PERSONAL IDENTITY AND SELF-CONSCIOUSNESS

INTERNATIONAL LIBRARY OF PHILOSOPHY

BRIAN GARRETT

ROUTLEDGE

PERSONAL IDENTITY AND SELF-CONSCIOUSNESS

> In eight, clear, careful and well designed chapters Brian Garrett analyses the central issues involved in the problem of personal identity. I found the central chapters, in which Garrett brings…his clarity of thought and a mastery of general logic, particularly penetrating and helpful. Garrett has written an intelligent, thoughtful and thought-provoking book…which significantly moves the debate along, and which should be read by students and by all philosophers interested in personal identity.
>
> Paul F.Snowdon
> Exeter College, Oxford

In *Personal Identity and Self-Consciousness,* Brian Garrett presents an original and comprehensive theory of persons: their nature, their values, and their self-consciousness. He begins by proposing a new theory of personal identity over time. Next, he defends the importance of personal identity against recent sceptical attack. Finally, Garrett explores the nature of self-consciousness by examining the elusive pronoun 'I' and the various grounds of our 'I' judgements.

Brian Garrett places recent discussions of personal identity in a broader context, and links issues in personal identity with other central issues in philosophy, notably the problem of self-consciousness and questions in ethics. Garrett manages to tackle a technical and complex discussion with jargon-free and elegant language.

This is the first book of its kind to bring together the many different issues that surround the discussion of personal identity. Brian Garrett makes an important and original contribution to the study of the philosophy of personal identity, the philosophy of mind, and to epistemology.

Brian Garrett is Lecturer in Philosophy at the Australian National University.

INTERNATIONAL LIBRARY OF PHILOSOPHY

Edited by Tim Crane and Jonathan Wolff

University College London

The history of the International Library of Philosophy can be traced back to the 1920s, when C.K.Ogden launched the series with G.E. Moore's *Philosophical Papers* and soon after published Ludwig Wittgenstein's *Tractatus Logico-Philosophicus.* After this auspicious start, the series has been edited by A.J.Ayer, Bernard Williams and Ted Honderich. Now under the joint editorship of Tim Crane and Jonathan Wolff, the I.L.P. will continue to publish the best of original research in philosophy.

Other titles in the I.L.P. include:

THE SCEPTICAL CHALLENGE
Ruth Weintraub

THE IMMATERIAL SELF
John Forster

DISPOSITIONS: A DEBATE
D.M.Armstrong, C.B.Martin and U.T.Place

PSYCHOLOGY FROM AN EMPIRICAL STANDPOINT
Franz Brentano, Auto Rancurello, D.B.Terrel, Londa L.McAllister and Peter Simons

G.E.MOORE: SELECTED WRITINGS
Edited by Thomas Baldwin

THE FACTS OF CAUSATION
D.H.Mellor

PERSONAL IDENTITY AND SELF-CONSCIOUSNESS

Brian Garrett

London and New York

First published 1998 by Routledge
11 New Fetter Lane, London EC4P 4EE

This edition published in the Taylor & Francis e-Library, 2002.

Simultaneously published in the USA and Canada
by Routledge
29 West 35th Street, New York, NY 10001

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalogue record for this book has been requested

ISBN 0-415-16573-3 (Print Edition)
ISBN 0-203-01566-5 Master e-book ISBN
ISBN 0-203-22346-2 (Glassbook Format)

For RG and SG

CONTENTS

PREFACE

This book is aimed mainly at professional philosophers. It is intended as a contribution to the research industry which sustains the topic of personal identity. However, I hope the book will also introduce this topic to a wider graduate audience. Accordingly, I have tried to make the book as accessible as possible, and I have attempted to explain any terms of jargon that might have crept in.

Many people have given me comments over the past few years which have helped improve earlier drafts. In particular, thanks to David Braddon-Mitchell, Tim Crane, Garrett Cullity, Richard Holton, Frank Jackson, Kevin Mulligan, Michael Smith, Natalie Stoljar, Philip Pettit, Paul Thorn, Timothy Williamson, Crispin Wright, and an anonymous referee for Routledge. Let me also acknowledge long-standing debts to Derek Parfit and Paul Snowdon, whose clarity of thought provided a model of philosophical theorising.

Finally, I am grateful to the Australian Research Council for awarding me a Queen Elizabeth II Research Fellowship, which enabled this book to be completed in the most congenial of circumstances.

1

THE PROBLEM AND ITS PLACE IN PHILOSOPHY

The problem of personal identity

This book is intended as an overview of issues in the philosophy of persons and personal identity. In the first five chapters, we will be concerned with questions dealing with the nature (or metaphysics) of persons and personal identity. In the sixth chapter, we address the question of whether the value or importance that we attach to persons and personal identity is justified. In the final two chapters, we shall assess the extent to which a proper understanding of the semantic (that is, meaning-related) and epistemic (that is, knowledge-related) features of first-person judgements—judgements of the form 'I am F' —can shed light on the concept of self-consciousness. This concept is a key constituent of our concept of a person.

The concerns of this book are strictly philosophical. We are not concerned with issues of 'personal identity' as this phrase is colloquially understood (in terms of a person's self-image or fundamental values and beliefs). Rather, we are concerned with personal identity in an abstract way, where what matters is not the particular characteristics that distinguish us, but those characteristics we all (or most of us) have in common.

Moreover, the word 'identity' should be taken to connote strict numerical identity, not mere qualitative identity (that is, exact similarity). The distinction between numerical and qualitative identity is crucial in what follows. We are not concerned with identity in the sense of qualitative identity or exact similarity, as when we talk of identical twins or identical billiard balls. Rather, we are concerned with identity in the sense of 'numerical identity'. In this sense, twins are not the same, they are two different people. Throughout the book, 'identity' should always be understood in this second, numerical, sense, unless otherwise indicated.

One might wonder why there should be a problem about personal identity. Is the relation of personal identity not simply an instance of the relation of identity, and so defined by the formal properties of reflexivity ($(\forall x)\ (x=x)$) and congruence ($(\forall x)(\forall y)(x=y \rightarrow$ x and y share all their properties))? A relation R is the relation of identity just if R is reflexive and congruent. What more needs to be said?

The answer, fortunately, is that a lot more needs to be said. The formal properties of identity tell us absolutely nothing about why we are right to make many of the judgements of personal identity that we do make, both in ordinary cases and in more outlandish fantasy cases.

For example, suppose we rightly judge Moriarty to be the murderer. We can ask why this is true. Someone might respond: Moriarty is the murderer because Moriarty stands to the murderer in the relation of identity, defined as above. However, it would be a fallacy to think that the availability of such an unilluminating response implies that there are no non-trivial necessary and/or sufficient conditions ('criteria') for the truth of judgements of personal identity.[1] Such judgements are subject to material conditions of correctness, and the formal properties of identity can tell us nothing about those conditions.

We can think of the matter as follows. The sentence 'A is the same person as B' is equivalent to the sentence 'A is B, and A and B are persons'. The truth of such sentences is subject to two sets of constraints: the formal constraints of identity, and constraints that follow from what it is to be a person. The task of the first five chapters of this book is to elucidate these latter constraints. The methodology employed is unrepentantly a priori.

In the chapters that follow we will be concerned to answer the following questions:

- What is a person? Spirit, animal, body, brain?
- What is it for the same person to persist through time? Can I survive the destruction of my body and brain? Can I survive the extirpation of my mental life?
- What does the possibility of fission show about the nature and importance of personal identity? The example of fission which will concern us is an imaginary case in which surgeons bisect my brain and transplant each hemisphere into its own body, resulting in the creation of two people, both of whom are psychologically very like me.
- Is personal identity an all-or-nothing matter? Or can it sometimes be *vague* or *indeterminate* whether a person at one time is the same

as some person at a later time? Can it be a vague matter whether I will exist tomorrow?

- Is the special importance we each assign to our own futures irrational? That is, is personal identity really the justifier of the 'special' concern which we have for ourselves in the future, or is the justifier some other relation which accompanies personal identity in the normal case?

  For example, Derek Parfit thinks that the relation of *psychological continuity* is the justifier of the 'special' concern we have for ourselves in the future.[2] This relation is composed of a number of chains, or strands, of interlocking direct psychological connections, such as those which hold between an experience-memory and the experience-remembered, or between an intention and the action which manifests it, or the chain consisting of the retention of beliefs, desires, memories, character, etc., over time. The relation of psychological continuity is not the same as the relation of personal identity, as the possibility of the fission and fusion of persons makes clear.
- Is the first-person singular ('I') a device of reference to an object, or does it have a different function?
- What is the link between the reference-fixing rule for 'I' (viz., 'A token of "I" refers to whoever produced it') and the fact that 'I' -judgements are expressions of self-consciousness?
- What is shown about our concept of self-consciousness by the fact that a certain (fundamental) class of 'I' -judgements are said and thought directly, and not said or thought on the basis of inference or observation?

What is a person?

In asking a question of the form 'What is an F?', we are asking a question in ontology. It is a question about the nature of Fs, not a question about the meaning of 'F' or the concept of F-ness. However, a question of the form 'What is an F?' is often ambiguous. It can mean: 'What conditions does something have to satisfy in order to be an F?' (call this the satisfaction question). Alternatively, it can mean 'Of what kind of stuff (animal, vegetable, mineral, etc.) are Fs composed?' (call this the nature question).

Thus, the question 'What is a table?' can be disambiguated in either of these two ways. In the first way, taken as a satisfaction question, the appropriate answer would be 'A table is an object, typically man-made, and typically having four legs, which is used for putting coffee cups on,

working on, eating off, etc.'. In the second way, taken as a nature question, the best answer would be 'A table is an artefact which is not made out of any one kind of stuff—tables can be made out of no end of material (wood, aluminium, plastic, gold, ice, etc.)'.

Notice that in the case of 'What is a table?' the two answers are independent of each other. In particular, the answer to the satisfaction question does not determine any particular answer to the nature question. The knowledge that something is a table (in the satisfaction sense) does not allow us to form any expectations about its composition. This is not so, however, in the case of a question such as, for example, 'What is a tree?'. The answer to this question, understood as a satisfaction question, cannot be separated from the answer to the question, understood as a nature question. Trees are necessarily made of wood, and a full answer to the satisfaction question will have to make reference to this fact.

Consider now the question 'What is a person?'. This can be understood either as a satisfaction question or as a nature question. In this and subsequent chapters, we shall be concerned to answer both the satisfaction and nature questions, and to assess the relation between them. We should then be in a position to determine whether an answer to the question 'What is a person?', understood as a satisfaction question, is independent of the answer to that question, understood as a nature question.

Some philosophers believe that the best answer to the satisfaction question is not independent of the best answer to the nature question. According to the animalist of Chapter 2, for example, the best answer to the satisfaction question will have to refer to our nature as human beings. However, as we shall see, there are good reasons to doubt the truth of animalism, and those reasons also suggest that the two answers to 'What is a person?' are largely independent.

