same problems as the Lewisian approach. The Synchronic Form Constitution View, on the other hand, is not able to deal with the problem of non-transferability of token-capacities between substances. It

[415] Sider (2001), 188-208 offers a different argumentation to the same effect.

must collapse into the Mixed Psycho-Physical Approach. But such account is less plausible than a Physically Based Approach or the Dualist Simple View. Within the PBA camp, the Serial Realization view is structurally identical to the Synchronic Form Constitution View PT, and untenable for the same reasons. Among the PBA theories which identify a person with a particular realizer of mentality, the biological approaches: Animalism and the Brain View are also untenable. Animalism relies on an implausible theory of composition. It also cannot satisfactorily deal with the challenges posed by problems related to thinking-subject minimalism. Finally, the Dicephalic Twins case is a strong counterexample to Animalism. The Brain View, on the other hand, seems unable to deal with the *Nanorobotic Commissurotomy* case; since in this case the very reasons we have for entertaining the Brain View force us to conclusions incompatible with this view. Two theories emerged as the strongest potential rivals of the Simple View. In PBA camp this is the One Realizer view. This theory is in a way closest to Cartesian SV: both theories identify persons with substances which essentially possess basic mental capacities. In the PT camp, the Parfitian Reductionism - most radically opposed to Cartesianism - turned out to be least objectionable from the ontological point of view. My charges against the One Realizer view do not concern its specific features; they are generally anti-Reductionist. First, there are well-known problems about the unity of the subject like Homunculism and the Problem of the Many. The new challenge comes from my argument that fission of consciousness is impossible. This seems to imply that the fission of the self is likewise impossible. Reductionism seems unable to account for either fact. Finally, indeterminacy of identity, unavoidable in Reductionism, creates a gap between identity and anticipation. Once this gap sets in, there is no arbitrary stopping point in extending anticipation; and this, I argued, undermines the appropriateness of anticipation. This suggests that there is no genuine survival if Reductionism is correct. This argument has a special force against Parfitian Reductionism which collapses into

Eliminativism. Considerations of fission and anticipation have thus served to establish Consciousness-Based SV characterized by (I&S) and (DT).

I argued that Cartesian Dualism characterized by (MS), (NC), (D) and (DC) is the best theory of the nature of persons consistent with SV. Besides the traditional arguments concerning the unity of the subject, it can be supported by the new argument based on consideration of Fission. It can answer to challenges to Persistence Views posed by Strawson; it is free from the problems troubling other versions of SV; and it can face up to challenges usually issued.

I have noted in the Introduction that dissatisfaction with the thought-experiment-based method in personal identity inquiry has led to three different reactions. One is to approach personal identity predominantly through investigation of first-person perspective, consciousness and self-experience. The second is to make the account of personal identity depend on solutions in general metaphysics. The third is to take our everyday concerns, practices and ethics as the guide. I tried to take account of each of these perspectives. In accordance with the methodological principles proposed in ch. 2, I took the approach to metaphysics which treats phenomenology seriously and is sensitive to practical implications. Nevertheless, the metaphysical issues were at the centre of my considerations; and the ethical aspects of personal identity were not systematically investigated. I want now to say a few words why this procedure seems to me justified (not only by the limits of size of my work) and why it does not make my findings overly vulnerable. First, I have some sympathy with what I call Methodological Pluralism with regard to personal identity.[416] We could say that the whole problem of "personal identity" runs just too many issues together. There are many different questions and concerns about persons; and there are many non-coinciding notions of persons engendered by our different concerns. So perhaps in all dubious cases the only

[416] I take Marya Schechtman's works as exemplifying the spirit of Methodological Pluralism.