The satisfaction question

If we temporarily assume that the answer to 'What is a person?', understood as a satisfaction question, need not make reference to the fact that we are human beings, how should that answer best proceed? That is, what conditions does something have to satisfy in order to qualify as (to be) a person? We may take it that, whatever else must be true, a person is a *mental* being. A person possesses a mind. The mind does not have always to be conscious—a sleeping or comatose person is still a person—but there must at all times be the capacity for mentality.

However, not just any mental being is a person. My cat is a mental being—he can feel pain or hunger, for example—yet he is not a person. So a person is (at least) a being that possesses a particular sort of mind. A person does not just have sensations of pain and pleasure, but also world-directed mental states like the belief that it will rain tomorrow, or the desire that a certain political party get elected.

Indeed, persons possess a range of particularly sophisticated mental states, including—most crucially—self-reflective mental states. I am capable of having not just the belief that it is raining or that Clinton is President, but also beliefs about myself. These are not just beliefs about someone who happens to be me, as when I think 'the person born on 16 May 1961 is Scottish', referring to myself but forgetting that *I* am that person. They are fully self-conscious beliefs about myself, the sort of beliefs I have when I say 'I remember that it snowed last Christmas' or 'I intend to holiday in Thailand when this term is over'.

Persons are self-conscious mental beings. Self-consciousness is what distinguishes us from other mental beings, such as cats and rabbits, and from everything else. This is confirmed when we reflect on how much of what matters in our mental life and social interactions presupposes the self-consciousness of ourselves and others. For example, we value our own autobiographical memories and our own future plans. This would be impossible if we were not self-conscious. Or again, we praise and blame other people because we take them to be self-consciously aware of their own responsibilities. If other people were not self-conscious, the rationale for most of our attitudes to others would simply be lost.

Hence, the best short answer to the satisfaction question is that persons are self-conscious mental beings. This common-sense answer to the satisfaction question was expounded clearly by Locke in the seventeenth century. Locke wrote that a person is: 'a thinking, intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places…'.[3] In this definition, Locke specifies some of the elements that comprise our concept of self-consciousness. In particular, he cites thinking, intelligence, reason, reflection, and the ability to engage in tensed first-person judgements ('I was F', 'I will be F there', etc.).

Locke held these features to be *constitutive* of our concept of a person. That is, a creature's possession of these features is not merely good evidence that the creature is a person, it is what it *is* to be a person. In addition, of course, there are ethical and social dimensions to

persons. Persons are free, rational, moral agents, living lives that depend for their richness on interaction with others. However, as suggested above, those dimensions are possible only because persons are rational and self-conscious. Self-consciousness is thus the core of personhood.

The nature question

The common-sense, Lockean answer to the satisfaction question does not help us to settle the nature question. As should be evident, Locke's definition does not presuppose any particular view about what kind of substance, if any, persons are.

How then should we attempt to answer the nature question? What is the range of available answers? Traditional answers to the nature question fall into two basic categories, immaterialist and materialist.

The immaterialist answer

According to this answer (associated with Plato, Descartes, and much, though not all, of the Christian tradition), a person possesses an immaterial soul, an entity with no extension in space.[4] The soul, in some way, interacts with the body. The soul is the seat of our mental life, and the activities of the embodied mind manifest themselves in action.

Different versions of this immaterialist or dualist view are possible. On one version, a person is to be identified with their immaterial soul. On this version, a person, strictly, has no physical parts, although their body does. Hence, if the name 'Smith' refers to a person, a judgement such as 'Smith is six feet tall' should not be taken at face-value, if it is to express a truth. Rather, it should be understood as elliptical for the judgement 'Smith's body is six feet tall'.

On a more liberal version of dualism (in fact, Descartes' version), a person should be regarded as a composite or 'union' of soul and body. On this version, it is true that Smith is six feet tall, in virtue of the fact that a part of Smith (his body) is six feet tall. (In general, where F is a proper part of G, it is not peculiar or unusual for the truth-maker for 'F is H' to be a truth-maker for 'G is H'. For example, it is true that the union voted for the motion in virtue of the fact that a part of (or member of) the union (its President) voted for it.) Nonetheless, on this version, even though body and soul are both parts of me, only my soul is essential to my identity. I continue to exist if and only if my soul continues to exist.

As stated, neither version of dualism strictly implies that it is possible to survive bodily death. For all that has been said so far, the soul may depend for its existence upon the existence of the body. However, traditional defenders of dualism (most notably, Descartes) have taken it that the soul can exist in the absence of the body. Indeed, in speaking of mind and body as 'distinct substances', Descartes meant to imply that either could exist without the other. On this more traditional conception of dualism, a person can exist without their body.

The motivation for dualism derives largely from two sources. Descartes' source was conceivability. We can coherently conceive of ourselves surviving into the future without a physical body; therefore, such survival is a 'real' possibility for us; therefore, we must currently have an immaterial part that is essential to our identity. In order to be maximally compelling, such conceivings are typically presented in the first-person. Thus, we each think: 'I conceive that *I* will survive into the future, in the absence of any physical support'. This seems a coherent speculation since nothing in the meaning of 'I' counts against it.

Arguably, however, this imaginative exercise only appears to imply a 'real' possibility because we read a metaphysical thesis into the epistemological fact that many 'I' -judgements (for example, 'I have a headache') are made 'directly', and not on the basis of bodily observation or inference. We are tempted to conclude from this that our identity over time must be metaphysically unconstrained by physical continuities. But it is fallacious to infer the absence of such constraints from the 'directness' of certain 'I' -judgements (see Chapter 8).

Moreover, conceivability does not seem to offer a stable rationale for dualism. If it is allowed that I can conceive of existing without any physical substance, why can I not equally conceive of existing without any mental substance (that is, without a soul)? But, in that case, I would have imagined away immaterial substance. We would then be left with a version of the 'bundle' theory according to which the self or person is composed of a 'bundle' of immaterial events, which do not inhere in any substance.

The second source for dualism is the belief that nothing purely physical (composed only of swirling atoms, electrons, quarks, etc.) could possibly have an 'inside' or 'conscious interior'. This is a deep-rooted belief. But it is too contentious to motivate dualism. To start with, the belief begins to seem much less compelling once we imagine the atoms organised in a law-like way. Why could the organisation of

the atoms not be responsible for the generation of an inner life? Further, this second source suffers from the same instability alluded to in the previous paragraph: it's not clear how an immaterial or purely mental substance could have an 'inside' either.

Dualism has few philosophical adherents today.[5] In addition to the problems with its motivation, it faces severe difficulties that serve to undermine its credibility. To start with, much of our mentality is essentially bound up with our biological nature (for example, our sensations). Further, if the soul exists, it must interact with the body and with the immediate physical environment. The soul perceives the physical world, and acts on its perceptions via the body. Perception and action imply interaction between soul and matter.

But how are we supposed to make sense of an *immaterial* soul's interaction with a *material* world? In particular, why does the mind, the essence of the soul, directly depend for its normal functioning on the normal functioning of the brain, if the mind is immaterial and the brain is material? More particularly still, if the mind is immaterial, why does it turn out that we generate a bizarre psychological disruption when we divide the two upper hemispheres of the brain?[6]

These worries are metaphysical in character. That is, they are worries about how such disparate substances as soul and matter could possibly interact with each other.

A different set of worries about dualism is epistemic in character. If dualism were true, how could I know that you have a soul and that you are not simply behaving *as if* you had one? Immaterial souls are invisible to the senses, and it is by the senses that we gain knowledge of the world around us, including knowledge of other people.

These questions may not be unanswerable. For example, with respect to the metaphysical worry, perhaps there just is an inexplicable, brute causal link between the material and the immaterial. There is no apparent contradiction in the idea of such a 'brute' relation. Indeed, not all causal laws within the physical realm are explicable. Is there any explanation of why the universe is governed by one set of fundamental laws and not another?

With respect to the epistemic worry, perhaps we can know, on non-sensory or theoretical grounds, that other people have souls. For example, maybe the assumption that they have souls is the best explanation of their behaviour. Science often postulates the existence of the unobservable in order to explain the behaviour of what is observed. Such an 'inference to the best explanation' account may not yield conclusive knowledge of the existence of other souls, but then, it might

be urged, when do we ever have conclusive knowledge about the mental life of other people, or of the external world more generally?

Thus, the standard objections to dualism are not decisive. Nonetheless, dualism is uncongenial to both science and commonsense. We find the idea of interaction between the material and the immaterial hard to comprehend, and increasingly we are able to understand mental capacities (such as memory and vision) in terms of the functioning physical brain. Consequently, we simply have no reason to believe that we are or have immaterial souls. We should reject dualism. Notice, however, that if we believe dualism to be false, we do not have to believe it to be incoherent or necessarily false. It might have been true. That is, perhaps in some other possible scenario (or 'possible world'), persons are immaterial, immortal souls communicating telepathically with each other; but our world is not such a world.

The materialist answer

There are many different versions of materialism about persons. What is common to them all is the rejection of the view that persons are immaterial souls or have any immaterial parts. The most orthodox and familiar versions of materialism each *identify* a person with some particular biological entity. The word 'identify' here should, of course, be taken to denote strict numerical identity, not mere qualitative similarity.

There are three familiar versions of materialism about persons. On the animalist theory, a person is identical to an animal, viz., a human being. On the body theory, a person is identical to a human body. (If we should decide that the human being is not distinct from the human body, then the animal theory and the body theory will simply collapse into one another.) Third, on the brain theory, a person is identical to the physical seat of his mental life, which we have discovered to be the brain and central nervous system.

The above versions of materialism all identify persons with some biological entity. However, there are other, unorthodox, versions of materialism which make no such strict identification. In fact, to each of the above theories there corresponds a theory which refuses to identify the person with any particular biological entity, yet which maintains that the conditions of identity over time of some biological entity (animal, body, brain) trace out the conditions of identity over time of persons.

It might be thought that there is no logical space for such unorthodox theories. How could the identity conditions of a person be determined by those of, for example, a brain, and yet the person not be

strictly identical with the brain? There is, however, no contradiction in the idea that the identity over time of Fs might be fixed by the identity over time of Gs, even though Fs are not identical to Gs.

An example may help. Consider a gold statue. Arguably, the statue is not identical to the lump of gold that composes it. If we melt down the statue, we destroy the statue, but we do not destroy the lump. Since, prior to meltdown, the lump has the property *will exist in a melted state,* which the statue lacks, it follows that the statue and the lump cannot be identical.[7] Still, the identity conditions for the lump fix the identity conditions for the statue in the following way: necessarily, if the statuesque lump continues to exist with its shape pretty much unaltered, then the statue continues to exist.

It's true that the identity conditions of the statue are not the same as those of the lump (as the possibility of meltdown shows), but we still have a necessary connection between identity conditions, despite the numerical distinctness of the statue and the lump. Non-standard or unorthodox materialist theories of personal identity take an analogous form.

In a recent book, Thomas Nagel defends such a theory. He writes that: 'I am whatever persisting individual in the objective order underlies the subjective continuities of that mental life that I call mine…. If my brain meets these conditions then the core of my self—what is essential to my existence—is my functioning brain'.[8] However, he says: 'I am not just my brain: I weigh more than three pounds, am more than six inches high, have a skeleton, etc. But the brain is the only part of me whose destruction I could not possibly survive. The brain, but not the rest of the animal, is essential to the self'.[9]

As the last quote reveals, Nagel is not an orthodox brain theorist. His non-standard counterpart of the brain theory requires the following distinction to be drawn. On the one hand, 'I am identical to my brain' is judged to be false for the sort of reason just given in the statue example, viz., that I and my brain have different properties (for example, my brain weighs only three pounds, and I weigh considerably more). On the other hand, 'The conditions of identity over time of my brain are my conditions of identity over time' is judged to be true, given Nagel's theory and the empirical fact that my brain causally supports my mental life.

Nagel's theory is thus an unorthodox counterpart of the brain theory. The animal and body theories of personal identity also have counterparts according to which the identity of a person over time is fixed by the identity over time of an animal or body, yet the person is not strictly identical to either.

There is a further materialist theory of persons which neither identifies a person with some biological entity nor (like Nagel) holds that a person's conditions of identity over time are just those of some biological entity. On this psychological theory, what is crucial to a person's survival is the survival of their mental life or stream of consciousness (beliefs, memories, character, sense of humour, desires, long-term plans, etc.), and this stream can continue even after the body and brain have been replaced with a new (for example, bionic) body and brain. If I can survive the destruction of my present body and brain, then plainly I cannot be identical to my present body and brain. (If X survives in a given situation and Y does not, X cannot be identical to Y.)

Indeed, on this psychological theory there is no internal requirement that we are essentially material beings. All the evidence suggests that we are material. But it is quite consistent to endorse the psychological theory (and endorse materialism) *and* maintain that we are not essentially material. It may not be physically possible for us to engineer it so that our stream of mental life continues in an immaterial or ghostly substance. But it does seem to be logically possible. If it is logically possible that we exist without a physical body then, although we are embodied creatures, we are not essentially embodied.

It might be thought that this cannot be right. If persons are material, surely they are essentially material? Here we have to be careful. The conditional 'If X is material, X is essentially material' is plausible only on the following reading: if X is identical to a (wholly) material object, then X is essentially material. But this conditional is not violated by the psychological theory since one of its tenets is precisely that a person is not identical to any biological object (nor to any other material object). The psychological theory does not conflict with any plausible version of composition essentialism.

Finally, it is worth mentioning one specific version of the psychological theory which is much more ambitious than the theory just sketched. This is reductionism about persons, a view traditionally associated with Hume and, more recently, with Derek Parfit.[10] There are many ways of understanding reductionism about persons, but one central thought underlying the doctrine is that the concept of a person is a derivative concept, built up primarily out of psychological concepts (memories, intentions, desires, etc.) which, despite appearances to the contrary, can be fully understood in essential respects without reference to the concept of a person or of personal identity.

Whether we can advance from the psychological theory to psychological reductionism is a topic of current debate. This topic is the focus of the second half of Chapter 2. My aim in this chapter, however, is not to evaluate any of the above theories of persons, but merely to indicate the range of theories that will be considered in this book.

What is it for the same person to persist through time?

The traditional question in personal identity is the question of what distinguishes the sorts of changes we can survive from the sorts of changes which constitute our death. We can call this the identity question. Clearly, the best answer to the identity question will be closely connected to the best answer to the nature question. After all, something is an F (has the nature of an F) only if it has the identity or survival conditions appropriate to Fs. Thus there is a sense in which the nature of something determines, and so is prior to, its identity conditions. Indeed, it seems a platitude that the nature of a thing determines the conditions under which it persists. If an F survives a given process, then, *ceteris paribus,* it does so in virtue of the nature of Fs. We might say that the nature of a thing is *metaphysically* prior to its conditions of identity.

Despite this metaphysical priority, the most effective way to answer the nature question may be to adjudicate between rival answers to the identity question. If it can be shown that a person can survive without some particular feature, possession of that feature cannot be essential to that person.[11] It cannot be part of the essence or nature of persons. In this way, the identity question may be *methodologically* prior to the nature question.

Indeed, famously, Descartes gave his answer to the nature question via his answer to the identity question. According to Descartes, it is precisely because we can 'clearly and distinctly' conceive ourselves surviving without a body, that we are entitled to conclude that we are immaterial.

The answers to the nature question sketched in the preceding section all have counterparts amongst answers to the identity question. For the dualist, the identity of a person over time consists in the continued existence of an immaterial soul. According to each of the three materialist views, the identity of a person over time consists in the continued existence of some biological object (human being, human body, or human brain). These three materialist answers to the identity question constitute different versions of the physical criterion of

personal identity over time. (We can call them the animal, body, and brain criteria, respectively.)

Other answers to the identity question are possible. For example, according to the psychological theory mentioned earlier, and without embracing dualism, a person's identity over time can be captured entirely in terms of psychological continuity, that is, overlapping chains of psychological connections (belief, memory, desire, character, etc.) holding between a person at different times. According to Locke, for example, the identity of a person over time is constituted by direct memory connections, independently of whatever substance might or might not support that stream.[12]

The psychological criterion of personal identity has a number of variants, which differ in their specification of the *cause* of the psychological continuity if such continuity is to preserve personal identity. On one version, the cause must be *normal* (that is, the continued existence of the brain and central nervous system) if it is to preserve identity.

On a second version, the cause of psychological continuity merely has to be *reliable* if it is to preserve identity. Thus, consider the teletransporter. A scanner records the exact state of all my cells, painlessly destroys me, and then sends the information to a distant planet, where a molecule-for-molecule replica of me is created. The successful operation of the teletransporter, which ensures psychological continuity in the absence of any continuity of material structure, would preserve personal identity on the second version of the psychological criterion. The cause of the psychological continuity linking me to my replica, though abnormal, is reliable.

On a third version of the psychological criterion, *any* cause of psychological continuity will do. Even if the teletransporter were unreliable (say, only working one time in ten), my identity would be preserved on those occasions when it did work properly and there was full psychological continuity between me and my replica.

Finally, there is a fourth version of the psychological criterion according to which the identity of a person over time has to be understood in terms of psychological continuity, caused in a way which does not correspond to any of the three ways mentioned. This version of the psychological criterion will be defended in Chapter 3.

The methodology of thought-experiments

As argued above, the best way to answer the nature question is by answering the identity question. But how should we answer the iden-

tity question? Evidently, consideration of ordinary cases will not help us to decide the issue. For example, when I judge that the lecturer before me now is identical to the lecturer who began speaking an hour ago, I typically make this judgement of identity under pretty much optimal conditions. In such a case, I can observe that the earlier person is both physically and psychologically continuous with the later person. The very same brain and body has persisted for one hour, and that brain (we may suppose) has directly supported the very same beliefs, character, desires and memories (with only very slight changes).

In this everyday case, my judgement of identity is based on the obtaining of both physical and psychological continuities. Reflection on such a case evidently will not help to determine which continuity (if either) is more important or central to the identity of a person over time. We will need to consider *thought-experiments* in which these continuities come apart. The events depicted in the thought-experiments in this book are all technically impossible at present, and may always be so. But we have no reason to think that any of the thought-experiments is physically impossible (that is, inconsistent with the laws of nature). And, certainly, none is logically impossible.

The use of thought-experiments in philosophy has been subject to a number of criticisms. It has been claimed that we should not take our intuitions about thought-experiments as guides to philosophical truth, since such intuitions may be prejudiced and unreliable. This criticism is, I think, over-stated. For one thing, it ignores the frequent and legitimate use of thought-experiments in virtually all traditional areas of philosophy (most notably, for example, in theories of knowledge and in ethics).

Second, and more important, thought-experiments can be useful in understanding the structure of a concept and the relative importance of its different strands, provided that there is general agreement about the best description of the thought-experiment. It's true that some philosophers have tried to gain mileage from thought-experiments in the absence of such general agreement. But it would be unwarranted to infer from the existence of such abuses that thought-experiments can never perform any useful function in philosophy.

Thus, consider Wittgenstein's verdict on the following thought-experiment:

> Imagine a man whose memories on the even days of his life comprise the events of all these days, skipping entirely what

> happened on the odd days. On the other hand, he remembers on an odd day what happened on previous odd days, but his memory then skips the even days without a feeling of discontinuity…. Are we bound to say that here two persons are inhabiting the same body? That is, is it right to say that there are, and wrong to say that there aren't, or vice versa? Neither. For the *ordinary* use of the word 'person' is what one might call a composite use suitable under ordinary circumstances. If I assume, as I do, that these circumstances are changed, the application of the term 'person' or 'personality' has thereby changed; and if I wish to preserve this term and give it a use analogous to its former use, I am at liberty to choose between many uses, that is, between many different kinds of analogy. One might say in such a case that the term 'personality' hasn't got one legitimate heir only.[13]

Wittgenstein has here described a nice case where neither the answer 'Only one person occupies the body throughout' nor the answer 'Two people alternately occupy the body' are correct or satisfactory. That is, Wittgenstein's thought-experiment exploits the vagueness or indeterminacy of our concepts *person* and *same person*. We may choose to stipulate a more precise meaning for the term 'person', allowing us to say, for example, 'the case involves two people'. But, if we do so, we must be aware that that is what we are doing. We are not reading-off a definite answer from our concept of a person—a concept clearly not designed to yield a yes-or-no answer to questions of personal identity in all possible cases. (The concept *person* is vague in another way too: it can sometimes be vague whether a given entity (for example, a neonate) is a person. But such vagueness is not relevant to the present discussion.)

However, none of this tells against the methodology of thought-experiments. It just shows that, in some thought-experiments, there is no definite answer to questions of personal identity. This is a result that no one ought to dispute.

In this book, we will appeal to a number of thought-experiments to help decide the identity question. The point of these thought-experiments is to enable us to extract a core (that is, minimally controversial) set of common-sense beliefs about the conditions of personal identity over time. In all these thought-experiments, unlike in the Cartesian thought-experiment of a soul floating free of a dead body, we respect the empirically supported fact that states of the mind depend

upon states of the brain. This gives our thought-experi-ments a grounding that Descartes' conceivings lacked.

Here, briefly, are some of the thought-experiments which will feature in subsequent chapters:

Brain Transplant

My brain is removed from my body, kept alive, and then hooked up inside a new skull and body, exactly similar to my old skull and body. My old body is destroyed. The resulting person has my brain and a new body. Since my brain directly supports my mental life, the new person is psychologically continuous with me.[14]

Scattered Existence

My brain is removed from my body and stored in a vat. It is 'connected' to my now brainless body by radio links. I can 'see' and 'hear' appropriately placed objects in the vicinity of my body, yet my brain is hundreds of miles from my body. Suddenly, an avalanche destroys my body. I am still conscious, but receiving no sensory input….[15]

Bionic Replacement

My brain develops cancer. Technology has reached the stage where any human brain function can be mimicked by an appropriate collec-tion of silicon chips. So my surgeons offer to carry out the following operation: they will gradually replace all my biological brain with silicon parts. I will end up with an entirely bionic brain. The new bionic brain will subserve the very same psychological functions as the original. In other words, I will be psychologically continuous with the resulting individual composed of a flesh and blood body and a bionic brain.