sensible answer is "Well, in one sense this is the same person, in another sense it is a different person". We can talk of persons$_1$, persons$_2$, persons$_3$ and so on. I think that Methodological Pluralism correctly diagnoses an important problem of contemporary philosophy of personal identity. We could make more progress if we abandoned the over-ambitious project of building one comprehensive theory of persons into which all intuitions must fit or be discounted. Nevertheless, Methodological Pluralism is not fully satisfying. What I take to be its major flaw is that it leaves us with no answer to the question: "But what *I* really am?". I think a viable alternative to Pluralism lies in the Stratification approach. The idea is to organize various conceptions of persons, and various senses of "is" as used in identificatory statements, into a multi-level structure. The use of distinctions between persons and *personae* or between substance- and phase-sortals is an example of such stratification. Of course, what may be "deepest" from the viewpoint of ontology may not be "deepest" from the viewpoint of ethics and vice-versa. But we can plausibly give priority to self-experience and to acts of self-identification. We can stratify them and then stratify the various senses of "self" and "identity" accordingly. Now, the theory of the self presented in this work is supposed to show what kind of thing is the self which is the object of the basic form of self-experience: one which provides the core of our concept (or the family of concepts) "person"; and which seems central to our acts of anticipation.[417] In this sense it is meant to show what we *fundamentally* are. Other conceptions and intuitions may be accommodated as relevant and correct on different levels. Development of Stratification approach would require systematic assessment of experiences, beliefs and attitudes and asking questions like: what are they really about? what are their sources? what is their relevance? how are they related to one another? The project of logically organizing (or perhaps re-organizing) our divergent intuitions and conceptions is ambitious but feasible. Thus I feel justified in hoping that my

[417] Cf. ch. 1, 8-12.

findings are not just a vulnerable fragmentary view but a completed stage of a definite research project. Of course, the proof of the pudding is in the eating. But I think we can get some foretaste of possible developments. Inquiry into anticipation provides a natural starting point for addressing ethical issues in "what matters in survival" and in our attitudes towards death. I argued that the concept of anticipation is reason-responsive, and so sensitive to metaphysical argumentation. This conclusion can be extended to other ethically important concepts. I thus come squarely on Parfit's side in the *meta*-ethical debate around "what matters" concerning the link - or insulation - of practice and metaphysics. On the other hand, further inquiry into the way our attitudes presuppose particular experiences and conceptualizations of our existence in time can lead to anti-Parfitian conclusions with regard to bias for the future and reasonableness of the fear of death.[418] The fact that Consciousness-Based SV is not only consistent with our important attitudes but also fruitful in their explanation further enhances its plausibility. However, the nature described by this view, while fundamental, is also thoroughly generic. And while this nature and its survival are relevant for some basic attitudes, it is surely not all that we care about. Our "concrete existence" - ethically significant properties which from the viewpoint of ontology may be merely contingent - is what distinguishes us as individuals and "makes us who we are". The interplay between these levels of identity - as well as confusions arising from the lack of distinction between them - should be analyzed by future Simple View theories. In this way, the Simple View will be able to offer a package-deal of solutions in phenomenology, ontology and ethics. I believe this is the way to go.

[418] Using the idea of dynamic continuity of consciousness one can show that concern for the present is not separable from the concern for the near future; and that in turn from the concern for the future in general. This, contrary to Parfit's view, makes bias for the future justified. This disarms the Symmetry Argument against the rationality of the fear of death.

Bibliography

Annas, J. (1993), *The Morality of Happiness*, Oxford University Press, New York.

Augustine of Hippo, *De Trinitate*, translated by A. W. Haddan, available at: http://www.newadvent.org/fathers/1301.htm

Baker, L. R. (2000), *Persons and Bodies: A Constitution View*, Cambridge University Press, Cambridge.

Baker, L. R. (2007), "Persons and the Natural Order", in van Inwagen and Zimmerman (eds) (2007), 261-278.

Belshaw, Ch. (1993), "Assymetry and Non-Existence", Philosophical Studies, 70, 103-116.

Blackburn, S. (1997), "Has Kant Refuted Parfit?", in Dancy (ed.) (1997), 180-201.

Boethius, *De persona et duabus naturis,* available at: http://www.ihaystack.com/authors/b/anicius_manlius_severinus_boethius/

Braude, S. E. (1991), "Multiple Personality and the Structure of Self", in Kolak and Martin (1991), 134-143.

Butler, J. (1736/1975), "Of Personal Identity", in Perry (ed.) (1975), 99-106.

Cassam, Q. (1997), *Self and World,* Oxford University Press, New York.

Chisholm, R. (1989), *On Metaphysics,* University of Minnesota Press, Minneapolis.

Clark, A. and Chalmers, D. J. (1998), "The Extended Mind", Analysis 58, 7-19.

Dainton, B. (2000), *Stream of Consciousness*, Routledge, London.

Dainton, B. (2005), "The Self and the Phenomenal", in Strawson (2005), 1-25.

Dainton, B. (2008), *The Phenomenal Self*, Oxford University Press, Oxford.

Dancy, J. (ed.) (1997), *Reading Parfit*, Blackwell, Oxford.

de Sousa, R. (1987), *The Rationality of Emotions*, MIT Press, Cambridge, Mass.

Dennett, D. C. (1991), *Consciousness Explained*, Little, Brown and Company, Boston.

Emilsson, E. (1988), *Plotinus on Sense-Perception: A Philosophical Study*, Cambridge University Press, Cambridge.

Evans, C. O. (1970), *The Subject of Consciousness*, George Allen & Unwin Ltd., London.

Forman, R. (1999), "What Does Mystycism Have to Teach Us About Consciousness", in Gallagher and Shear (eds) (1999), 361-378.

Foster, J. (1991), *The Immaterial Self: A Defence of the Cartesian Dualist Conception of the Mind*, Routledge, London.

Frankfurt, H. (1999), *Necessity, Volition, and Love,* Cambridge University Press, Cambridge.

Gallagher, S. and Shear J. (eds) (1999), *Models of the Self*, Imprint Academic, Thorverton.

Garrett, B. (1998), *Personal Identity and Self-Consciousness*, Routledge, London.

Gendler, T. Sz. (1999), "Exceptional Persons: On the Limits of Imaginary Cases", in Gallagher and Shear (eds) (1999), 447-466.

Giles, J. (1997), *No Self to Be Found: The Search for Personal Identity,* University Press of America, Lanham MD.

Glover, J. (1988), *I: The Philosophy and Psychology of Personal Identity,* The Penguin Press, London.

Haksar, V. (1991), *Indivisible Selves and Moral Practice*, Edinburgh University Press, Edinburgh.

Haldane, J. (1999), "A Return to Form in the Philosopy of Mind", in Oderberg (ed.) (1999), 40-64.

Hart, W. D. (1988), *The Engines of the Soul*, Cambridge University Press, Cambridge.

Hart, W. D. and Yagisawa T. (2007), "Ghosts Are Chilly", in van Inwagen and Zimmerman (eds) (2007), 166-168.

Hawthorne, J. (2007). “Cartesian Dualism”, in van Inwagen and Zimmerman (eds) (2007), 87-98.

Hazzlit, W (1805), *An Essay on the Principles of Human Action*, J. Johnson, London. Available at: http://books.google.pl/

Hershenov, D. (2006), “Shoemaker”s Problem of Too Many Thinkers”, Proceedings of the American Catholic Philosophical Association, 80, p 225-236.

Hilgard, E. (1991), “Dissosiative Phenomena and the Hidden Observer”, in Kolak and Martin (eds) (1991), 89-114.

Hoffman, J. and Rosenkrantz, G. S. (1997), *Substance: its nature and existence*, Routledge, London.

Hudson, H. (2007), “I Am Not an Animal!”, in van Inwagen and Zimmerman (eds) (2007), 216-234.

Humphrey, N. and Dennett, D. C. (1989), “Speaking for Ourselves: An Assessment of Multiple Personality Disorder”, Raritan 9, 68-98. Reprinted in Kolak and Martin (eds) (1991), 144-162.

Johnston, M. (1987), “Human Beings”, The Journal of Philosophy 84, 59-83.

Johnston, M. (1989), “Fission and the Facts”, Philosophical Perspectives, 369-397.

Johnston, M. (1992), “Constitution Is Not Identity”, Mind 101, 89-105.

Johnston, M. (1997), “Human Concerns without Superlative Selves”, in Dancy (ed.) (1997), 149-179.

Kolak, D. and Martin, R. (eds) (1991), *Self and Identity: Contemporary Philosophical Issues*, Macmillan Publishing Company, New York.

Korsgaard Ch. (1989), “Personal Identity and the Unity of Agency: A Kantian Response to Parfit”, Philosophy and Public Affairs 18, 103-31. Reprinted in Martin and Barresi (eds) (2003), 168-183.

Le Poidevin, R., Simons, P., McGonigal, A., Cameron, R. (eds) (2009), *The Routledge Companion to Metaphysics*, Routledge, London.

Lewis, D. (1971), “Counterparts of Persons and Their Bodies”, Journal of Philosophy 68, 203-11. Reprinted in: Lewis (1983*a*), v. I, 47-54.