Teletransportation

On Earth, I step into the scanner. The function of the scanner is to create an exact atom-for-atom blueprint of me, and then painlessly to destroy me by vaporisation. On the surface of a distant planet, out of different matter, a replicator receives the blueprint and creates an exact replica of me. The replica looks like me, and has all my physical characteristics. He also has all my mental characteristics,

since mental properties depend significantly on physical properties of the brain, and the replicated brain is physically identical to my original brain. Yet my replica has no material substance in common with me.

Branch-line

I am replicated on the distant planet's surface, but the scanner on Earth is now programmed not to vaporise me. However, the operation of the scanner causes me to have cardiac failure on Earth. I am still conscious, and know that I have only a few days to live.[16]

Accident

I am in a horrendous car accident, and suffer massive brain damage. In fact, my psychological life has been completely destroyed, but my body and brain are artificially kept alive. The surgeons find a way to make my brain function again. But complete re-training is necessary. It takes years to advance from the psychological level of a newborn infant to that of a normal adult. The resulting person is quite unlike me psychologically. He and I are not at all psychologically continuous.[17]

Indeterminacy

An alteration machine changes me physically and psychologically. My brain is refigured so that roughly half of my memories, beliefs, desires, and character traits are replaced with new and very different ones. It is vague or indeterminate whether I am psychologically continuous with the resulting person.

Fission

My body is riddled with cancer. The surgeons want to try out a new technique: hemisphere transplant. They have two brainless donor bodies available, cloned years ago from my body. Each of my two brain hemispheres is removed and placed in its own body. Two persons result. Since I am one of the few people whose brain hemispheres are functionally equivalent (that is, they support the very same mental capacities), both resulting persons will think they are me, and they will both have my character, apparent memories, and all my other psychological features.[18]

In these thought-experiments, the first question to ask is: What has happened to me? Have I survived? Have I died? Or is there no definite answer? We shall address these questions in coming chapters.

Why is personal identity important?

The topic of personal identity lies at the intersection of metaphysics and morals. For many philosophers, therein lies its real importance. It is because the concept *person* is a moral and legal concept (or a 'forensic' concept, as Locke described it) that we must be clear about our identity and what it involves. The concept of a person is loaded with assumptions of duties and rights, and hence its proper construal is of obvious moral importance. For example, many familiar positions on abortion and euthanasia presuppose particular conceptions of persons.

More recently, some philosophers—in particular, Derek Parfit—have tried to forge a more interesting connection between theories of personal identity and value theory (ethics and rationality).[19] The possibility of such a connection had not previously been investigated in any detail. Parfit has argued that, on the best theory of personal identity (which, for Parfit, is psychological reductionism), identity is not what matters. What matters is the preservation of psychological relations such as 'apparent' memory, belief, desire and character, etc. Unlike identity, these relations can hold between one earlier person and two or more later persons (as in *Fission*). They can also hold to varying degrees (for example, I can acquire a more or less different character over a period of years, or more quickly, as in *Indeterminacy*).

This view of what matters has implications for theories of compensation and punishment. For example, a now reformed criminal may deserve less or no punishment for the crimes of his earlier criminal self, provided that there have been sufficient and appropriate psychological changes. More recently, Parfit has argued for the even more radical conclusion that no one *ever* deserves to be punished for what they did, even in the absence of any psychological changes.[20]

Another important effect of discussions of the importance of personal identity has been to provide a new perspective on the debate between utilitarianism and its critics. Parfit has argued that reductionism lends support to a more impersonal ethic which ascribes no weight to distributive principles (principles of just distribution). For example, an impersonal ethic would justify

assigning a pain to person A who has suffered much in the past in preference to assigning a slightly greater pain to person B who has led a relatively pain-free life. Such an ethic gives no weight to the distributive principle: distribute pain fairly between lives. The utilitarian ethic ascribes no weight to distributive principles. It aims simply to maximise the *net* sum of happiness over suffering. The connection between personal identity and value theory is the topic of Chapter 6.

2

ANIMALISM AND REDUCTIONISM

Animalism

In Chapter 1 we mentioned animalism as one answer to the question 'What is a person?'. A recent defender of animalism, David Wiggins, has stated his preferred version of this theory as follows:

> *x* is a person if and only if *x* is an animal falling under the extension of a kind whose typical members perceive, feel, remember, imagine, desire, make projects…have, and conceive of themselves as having, a past accessible in experience-memory and a future accessible in intention,…etc.[1]

The basic doctrine of animalism is defined as follows: *x* is a person only if *x* is an animal. Note that it is not required by this doctrine that all persons have to be human beings. Chimpanzees and dolphins, for example, could qualify as persons if their behaviour revealed a suitably impressive mental life.[2]

On Wiggins' version of animalism, the doctrine is *relational* in character. That is, whether a particular animal is a person depends upon the psychology of typical members of its kind (that is, upon the psychology of *other* individuals). If we assume, perhaps uncontroversially, that human foetuses and the irreversibly comatose are human beings, the relationality clause allows foetuses and the comatose to count as persons in virtue of the fact that typical adult human beings are rational and self-conscious.

On the other hand, the relationality clause would exclude an intelligent, self-conscious creature from the extension of person if typical members of its kind happened not to be rational and self-conscious. Thus, if self-consciousness were induced into a single

orang-utan by some bizarre experiment in neuro-engineering, the resulting creature would not count as a person by the lights of Wiggins' relationality clause, since typical orang-utans are not self-conscious. (Assuming, what is not uncontroversial, that the resulting creature could rightly be said to be an orang-utan.) Some would find this consequence implausible. Fortunately, inclusion of the relationality component is no essential part of the formulation of animalism.

The animalist's claim that all persons must be animals, if true, will have to be an a posteriori truth. Nothing in our concept of a person supports the constraint that all persons must be animals. It is neither analytic ('true in virtue of meaning') nor a priori ('known independently of experience') that we are human beings. Rather, it is an empirical fact that we belong to the biological kind *human being*. Hence, people who believe in the possibility of synthetic or robot persons are not committing any conceptual error—as they would if it were a conceptual truth that all persons are animals.

Animalists must therefore conceive of their definition as a necessary a posteriori truth. And its source must lie in a further supposed truth: that animality is a necessary a posteriori constraint on possession of the self-conscious mental life characteristic of persons. However, we simply have no reason to believe in the existence of such a constraint. The fact that all actual self-conscious beings are animals gives us no reason to think that all possible self-conscious beings are animals. And it is the latter thesis that must be defended if animality is deemed a necessary condition of self-consciousness.

Animalism is therefore unmotivated. Further, as we shall see in Chapter 3, the most plausible description of certain thought-experiments implies that animalism is false. First, however, I want to outline and criticise an interesting argument which purports to show that animalism must be the correct view of persons.

An argument for animalism

Animalists sometimes try to argue for their view by appealing to a thought-experiment such as *Accident*.[3] In this thought-experiment, I am involved in an horrendous car accident which irreversibly 'wipes out' all my mental states. The neurosurgeons repair my brain, but the resulting individual has the mental age of a one-year old, and has to be completely re-trained. A number of years later, a re-trained person occupies my body, and that person is psychologically quite unlike me.

To avoid begging any questions about who is who, let's call the human being who exists throughout 'Animal', and call the resulting post-accident person, 'Reverse'.

According to the animalist, the correct description of this case is quite straightforward. I am identical to Animal and to Reverse; we are all one and the same person. According to the anti-animalist, who will typically take psychological continuity to be a necessary condition of personal identity, I ceased to exist when I irretrievably lost all my mental states, and Reverse is a new person who occupies my old body.

The thought driving anti-animalism is that our concept of a person satisfies the moral, practical and theoretical need we have for a conception of ourselves that does not simply coincide with that of a human being. Persons are essentially psychological beings (or, at least, essentially have the capacity for self-conscious life). Human beings, in contrast, can still exist even when the capacity for self-conscious life has been extinguished. Human beings are only contingently psychological beings.

Thus, according to the anti-animalist, the best description of *Accident* is that Animal, Reverse, and me, are three numerically distinct entities. I cannot be identical to Animal if the latter survives while I die; and, for the anti-animalist, I am not the same person as Reverse since I am not psychologically continuous with Reverse.

What is the objection to anti-animalism? The animalist's counter-argument is a *reductio ad absurdum.* That is, he tries to show that anti-animalism is committed to an absurdity, and hence that we should reject anti-animalism and embrace animalism. The argument runs as follows. Suppose that the anti-animalist view is correct, and that Animal and I are distinct entities. If Animal and I are numerically distinct, then we were distinct even before the accident. (If X and Y are numerically distinct at one time, they are distinct at all times. Two entities cannot have been identical.)

Prior to the accident, Animal and I occupied the same body. And, at that time, Animal and I were both self-conscious subjects of experience (that is, persons). Hence, *two persons* then occupied the same body at the same time. What is true of Animal and me prior to the accident is true of all of us now. Each normal adult human body houses two people. The population of the world is twice what we thought it was. Is this consequence not plainly absurd? If it is, and it surely is, then we should reject the anti-animalist assumption that Animal and I are distinct, and embrace animalism.

The animalist's argument rebuffed

Fortunately, the anti-animalist has a response to the above argument. The animalist's argument could be generalised to undermine the most plausible view about, for example, the relation between a statue and the lump of matter which constitutes it. In which case, the animalist's argument must be too strong, and so unsound.

Thus, suppose that we have a bronze statue before us, and we call the statue 'Statue' and the lump 'Bronze'. It might be thought that Statue is Bronze, that they are one and the same. But, as we saw in Chapter 1, this would be a mistake.

Imagine that we take Statue and melt it down—call this thought-experiment, *Meltdown*. Then Statue has ceased to exist, though Bronze still exists in a formless lump. (Maintaining a statuesque shape is an essential property of Statue, but not of Bronze.) If there are situations in which Bronze exists whilst Statue does not, it follows that Statue is not identical to Bronze. Hence, even before the meltdown, Statue and Bronze are numerically distinct objects. And this is so, even though they are exactly spatially coincident during that earlier time. Prior to meltdown, Statue and Bronze occupy the same space, but they are numerically distinct since governed by different criteria of identity.

This description of *Meltdown* is very persuasive. Yet we could run an exact analogue of the animalist's argument to reduce this description to absurdity. We could reason as follows. Prior to meltdown, Bronze has 'all that it takes' to be a statue (what more could we demand?); hence, Bronze is a statue (at that time); yet Bronze is not identical to Statue; consequently, *two* statues occupy the very same space at the very same time. This conclusion is absurd, and hence we should reject the premise that Bronze is not Statue.