Lewis, D. (1976*a*), “Survival and Identity”, in Rorty (1976), 17-40. Reprinted in: Lewis (1983*a*), *v. I*, 55-72.

Lewis, D. (1976*b*) “The Paradoxes of Time Travel”, American Philosophical Quarterly 13, 145-152. Reprinted in: Lewis (1983*a*), v. II, 67-80.

Lewis, D. (1983*a*), *Philosophical Papers*, Oxford University Press, Oxford and New York.

Lewis, D. (1983*b*), “Postscripts to Survival and Identity”, in Lewis (1983*a*), v. I, 73-77.

Lizza, J. P. (2006), *Persons, Humanity and the Definition of Death*, The John Hopkins University Press, Baltimore.

Locke, J. (1690/1975), *An Essay Concerning Human Understanding*, Oxford University Press, Oxford.

Loux, M. J. and Zimmerman, D. W. (eds) (2003), *The Oxford Handbook of Metaphysics*, Oxford University Press, Oxford.

Lowe, E. J. (1996), *Subjects of Experience*, Cambridge: Cambridge University Press.

Lowe, E. J. (1998), *The Possibility of Metaphysics. Substance, Identity, and Time*, Clarendon Press, Oxford.

Lowe, E. J. (2003), “Individuation”, in: M.J. Loux and D. W. Zimmerman (eds) (2003), 75-95.

MacIntyre, A. (1985), *After Virtue: A Study in Moral Theory*, Duckworth, London.

Mangan, B. (2001) “Sensation’s Ghost: The Non-Sensory ‘Fringe’ of Consciousness”, Psyche 7, available at:

http://journalpsyche.org/ojs-2.2/index.php/psyche/article/view/2592

McMahan, J. (2002), *The Ethics of Killing. Problems at the Margins of Life*, Oxford University Press, Oxford.

Maddell, G. (1981), *The Identity of the Self*, Edinburgh University Press, Edinburgh.

Martin, R. (1998), *Self Concern*, Cambridge University Press, Cambridge.

Martin, R. and Barrese, J. (eds) (2003), *Personal Identity*, Blakwell, Oxford.

Nagel, T. (1971) "Brain Bisection and the Unity of Consciousness", Synthese 22, 396-413. Reprinted in Nagel (1979), 147-164.

Nagel, T. (1979), *Mortal Questions*, Cambridge University Press, Cambridge.

Nagel, T. (1986), *The View from Nowhere*, Oxford University Press, Oxford.

Neisser, U. (1988), "Five Kinds of Self-Knowledge", Philosophical Psychology 1988, 35-29. Reprinted in Kolak and Martin (eds) (1991), 386-406.

Noonan, H. (2003), *Personal Identity*, 2nd edition, Routledge, London.

Nozick, R. (1981), *Philosophical Explanations*, Harvard University Press, Cambridge Mass.

Oderberg, D. (1993), *The Metaphysics of Identity over Time*, St. Martin"s Press, New York.

Oderberg, D. (ed.) (1999), *Form and Matter*: *Themes in Contemporary Metaphysics*, Blackwell, Oxford.

Olson, E. T. (1997), *The Human Animal: Personal Identity without Psychology*, Oxford University Press, Oxford

Olson, E. T. (2007), *What Are We?*, Oxford University Press, Oxford.

Parfit, D. (1984), *Reasons and Persons*, Oxford University Press, Oxford.

Parfit, D. (1995), "The unimportance of identity", in Harris, H. (ed.), *Identity: essays based on Herbert Spencer lectures given in the University of Oxford*, Clarendon Press, Oxford, 13-45. Reprinted in: Martin and Barresi (2003), 292-317.

Parfit, D. (1999), "Experiences, subjects and conceptual schemes", Philosophical Topics 26, 217-270.

Perry (1975), "Personal Identity, Memory and the Problem of Circularity", in Perry (ed.) (1975), 135-155. Reprinted in Perry (2002), 84-99.

Perry, J. (ed.) (1975), *Personal Identity*, University of California Press, Berkeley.

Perry, J. (1976), "The Importance of Being Identical", in Rorty (ed.) (1976), 67-90.

Perry, J. (1978), *A Dialogue on Personal Identity and Immortality*, Hackett Publishing Company Inc., Indianapolis.

Perry, J., (2002), *Identity, Personal Identity and the Self*, Hackett Publishing Company Inc., Indianapolis.

Persson, I. (2005), "Self-Doubt: Why We are not Identical to Things of Any Kind", in Strawson (2005), 26-44.

Plantinga, A. (2007), "Materialism and Christian Belief", in van Inwagen and Zimmerman (eds) (2007), 99-141.

Puccetti, R. (1973), "Brain Bisection and Personal Identity", British Journal for the Philosophy of Science 24, 339-55.

Puccetti, R. (1989), "Two Brains, Two Minds? Wigan's Theory of Mental Duality", The British Journal for the Philosophy of Science 1989, 137-144. Reprinted in Kolak and Martin (eds) (1991), 68-74.

Radden, J. (1999), "Pathologically Divided Minds, Synchronic Unity and Models of Self", in Gallagher and Shear (eds) (1999), 343-358.

Railton, P. (1989), "Naturalism and Prescriptivity", *Social Philosophy and Policy* 7, 151-174.

Rist, J. M. (2002), *Real Ethics: Reconsidering the Foundations of Morality*, Cambridge University Press, Cambridge.

Robinson, H. (2007), "The Self and Time", in van Inwagen and Zimmerman (eds) (2007), 55-

83.

Robinson, H. (2009), “Supervenience, Reductionism and Emergence”, in Le Poidevin et al. (eds) (2009), 527-536.

Rorty, A. O. (ed.) (1976), *The Identities of Persons*, University of California Press, Berkeley.

Rovane, C. (1998), *The Bounds of Agency,* Princeton University Press, Princeton.

Schechtman, M. (1990), “Personhood and Personal Identity”, Journal of Philosophy 87, 71-92.

Schechtman, M. (1996), *The Constitution of Selves*, Cornell University Press, Ithaca and London.

Schechtman, M. (1997), “The Brain/Body Problem”, Philosophical Psychology 10, 149-164.

Schechtman, M. (2001), “Empathic Access: The Missing Ingredient in Personal Identity”, Philosophical Explorations 4, 95-111. Reprinted in Martin and Barresi (eds) (2003), 238-259.

Seager, W. (1999), *Theories of Consciousness. An Introduction and Assessment*, Routledge, London.

Shear, J. (1999), “Experiential Clarification of the Problem of Self”, in Gallagher and Shear (eds) (1999), 407-420.

Shoemaker, S. (1963), *Self-Knowledge and Self-Identity*, Cornell University Press, Ithaca and London.

Shoemaker S. (1984), “Personal Identity: A Materialist Account”, in Shoemaker and Swinburne (1984), 67-132.

Shoemaker, S. (1997), “Parfit on Identity”, in Dancy (ed.) (1997), 135-148.

Shoemaker, S. (2004), “Functionalism and Personal Identity - A Reply”, Noûs 38, 525-533.

Shoemaker, S. (2008), “Persons, Animals, and Identity”, Synthese 162, 313-324.

Shoemaker, S. and Swinburne, R. (1984), *Personal Identity*, Basil Blacwell, Oxford.

Sider, T. (2001), *Four-Dimesionalism. An Ontology of Persistence and Time*, Clarendon Press, Oxford.

Slors, M. (2001), *The Diachronic Mind*, Kluwer Academic Publishers, Dordrecht.

Sorabji, R. (2006), *Self. Ancient and Modern Insights about Individuality,Life and Death*, Clarendon Press, Oxford.

Sperry, R. W. (1968) “Hemisphere Deconnection and Unity in Conscious Awareness”, American Psychologist 23,723-733. Reprinted in Kolak and Martin (eds) (1991), 55-68.

Stone, J. (1988), “Parfit and the Buddha: Why There Are No People”, Philosophy and Phenomenological Research 48, 519-532.

Stone, J. (2005), “Why There Still Are No People”, Philosophy and Phenomenological Research 70, 174-191.

Strawson, G. (1997), “The Self”, Journal of Consciousness Studies, 1997, 405-428. Reprinted in Martin and Barresi (eds) (2003), 335-377.

Strawson, G. (1999), “The Self and SESMETS”, Journal of Consciousness Studies 6, 99-135.

Strawson, G. (ed.) (2005*a*), *The Self?*, Blackwell, Oxford.