I agree that the 'two statue' conclusion is absurd. But the reasoning that led to it is faulty, and there is an analogous flaw in the animalist's *reductio* argument. The mistake is to think that judgements of identity, 'x is (identical to) an F', can reasonably be made on purely synchronic grounds, that is, grounds relating to how things are at just one time, rather than to how things are, or might be, over time.

We cannot judge Bronze to be identical to a statue by considering only its *intrinsic* properties (weight, height, colour, etc.) at some time prior to meltdown. An object can be (identical to) a statue only if any of its possible futures is a future *for a statue*. Since some of Bronze's possible futures are not futures for a statue, it follows that Bronze is not (ever) identical to a statue. We are not forced to the absurd conclusion that two statues occupy the same space at the same time.

Similarly, the anti-animalist will insist that Animal is not identical to a person. Prior to the accident, Animal has all of the intrinsic properties that I then have, but that does not make him (identical to) a person. According to the anti-animalist, the thought-experiment *Accident* shows precisely that one of Animal's futures is a not a possible future for a person. Hence, Animal is (never) identical to a person. Prior to the accident, Animal and I occupy the same space, but Animal is not a self-conscious subject of experience. Hence, the anti-animalist is not forced to the absurd conclusion that two persons occupy my body before the accident.

Wiggins' charge of relative identity

David Wiggins has offered the following additional argument against the anti-animalist. In his description of *Accident,* the anti-animalist must say that I am the same animal as Reverse, but a different person. In which case, the anti-animalist, and anyone else who believes that psychological continuity is necessary for personal identity over time, is committed (absurdly) to the *sortal relativity* of identity.[4]

This is a false charge. Sortal relative identity arises where objects X and Y are held to be numerically identical qua Fs, but numerically distinct qua Gs. ('F' and 'G' stand for sortal concepts, that is, concepts of a kind or sort of object—man, crocodile, tree, etc.). For example, it might be claimed that A and B are identical qua office-holders, yet distinct qua men. Such a claim does indeed verge on incoherence. How can we have a consistent conception of what X *is,* if it falls under sortal concepts that yield mutually incompatible criteria of identity over time?

However, the anti-animalist will deny that the sentence 'I am the same animal as Reverse' is an expression of numerical identity. Rather, he will hold that the sentence 'I am the same animal as Reverse', if it is to express a truth, means that Reverse and I *share our matter* with the very same animal (namely, Animal). Since there is no claim to numerical identity, the anti-animalist cannot be committed to the sortal relativity of numerical identity. Similarly, someone who asserts, for example, 'the boat (at t_1) is the same as the wooden hut (at t_2)', may charitably be interpreted, not as making a (false) statement of numerical identity (a boat cannot become a wooden hut), but as claiming (truly) that the boat and the hut are, at different times, composed of the very same planks.

Consequently, neither the anrmalist's argument nor Wiggins' additional argument are persuasive. We have not been forced to acknowledge any incoherence or absurdity in the anti-animalist position. Moreover, that position, unlike the animalist's, is well-motivated. It respects the intuitive differences between the concepts *person* and *human being*—most importantly, that *person,* unlike *human being,* is the concept of an essentially psychological being. And as we shall see in Chapter 3, the most plausible description of certain thought-experiments provides counter-examples to animalism.

Models of reductionism

Reductionism about persons is an ontological doctrine that is hard to elucidate and, therefore, hard to evaluate. Part of the difficulty is that there is no single doctrine that goes under the name 'reductionism'. It is an umbrella term covering a wide variety of doctrines. Nonetheless, there is a common motivating idea which these doctrines attempt to express. This is the idea that, once we have given up the immaterialist or dualist conception of the person, we should see the ontological status of persons as somehow secondary or derivative relative to the supposedly primary elements of our ontology, for example, physical bodies or mental and physical events.

However, it should be noted that the rejection of dualism, in itself, does not entail the truth of any version of reductionism. There is no incoherence in the conception of persons as entirely material, yet ontologically irreducible relative to the stream of mental and physical events that compose the lives of typical adult humans. In fact, this may well be the best conception of persons. But, obviously, we are not entitled to that conclusion until we have assessed the versions of reductionism about persons that have been proposed. We will consider seven models of reduction, and we will try to discern which model, if any, provides the most plausible model for reductionism about persons.

The eliminativist model

Some versions of reductionism are straightforwardly eliminativist. For example, reductionist theories about phlogiston, the average man, or abstract objects, are typically eliminative in character. That is, such theories maintain, for a variety of reasons, that there is no phlogiston, no average man, and that there are no abstract objects.

(Of course, there are differences between the theories: for example, eliminativism about the average man does not impugn the truth of sentences containing 'the average man', whereas reductionism about phlogiston does imply that sentences containing 'phlogiston' —apart from 'there is no phlogiston' —are all false. Nonetheless, the theories are eliminative: both deny that 'the average man' and 'phlogiston' refer.)

Applying this model to persons, we arrive at the view that, despite the way we talk, there are no persons. There are simply mental and physical events, particular collections of which we lump together in thought and speech, and denote using personal proper names and personal pronouns. But these linguistic items correspond to nothing in reality. Talk of persons is merely a conventionally determined way of talking about collections of mental and physical events.

Many have thought that eliminativism about persons receives support from Hume's famous remark about the unencounterability of the self in introspection. Hume wrote:

> When I enter most intimately into what I call *myself* I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never catch myself at any time without a perception, and never can observe anything but the perception.[5]

However, Hume's remark does not imply that there is anything 'fictional' about persons or personal identity. The unencounterability of oneself in introspection does not exclude the possibility of other modes of access to oneself (for example, seeing oneself in a mirror); *a fortiori,* it does not imply the non-existence of the person. Moreover, as we shall see in our discussion of the epistemic model, there may be weighty and countervailing *theoretical* reasons why belief in the existence of persons is indispensable.

The scientific identification model

A different model of reduction is that provided by scientific identifications, such as those of water with H_2O or lightning with electrical discharge. The identification of water with H_2O is clearly not intended to be eliminative (water undeniably exists). Just for that reason, it might be wondered how the identification of water with H_2O can be a reduction. This is not merely a terminological quibble. The

relation of 'identification' (that is, 'identity') is symmetric (if A is B, B is A). Yet reduction is an asymmetric relation (if As reduce to Bs, Bs cannot reduce to As). A symmetric relation cannot ground an asymmetric relation, since the former cannot account for the direction of the latter.

The best response to this worry is to see the reduction as theoretical. The identification succeeds in giving a direction to the reduction (water to H_2O, and not vice versa) by facilitating the description of yet another natural substance in the terms of a basic scientific vocabulary. It is this latter consideration which justifies talking of a reduction of water to H_2O.

Do scientific identifications provide a plausible model for reductionism about persons? Clearly, only if there is some scientifically describable entity or substance with which a person might be identified. The candidates are presumably biological entities: either the human being, the human body, or human brain. Unfortunately, no such identification is plausible, given the dominant reaction to certain thought-experiments, discussed in the next chapter. These are thought-experiments the best description of which is that a person survives in a given scenario, even though no biological entity with which he might be identified survives. It follows from such descriptions that a person is not identical to, and so cannot be identified with, any biological entity.

The entailment model

This model of reductionism is a popular one. The leading idea is that we can reduce Fs to Gs—nations to people and territory, artefacts to their constituting matter, moral properties to natural ones, mental properties to physical ones, etc. —if the existence of Gs entails the existence of Fs, or if the Gs' possession of certain properties entails the Fs' possession of certain properties.

Granted the existence of the Gs, the thought runs, such an entailment gives us the Fs at no additional ontological cost. So, for example, if our ontology includes natural properties, and an object's possession of certain natural properties entails its possession of certain moral properties, then the latter properties can enter our ontology for 'free'. A 'complete' description of the world (one that entails all truths) need not make explicit reference to moral properties provided it explicitly refers to natural properties.[6]

To begin with, we will need to make a couple of relatively minor modifications to the leading idea. We should require that the Gs constitute or compose the Fs in order for the reduction to go through.

Without this qualification, we would be able to conclude that holes can be reduced to doughnuts since the existence of doughnuts entails the existence of holes.

Second, the Fs in question will have to be contingent existents, otherwise the model would imply that necessary beings (if such there be) can be reduced to everything else. (If Fs are necessary existents, then 'Fs exist' is a necessary truth, and a necessary truth is entailed by any statement whatsoever.) So, for example, unless we exclude necessary existents, God would be reducible to anything at all—a conclusion which obviously does not fit with the idea of reduction.

The leading idea has to be further refined in more substantial ways. Where the existence of Gs entails the existence of Fs, two further constraints will have to be met if the reduction of Fs to Gs is to succeed:

(i) the existence of Fs must not entail the existence of Gs;
(ii) the concept of a G must be intelligible independently of the concept of an F.

The first constraint is needed in order to respect the asymmetry of the relation of reduction. If this constraint were not met, we could not give a reduction of Fs to Gs the direction it requires. The second constraint is needed in order to ensure that Gs are an 'appropriate' reductionist base for Fs. (For example, we could not 'reduce' monetary systems to coins, if the concept of a coin presupposed the concept of a monetary system. Such circularity would undermine any possibility of reduction.)

The central motivation

The fundamental motivation lying behind the entailment model is a belief in Hume's Principle that there cannot be 'necessary connections between distinct existences'.[7] Since entailment is a necessary connection, this principle implies that if the existence of Gs entails the existence of Fs, then Fs and Gs cannot be 'distinct existences'. Of course, to claim that Fs and Gs are not 'distinct existences' is not to claim that they are identical—they may overlap like Siamese twins. But such overlap is not a relevant possibility in the cases under consideration here. Ignoring this possibility, if Fs and Gs are not 'distinct existences', then they should be identified

(exactly the conclusion of the scientific identification model of reduction).

However, the appeal to Hume's Principle provides at best a limited motivation. Hume's Principle is plausible in the case of necessary connections which are analytic, but not otherwise. If it is analytic that the existence of Gs entails the existence of Fs, then, plausibly, Fs cannot be anything distinct from Gs. A complete description of reality need not explicitly refer to Fs, it need only explicitly refer to Gs. It is analytic that unmarried men are bachelors; hence, bachelors are not distinct from unmarried men. It is analytic that truths about concrete men imply truths about the average man; hence, truths about the average man are not distinct from truths about concrete men.

Unfortunately, many putative connections that are of philosophical interest (physical to mental, natural to moral, wavelength to colour, and so on) are typically synthetic and a posteriori necessities, if necessities at all. In such cases, Hume's Principle has no intuitive force. Why can there not be non-analytic necessary connections between distinct and mutually irreducible entities?

Certainly, believers in the doctrine of the 'necessity of origin' are committed to such non-reductive, metaphysical necessities: on this view, the existence of Smith necessitates the existence of his father, yet Smith's father is not 'reducible' to Smith. A description of reality that referred to Smith, but failed to refer to Smith's father, would not be complete. The notions of completeness and reduction are not fully captured by the notion of entailment. It seems, therefore, that the entailment model, applied to a posteriori necessities, is not adequately motivated.

Two questions

Even if we bracket these worries, the entailment model will provide a plausible model for reductionism about persons only if we can give affirmative answers to the following pair of questions. *First,* is there an entailment from the existence of mental and physical events/objects to the existence of persons? *Second,* are constraints (i) and (ii) above satisfied?

First question

There will be no entailment from the existence of appropriately arranged atoms and molecules to the existence of persons if there can

be 'bare' or 'ungrounded' transworld identities and non-identities between persons. That is, if two possible worlds might agree in the identity, existence and arrangement of atoms and molecules, yet differ in the identities of persons existing in the two worlds.

Could there be a world just like the actual world at the level of physical and mental events, but in which Richard Nixon occupies my body and I occupy his? If there can, then the transworld identity 'Garrett (in this world) is identical to Garrett (in that world)' and the transworld non-identity 'Garrett (in this world) is not identical to Nixon (in the other world)' will both be ungrounded.

It is hard to say why there is any incoherence in the supposition that there are such ungrounded instances of transworld identity and non-identity. Resistance to this view is often based on a conflation between conditions of identity over time and conditions of identity across possible worlds. Identities and non-identities over time of ordinary empirical objects do need to be grounded. If A at t_1 is identical to B at t_2, the truth-maker for this identity will be connected to more than the fact that A and B are identical. It will also implicate facts about continuities (physical and/or psychological). But why assume that there is any analogous constraint in the case of transworld identities and non-identities?

It might be thought that if we were to embrace ungrounded transworld identities and non-identities, we would have to accept that there are *no* non-trivial essential properties (so that Nixon could have been a poached egg, for example). However, belief in the sort of ungrounded transworld identities and non-identities countenanced above (for example, 'there is a world in which I have all of Nixon's logically shareable actual properties, and he has all of mine') is compatible with belief in fundamental kind essentialism (the view that, if x is an F, where F is a fundamental kind ('man', 'dog', 'tree', etc.), then x is essentially an F). So the metaphysical view on which the above entailment would fail need not be a radically anti-essentialist view.

The second question

Are constraints (i) and (ii) satisfied? What of constraint (i)? Even if we were to endorse the original entailment, is it consistent to accept that entailment, yet reject its converse? If the original entailment holds, then, for example, in every world in which Nixon's actual functioning body, brain, etc., exist, Nixon exists. If the converse entailment fails to hold, then there are some worlds in which Nixon exists, but in which

Nixon's actual functioning body, brain, etc., don't exist (Nixon occupies a different body in those worlds).

Conjoining these claims, it follows that in all worlds in which Nixon occupies a body different from his actual body, Nixon's actual functioning body doesn't exist in those worlds (assuming that Nixon cannot occupy two bodies simultaneously). But such ontological exclusion seems very odd. Why can't two these distinct and non-overlapping living bodies co-exist in one world? It seems that this oddness must have its source in our assumption that it is consistent to accept the original entailment yet reject its converse. Consequently, constraint (i) is not satisfiable: acceptance of the original entailment requires acceptance of its converse.

The second part of the second question—the satisfaction of constraint (ii)—is complex. If the reductionist base includes, not just atoms and molecules, but also episodes in the mental life of a typical person, there will be a legitimate worry whether constraint (ii) is satisfied. This will be at centre stage in our discussion of the next model for reductionism.

In sum: the entailment model of reduction is unmotivated in the cases that are of philosophical interest; it is doubtful whether the target entailment holds; and it seems that constraint (i), at least, is unsatisfiable. This model is not promising.

The epistemic model

According to the epistemic model, we can reduce Fs to Gs provided that we can understand the concept of a G without reference to the concept of an F. That is, provided that the concept of a G has no analytic links with the concept of a F. This model fits nicely with our conviction that an artefact such as a house is 'reducible' to its component bricks—the concept of a brick can, after all, be understood without reference to the concept of a house.

As stated, the epistemic model provides only a necessary condition for reduction. It says that we can reduce Fs to Gs *only if we* can understand the concept of a G without reference to the concept of an F. Clearly, as it stands, we cannot take this condition also to be sufficient. If we did, we could reduce tables to mountains since the concept of a mountain can be understood without reference to the concept of a table.

In order to have a sufficient condition that avoids this consequence, we could require that, for example, Fs be composed of Gs. Importantly, however, such an additional constraint need not appeal to the notion of entailment. Even if we took the view according to which the existence

of a given artefact is not entailed by the existence of its (appropriately arranged) parts, we would still have an interesting sense in which we ought to be reductionists about artefacts. Thus, the epistemic model does not collapse into the entailment model.

As noted above, an appealing feature of the epistemic model is that it underwrites our intuitive belief that an artefact is reducible to its component parts. A house can be reduced to its component bricks since a house is composed entirely of bricks, arranged in a certain way, and the concept of a brick can be understood without reference to the concept of a house.

Analogous considerations apply to social objects such as nations and committees. A committee can be reduced to its members since a committee is composed entirely of persons, organised according to certain rules of procedure, and the concept of a person can be understood without reference to the concept of a committee.

What of persons? I assume that concepts of physical objects and events (bodies, brains, neural firings, etc.) can be understood in impersonal terms. But persons are mental beings. Can mental events and happenings be understood in impersonal terms? Put thus, this question might prompt the following answer. Suppose some version of physicalism is true, say the token-token identity theory. Token mental events are identical to token physical events. Then, since physical events can be understood impersonally, and mental events are physical events, it follows that mental events can be understood impersonally. So our question is easily answered if such a version of physicalism is true.

I do not deny that if such a version of physicalism were true, and if persons were identical with physical objects, then some doctrine worth calling 'reductionism' would be true. However, such reductionism is tangential to the concerns of the epistemic model. This model is concerned with relations between concepts, not with relations between objects or events. Its target question is: can the concept of a G be understood without reference to the concept of an F? If so, then we can make logical space for the reductive claim that, for example, Fs are 'nothing but' Gs.

To see whether the epistemic model yields a sense in which we ought to be reductionists about persons, we should ask the following question: Is the relation between the concept *person* and our range of mental concepts analogous in relevant respects to the relation between our concept of a house and the concept of a brick? Can the mental concepts be understood in impersonal terms? This question is difficult to answer, and it has rightly been the focus of much contemporary debate. Let me begin by briefly outlining three arguments that have

been given for returning a negative answer, and so for concluding that we should not, in the sense currently entertained, be reductionists about persons.

First, there is the familiar observation that we find it hard to make anything of the concept of an unowned or subjectless experience. A particular toothache, for example, must be had by someone. How then could we understand the concept of an experience without reference to the concept of a subject of experience?

Second, the mental life of a normal adult human being does not simply consist in experiences of pleasure and pain. Such a mental life includes first-personal mental states with complex contents: for example, intending to visit Bangkok next month, or remembering how Alice Springs looked at dawn last year. These mental states are *ascribable* only to persons. (Cats and dogs can have toothache, but they cannot have complex memories or intentions.) How then could concepts of such mental states be understood without reference to the concept of a person?

Third, the *contents* of certain complex first-personal mental states appear to presuppose personal identity. In the case of experience-memory (for example, remembering the taste of yesterday's ice-cream), I can only be said to remember, from the inside, *my own* experiences. Similarly with first-person intentions: I can intend only that *I* do such-and-such. In which case, there is no prospect of a complete and impersonal description of these mental contents—personal identity is built into them.

However, these arguments have not gone unchallenged. It has been replied to the first argument that, although it is true that experiences require subjects, this truth is merely 'grammatical' or 'conventional'. Parfit defends such a view: 'A [r]eductionist can admit that…a person is *what has* experiences, or the *subject of experiences*. This is true because of the way in which we talk'.[8] However, this account would seem to collapse into eliminativism about persons. If 'there are persons' is true *only* because of the way we speak, then surely, in reality, there are no persons.

It has been suggested that, contrary to the second argument, complex memory-ascriptions can be relativised to human mouths or bodies, and not to persons.[9] Finally, and in response to the third argument, some philosophers have appealed to invented concepts such as that of *quasi-memory,* the contents of which are explicitly designed not to presuppose the concept of personal identity.[10]

This reply to the third argument is worth some elaboration. Quasi-memory is stipulated to be like ordinary memory in all relevant phenomenological and causal respects, except that quasi-memory does not presuppose personal identity. We can define q-memory as follows. *A* q-remembers F-ing if and only if (i) *A* has an apparent memory (or memory-like image) of F-ing; (ii) *A*'s apparent memory is caused in a way 'relevantly similar' to ordinary memory; and (iii) it is not presupposed that *A* F-ed.

The concept of q-memory, so defined, is perfectly coherent. In fact, ordinary memories are a sub-class of q-memories in the following sense: anyone who is in a state of remembering F-ing will be in a state of q-remembering F-ing (which is not, of course, to imply that the states are the same).

The cases of most interest are those in which someone q-remembers F-ing, but does not remember F-ing (that is, a case in which personal identity is lacking). So let us imagine a case in which I quasi-remember, from the inside, someone else's experiences. For example, imagine that, as the result of a brain-graft from an Indian, I come to have a quasi-memory of standing in front of the Taj Mahal (even though I know that I have never been to India). I take it that clauses (i)–(iii) of our definition are satisfied. (We can suppose that continuity of living brain matter ensures that the causal link is 'relevantly similar' to those involved in cases of ordinary memory.) I do not remember seeing the Taj Mahal, but I do q-remember seeing it, and quasi-memory does not presuppose personal identity. Consequently, it is claimed, we can re-describe my psychological life in terms of concepts such as q-memory, without presupposing personal identity.

Indeed, it might seem that such a re-description of our psychological life has to be possible, given the most plausible response to the thought-experiment *Fission,* viz., that I am identical to neither of my fission products, though I am, in some uncontroversial sense, psychologically continuous with both. For if I am identical to neither of my fission products, and yet I am psychologically continuous with both of them, then surely an impersonal (identity-free) description of psychological continuity must be available, utilising precisely concepts like that of q-memory.

Unfortunately, matters are not quite so straightforward. We should not confuse the question of whether the notion of q-memory is coherent (not seriously in dispute) with the crucial question of whether the concept of q-memory is genuinely intelligible independently of the concept of memory. The concept of q-memory may be, at a deep level, as identity-involving as the concept of memory.

John McDowell has emphasised the fact that q-memories and other such memory-like states present themselves as memories.[11] If I was initially unaware of my brain-graft operation, and falsely believed that I visited India, I will be disposed to utter 'I remember seeing the Taj Mahal'. The content of my q-memory will, falsely as it turns out, represent me as the seer of the Taj Mahal.