Strawson, G. (2005*b*), “Against Narrativity”, in Strawson (2005*a*), 63-86.

Strawson, G. (2008*a*), *Selves: An Essay in Revisionary Metaphysics*, manuscript.

Strawson, G. (2008*b*), “The Identity of the Categorical and the Dispositional”, Analysis 68, 271-282.

Swinburne, R. (2007), “From Mental/Physical Identity to Substance Dualism”, in van Inwagen and Zimmerman (eds) (2007), 142-165.

Taylor, C. (1989), *Sources of the Self*, Harvard University Press, Cambridge, Mass.

Thomson, J. J. (1997), “People and their Bodies”, in Dancy (ed.) (1997), 202-229.

Tye, M. (2003), *Consciousness and Persons. Unity and Identity*, The MIT Press, Cambridge, Mass.

Unger, P. (1979*a*), “There Are No Ordinary Things”, Synthese 41, 117-154. Reprinted in Unger (2006), 3-35.

Unger, P. (1979*b*), “I Do Not Exist”, in G. F. Macdonald (ed.) *Perception and Identity*, Macmillan, London, 235-251. Reprinted in Unger (2006), 36-52.

Unger P. (1979*c*), “Why There Are No People”, Midwest Studies in Philosophy 4, 177-222. Reprinted in Unger (2006), 53-109.

Unger P. (1980), “The Problem of the Many”, in Midwest Studies in Philosophy 5, 411-467. Reprinted in Unger (2006), 113-182.

Unger P. (2000), “The Survival of the Sentient”, Philosophical Perspectives 14, 325-348. Reprinted in Unger (2006), 265-292.

Unger P. (2004), “The Mental Problems of the Many”, Oxford Studies in Metaphysics 1, 195-222. Reprinted in Unger (2006), 183-208.

Unger P. (1990), *Identity, Consciousness and Value*, Oxford University Press.

Unger, P. (2005), *All the Power in the World*, Oxford University Press, Oxford.

Unger, P. (2006), *Philosophical Papers*, v. II, Oxford University Press, Oxford.

van Frassen, B. (2005), “Transcendence of the Ego (The Non-Existent Knight)”, in Strawson (2005*a*), 87-110.

van Inwagen, P.(1990), *Material Beings,* Cornell University Press, Ithaca N.Y.

van Inwagen, P. and Zimmerman, D. (eds) (2007), *Persons. Human and Divine*, Clarendon Press, Oxford.

Velleman, J. D. (1996), “Self to Self”, *The Philosophical Review* 105, 39–76. Reprinted in Velleman (2006), 170-202.

Velleman, J. D. (2001), “Identification and Identity”, in Buss, S. and Overton, L., *The Contours of Agency: Essay on Themes from Harry Frankfurt*, MIT Press, Cambridge, Mass., 91–123. Reprinted in Velleman (2006), 330-360.

Velleman, J. D. (2005), “The Self as Narrator”, in Anderson, J. and Christman, J. (eds), *Autonomy and the Challenges to Liberalism: New Essays*, Cambridge University Press, Cambridge, 56-76. Reprinted in Velleman (2006), 203-223.

Velleman, J. D. (2006), *Self to Self*, Cambridge University Press, Cambridge.

Whiting, J. (1986), “Friends and Future Selves”, The Philosophical Review 95, 547-580.

Wiggins, D. (1980), *Sameness and Substance,* Blackwell, Oxford.

Wiggins, D. (2001), *Sameness and Substance Renewed,* Cambridge, Cambridge University Press.

Wilkes, K. V. (1988), *Real People: Personal Identity without Thought-Experiments*, Oxford University Press, New York.

Williams, B. A. O. (1966), “Imagination and the Self”, Proceedings of the British Academy 70: 105–24. Reprinted in Williams (1973), 26-45.

Williams, B. A. O. (1970), “The Self and the Future”, Philosophical Review 79, 161-180. Reprinted in Williams (1973), 46-63.

Williams, B. A. O. (1973), *Problems of the Self: Philosophical Papers 1956-1972,* Cambridge University Press.

Wollheim, R. (1984), *The Thread of Life,* Cambridge University Press, Cambridge.

Zahavi, D. (2005), *Subjectivity and Selfhood*, MIT Press, Cambridge, Mass.