Importantly, the content of this q-memory will not change when I am informed that, actually, it was not me who saw the Taj Mahal. The content of my q-memory will not suddenly become impersonal (whatever that might mean). My q-memory will still present itself as a memory (and hence as identity-involving), even if I come to believe that its content is illusory in respect of its identity-involving component.

McDowell draws on a useful analogy with the Muller-Lyer illusion. The pair of lines in that illusion present themselves in perception as unequal in length, and they continue to do so, even when one comes to know that they are the same length. Analogously, there is no switch in phenomenology or content when I come to know that my memory-like state is not a memory. Thus, it is a misconstrual of q-memory to think that, when I have a q-memory, I have an impersonal or identity-free memory-like state. Q-memory is an illusion of memory, and so the former concept is not identity-free.

The conclusion McDowell draws from these observations is that the notion of q-memory (and, analogously, other q-notions) are not identity-free, and so cannot serve to vindicate reductionism.[12] More needs to be said, but the onus of proof lies with a defender of this version of reductionism about persons.

The no-substance model

According to our fifth model—the no-substance model—reductionism about Fs is the view that Fs are not substances. Paradigm examples of substances are biological entities such as horses and oak-trees; paradigm examples of non-substances are social objects such as nations and committees.

What, exactly, distinguishes a substance from a non-substance? On one standard view of substances, the difference between a substance, such as an oak-tree, and a non-substance, such as a committee, is that the latter is a 'dependent' or 'notional' being, whereas the former does not depend for its existence upon anything else (apart from the raw materials necessary for its sustenance—soil, CO_2, etc.). One mode of dependence is 'mind-dependence'.

What does it mean to think of dependence as 'mind-dependence'? I take it that an entity is 'mind-dependent' only if it exists, and continues to exist, because of our attitudes towards it. An entity, such as a committee, is mind-dependent in that it depends for its existence upon the existence of certain specific attitudes. If the members of a committee, and any relevant higher authorities, lost their beliefs that they were members of that committee, no longer intended to meet etc., then the committee would cease to exist.

However, a person who believes that he is no longer a person does not thereby cease to exist; indeed, in order to have that belief, he must be a person. Any others who came to regard him as a non-person would, assuming no other changes, simply be mistaken. So, in this sense, persons are not mind-dependent. Hence, the fifth model cannot underwrite reductionism about persons.

'Person' is a phased sortal

The concept *person* is a sortal concept. Within the class of sortal concepts we can distinguish between those which are temporary (for example, *child, sun-bather, acorn*) and those which are permanent (for example, *human being, poodle, horse*). Concepts of the first kind we call phased sortals; those of the second, substance sortals. Anything which falls under a phased sortal typically does so only for part of its existence. But anything which falls under a substance sortal does so for the entirety of its existence.

One expression of reductionism about persons, not unrelated to the previous model, is the claim that the concept *person* is a phased sortal concept, and not a substance sortal concept.

Before asking whether these notions can be characterised more precisely, it's worth appreciating the intuitive motivation for this model of reduction. The thought is that substance sortals characterise the world at its most fundamental level, whereas phased sortals merely serve to characterise temporary 'shapes' assumed by the more basic or substantial entities. Thus, *boy* is a phased sortal, and one mark of this is that when a boy becomes an adolescent, nothing goes out of existence. No substantial change occurs. (Contrast the case of melting down a statue. Arguably, an object—the statue—does go out of existence in this case.)

Not only does nothing go out of existence when a boy becomes an adolescent, but we can completely describe what persists without using the phased concepts *boy* or *adolescent*. A particular human being exists throughout. The human being is the substance, of which the boy and

the adolescent are phases in its history. Alternatively, we can say that 'boy' restricts 'human being', where F restricts G just in case, necessarily, everything F is G, but not conversely.

Thus to hold that *person* is a phased concept is to hold that, for example, when a person expires, leaving only a body, this is not relevantly different from a boy becoming an adolescent. In neither case does any object cease to exist. Rather, in each case, certain physical changes (which take place either instantaneously or gradually) occur in a human organism.

On this view, 'person' restricts 'human being'just as 'boy' restricts 'human being', albeit that 'person' typically denotes a longer-lasting phase than 'boy'. Phased sortals are not required in a complete description of the universe. We need only the underlying substance sortals. From a description in terms of substance sortals all else follows. Hence, just as the category 'boy' is not required in a description of the ultimate furniture of the universe, nor is the category 'person' required either.

I think there are two basic problems with this model of reductionism. One problem is quite general, the other is peculiar to persons.

First, although we have some intuitive grip on the phased/substance distinction, that grip loosens when we try to make the distinction precise. We can attempt first to define 'substance sortal', and then characterise phased sortals as those sortal concepts which are not substance sortals. David Wiggins offers two definitions: (i) 'F is a substance sortal if F present-tensedly applies to an individual x at every moment throughout x's existence'; and (ii) 'F is a substance sortal if *x is no longer F* implies *x is no longer*'.[13]

These definitions are not equivalent, and both are open to objection. The first implies that, for example, 'foetus' counts as a substance sortal since, in the case of an aborted foetus x, the sortal 'foetus' applies to x at 'every moment throughout x's existence'. But, intuitively, 'foetus' should be classified as a phased sortal. The second definition implies that, for example, *person who was born on 1 April 1997* counts as a substance sortal. In which case, we have severed the intended link between the notion of a substance sortal and the notion of a fundamental type of thing. (This link is anyway unclear. The *concept poodle* 'restricts' the concept *dog*. Yet both concepts count as substance sortals by the above definitions.)

Second, 'person' does not stand to 'human being' as 'boy' stands to 'human being'. In fact, there seems no substance sortal which 'person' restricts. In the next chapter, we will argue that there are possible scenarios in which a person survives a complete replacement of all his biological organs with bionic parts. Thus, the person survives, but no

human being survives; hence, 'person' cannot restrict 'human being'. It seems that 'person' is not a phased sortal since there is no substance sortal which it restricts.

It must be admitted that there are pressures to view 'person' as a phased sortal. Didn't I exist when I was a foetus? Since foetuses aren't persons, I wasn't always a person. So perhaps 'person' should be counted a phased sortal. But, as we have just seen, there is the problem of specifying which substance sortal it restricts. Perhaps the proper conclusion to draw is that 'person' is a sortal concept which is not happily seen either as phased or as substance. In which case, so much the worse for the phased/substance distinction, to the extent that it purports to be an exhaustive classification of sortal concepts. And also so much the worse for the attempt to use that distinction as a way of characterising reductionism about persons.

Parfit's reductionism

Parfit does not distinguish the six models outlined above, and at different places, he seems to endorse ingredients in each of the entailment, epistemic and the no-substance models.[14] In addition, he uses the term 'reductionist' in a very broad sense which includes both the thesis that personal identity is not all-or-nothing (there can be circumstances—as in the thought-experiment *Indeterminacy*—in which it is vague or indeterminate whether a single person has persisted), and the thesis that personal identity is not 'what matters'.

This usage is unhelpful. One could reject all the above models of reductionism about persons *and* concede both that personal identity is sometimes vague and that personal identity is not 'what matters'. Moreover, as Shoemaker has pointed out, one could be reductionist, in one or more of the above senses of that word, and think that personal identity over time is a perfectly determinate matter. This would be so, for example, if the conditions of personal identity over time were the same as those for the identity of brains over time and if, for some bizarre reason, it turned out that brain-identity was necessarily determinate.[15]

However, we can distil out these tangential theses, and simply focus on one of Parfit's clear statements of reductionism about persons. In *Reasons and Persons,* Parfit endorsed the following impersonality thesis:

> (1) [T]he fact of a person's identity over time just consists in the holding of certain more particular facts.

and

(2) [T]hese facts can be described without either presupposing the identity of this person, or explicitly claiming that the experiences in this person's life are had by this person, or even explicitly claiming that this person exists. These facts can be described in an *impersonal* way.[16]

and

(3) Though persons exist, we could give a *complete* description of reality *without* claiming that persons exist.[17]

What is it for a description of reality to be complete? According to Parfit: '[I]f our description of reality either states or implies, or enables us to know about, the existence of everything that exists, our description is complete'.[18]

The impersonality thesis combines features of both the entaihnent and epistemic models of reductionism, and does not appear to contain any features not contained in these models. Consequently, our earlier criticisms of those models carry over to Parfit's proposal.

More recently, Parfit has attempted to characterise reductionism as follows. To be a reductionist is to hold that, once we know all the relations of physical and psychological continuity holding between a person at one time and some person at a later time, it is a purely 'verbal' or 'conceptual' decision whether or not we choose to call them the same person. Other examples of verbal decisions are: deciding whether to call sea-sickness 'pain', or deciding whether to call a pile of sand a 'heap'.[19]

Parfit also talks of questions of personal identity as 'empty questions'. He distinguishes two senses of 'empty question'. A question is empty if it has no determinate answer (the question 'Will I survive the operation of the alteration machine?' is thought by many to be empty in this sense). But even where there is a clear answer, a question of personal identity is always empty in a different sense. 'The question is empty because it does not describe different possibilities, any of which might be true, and one of which must be true. The question merely gives us different descriptions of the same outcome.'[20]

This second sense of 'empty question' fits with the idea of a question or dispute as 'verbal' or 'conceptual'. Unfortunately, this way of characterising reductionism about personal identity is imprecise and

hard to apply. Further, Parfit's model would anyway appear to require the truth of the epistemic model if it is to succeed. Parfit's talk of 'different descriptions of the same outcome' appears to presuppose that the 'outcome' can adequately be described in identity-free terms. As we have seen, that is a controversial presupposition.

Conclusion

We have argued that the animalist view of persons is unmotivated. Further, none of the above models of reductionism about persons is satisfactory, although debate about the epistemic model is far from closed.

3

CRITERIA OF PERSONAL IDENTITY

The range of criteria

There are two broad accounts or 'criteria' of personal identity over time: the physical criterion and the psychological criterion. These criteria constitute different answers to the identity question posed in Chapter 1 —what is it for the same person to exist over time? What conditions must be satisfied in order for the same person to continue to exist? The criteria do not purport merely to offer quite general 'ways of telling' or of 'finding out' who is who. They purport to specify what the identity of persons over time consists in: what it is to be the same person over time.

According to the *physical criterion,* the identity of a person over time consists in the holding of some relation of physical continuity between a person at different times. On this view, to be the same person over time is to be the same biological object over time. Different versions of the physical criterion differ over which biological item grounds personal identity over time: the human being (animal), the body, or the brain and central nervous system.

All these versions hold that the conditions of identity of a person over time are fixed by the conditions of identity over time of some biological item. However, no version of the physical criterion by itself implies that a person is strictly or numerically identical with some biological object. As noted in Chapter 1, it is quite consistent to hold that, for example, the identity conditions for brains traces out, or fixes, the identity conditions for persons, and yet also hold that persons are not identical to their brains.

According to the *psychological criterion,* the identity of a person over time consists in the holding of the relation of psychological continuity between a person at different times. This relation is composed of a number of chains of interlocking direct

psychological connections, such as those which hold between an experience-memory and the experience-remembered, between an intention and the action which manifests it, or the chain consisting of the retention of beliefs, desires, memories, character, etc., over time. Like the physical criterion, the psychological criterion splits into different versions. It has a strong version and a weak version.

According to the strong version, the cause of the psychological continuity must be *normal* if it is to preserve personal identity. The normal cause of our psychological continuity is the continued existence of our brain. Hence, on the strong version, if psychological continuity is supported by some abnormal cause, the persons so linked cannot be identical.

Thus, consider the fantasy of *Teletransportation* in which a complete psycho-physical blueprint is taken of one person, who is then painlessly destroyed, and an exact physical and psychological duplicate of that person is created elsewhere out of different matter. Here psychological continuity is preserved, but the cause of that continuity is abnormal. The cause essentially involves the mechanism of the teletransporter, and does not involve the continued existence of a single brain. According to the strong version of the psychological criterion, the pre-teletransportation person and the post-teletransportation person are numerically distinct, however similar they may be.

According to the weak version of the psychological criterion, psychological continuity will suffice for personal identity over time, even if the cause is abnormal. Teletransportation preserves personal identity. Different versions of the weak version differ over whether the cause has to be reliable (that is, whether the mechanism that supports psychological continuity does so, or would do so, most of the time).

On the less extreme version, the cause has to be reliable. On the most extreme version, a person will continue to exist on a given occasion, even where the cause of psychological continuity is abnormal and generally unreliable, provided that, on that occasion, there is full psychological continuity linking the earlier person with the later person. So, according to the most extreme sub-version of the weak version, even if the teletransporter only works one time in ten, still, on the occasions when it does work, it preserves personal identity.

Finally, there is an intermediate criterion which lies between the strong and weak versions of the psychological criterion. I will argue that this criterion provides the best account of personal identity over time.

Versions of the psychological criterion also differ over the issue of whether any one psychological relation is privileged with respect to identity-preservation. Locke, for example, thought that memory was such

a privileged relation. According to Locke, direct memory connections (connections between a memory and the experience remembered) are necessary and sufficient for personal identity over time.[1]

However, Thomas Reid noticed an obvious objection to this account.[2] Memories fade as people get older. Thus the following scenario is quite common: where 'C', 'B' and 'A' name the same person at different times, C remembers B's experiences, B remembers A's experiences, yet C is too old to remember A's experiences. But Locke's theory cannot give a consistent description of this case. Since C remembers B's experiences, and B remembers A's, it follows, on Locke's theory, that C is B and that B is A. Since identity is transitive, it follows that C is A (C is the same person as A). But since C cannot remember A's experiences, it follows, on Locke's theory, that C is not the same person as A. Hence, on Locke's theory, C both is and is not the same person as A.

An obvious response to this objection is to give up the idea that direct memory connections are necessary for personal identity over time, and claim only that continuity of memory (overlapping chains of direct memory connections) is necessary. To do this, though, is to give up much that is distinctive about Locke's account.

I now want to examine in more detail some familiar versions of the physical and psychological criteria. My conclusion will be that all these versions are open to objection, and that we should accept the intermediate criterion. That criterion best captures our core (that is, minimally controversial) beliefs about personal identity.

The physical criterion

We can begin with the physical criterion. As noted, this criterion divides into three criteria: the animal criterion, the bodily criterion and the brain criterion. However, since there could not be a situation in which a human being survives, but in which neither his brain nor his body survives, the animal criterion will entail the disjunction of bodily and brain criteria. And since, as I shall argue, there are plausible counter-examples to that disjunction, so there are counter-examples to the animal criterion. Consequently, the animal criterion will not require separate treatment.

It is worth noting, at the outset, a popular objection to both the bodily and brain criteria. This objection, in the case of the bodily criterion, runs as follows. If I drop dead of a heart attack, my body will still exist. My body won't be a living body anymore, but the same body will still exist. The bodily criterion states that I continue to exist if and

only if my body continues to exist. From this it follows that I exist after I'm dead. Isn't this an absurd consequence? Shouldn't we therefore reject the bodily criterion?

One familiar attempt to rid the bodily criterion of the consequence of possible-personal-existence-while-dead is not fully coherent. Defenders of the bodily criterion sometimes say that they want to identify a person with his living body, not with his body *per se*. But this does not make sense. My present living body is not a different body from my body *per se*—I don't have two bodies. I can't be identical to my-body-only-when-living if *being alive* is an accidental property of my body.

In general, if F is an accidental or non-essential property of x, no sense can be given to the claim that y is identical to x-only-when-F. Either y is identical to x, and hence identical to x even when x is not-F, or y is not identical to x. There is no third alternative.

It might be thought that this cannot be right. For example, surely my fist is identical to my-hand-when-clenched, and not otherwise. However, we have to be careful. My hand is identical to my-hand-when-clenched. If my fist is identical to my-hand-when-clenched, then it follows that my hand is identical to my fist. But that seems wrong: when my hand is unclenched, there is no fist. The best response to this conundrum is to reject the assumption that my fist is identical to my-hand-when-clenched. To reject this assumption is not to contradict the obvious truth that my-hand-when-clenched 'constitutes' my fist. There is no contradiction since constitution is not identity, as our discussion of Statue and Bronze made clear (see *Meltdown* of Chapter 2).

Thus, adherents of the bodily criterion simply have to accept the possibility of personal-existence-while-dead and argue that it does not undermine their favoured criterion of personal identity.

Although accepting the possibility of personal-existence-while-dead hardly fits well with the point of our concept of a person (viz., to delineate a being of some psychological sophistication), it does not provide a conclusive reason to reject the bodily criterion. Appeals to ordinary language are not always decisive. We often speak of our survival in different and apparently conflicting ways. We say 'Bill's no longer with us, we might as well turn off the life-support machine'. But we also say things like 'Look what happened to old Fred, one bite and now he's dead' (that is, being dead is the state that Fred is now in). So ordinary language does not provide a knock-down argument against the bodily criterion.

We shall now assess the two versions of the physical criterion in more detail.

The bodily criterion

According to the bodily criterion, person A at time t_1 is identical to person B at time t_2 if and only if A and B have the same body (that is, they are bodily continuous). Note that on this, perhaps slightly artificial, use of the word 'body', A and B can be said to have the same body even if they have different brains. (Indeed, as we shall see, this possibility will count against the bodily criterion.) Artificial though this use of 'body' might be, it at least has the merit of enabling us to mark a sharp and useful contrast between the two leading versions of the physical criterion of personal identity.

Further, it should be noted that A and B can truly be said to have the same body, even though the body at the later time has no matter in common with the body at the earlier time. This, after all, is true of human bodies over fairly lengthy periods of time. In such cases, however, the identity of body is preserved since the replacement of matter is gradual, and the new matter is functionally absorbed into the living body. (See also the discussion of *Ship* at the end of this section.)

The bodily criterion accords with most of our ordinary, everyday judgements of personal identity. That is, the rubric 'same person, same body' is a perfectly reliable guide to ordinary cases of personal identity over time. However, there appear to be logically possible cases in which the deliverances of the bodily criterion conflict with our considered judgements. One such case is that of brain-transplantation. Such transplants are, of course, technologically impossible at present; but that is hardly relevant. The speculations of philosophers are not confined to what is technologically possible.

Sydney Shoemaker was the first to introduce the thought-experiment *Brain Transplant* into the philosophical literature. He wrote:

> It is now possible to transplant certain organs…. [i]t is at least conceivable…that a human body could continue to function normally if its brain were replaced by one taken from another human body…. Two men, a Mr Brown and a Mr Robinson, had been operated on for brain tumors, and brain extractions had been performed on both of them. At the end of the operations, however, the assistant inadvertently put Brown's brain in Robinson's head, and Robinson's brain in Brown's head. One of these men immediately dies, but the other, the one with Robinson's head and Brown's brain, eventually regains consciousness. Let us call the latter 'Brownson'…. When asked his name he automatically replies 'Brown'. He recognises

> Brown's wife and family…, and is able to describe in detail events in Brown's life…of Robinson's life he evidences no knowledge at all.[3]

We can suppose, in addition, that Brown and Robinson are physically very similar, and that their bodies are equally suited for the realisation of particular dispositions or abilities (for example, playing the piano, or hanggliding). Since the possession of certain psychological properties is intimately connected with the possession of certain physical abilities and dispositions, this latter supposition allows us to make sense most easily of the claim that Brownson is fully psychologically continuous with Brown.

The description of *Brain Transplant* which commands almost universal assent is that Brown is the same person as Brownson. Virtually no-one thinks that the correct description is: Robinson acquires a new brain. Receiving a new skull and a new body seems to be just a limiting case of receiving a new heart, new lungs, new legs, etc. Since grounded in both physical and psychological continuities, the judgement that Brown is Brownson is well-entrenched relative to our system of core beliefs about personal identity over time. Given that Brown is the same person as Brownson, and yet Brownson's body is not the same body as Brown's body, it follows that the bodily criterion is false.

Another thought-experiment which reveals the implausibility of the bodily criterion is *Scattered Existence*. In this thought-experiment, my brain is removed from my skull, placed in a vat, and connected to a machine. The inside of my skull is fitted with electrodes which transmit information to my disembodied, but still conscious, brain, kept alive by the machine. My body functions as normal, receiving information through the senses. My behaviour and appearance are as before, only now my brain is outside my body and sensory input is relayed to and from my brain by radio links. An observer of my body would notice no change in behaviour, and nor would there be any change from the first-person perspective.

This thought-experiment, though technically impossible at present, is perfectly imaginable. There is no more difficulty imagining a disembodied brain still being conscious than there is imagining a totally sensorily deprived person still being conscious. There is no principled objection to mimicking the inputs which our brains currently enjoy in the way envisaged.

In this thought-experiment, it is natural for me to ask the question 'Where am I?'. However, arguably we ought not to find the question as puzzling as have some philosophers.[4] As the name of the thought-

experiment suggests, my existence is scattered—I am in two places at once, in virtue of my brain being in one place and my body being in another, distant, place. But this is not paradoxical. After all, we are all now in many places at once. My head and my feet are presently in different places. (In fact, any decent-sized space-occupying object will be in more than one place at the same time, in virtue of its spatial parts occupying different places.)

It's true that, in *Scattered Existence,* my brain and my body are not connected to each other by intervening matter. But this does not seem to be a relevant consideration. It is not the absence of intervening matter which, in general, makes for paradox. For example, the existence of a scattered edition of *Encyclopaedia Britannica* gives rise to no metaphysical puzzles.

After my brain is removed from my body, my existence is very scattered. Suppose that our thought-experiment continues in the following way. An accident befalls my body. It is destroyed in an avalanche. All communication from my body is cut off, and I am sensorily deprived. (The scientists may later choose to give me the illusion of sensory input from the external environment.) Despite the darkness, however, I know that I still exist. I can still reason, remember, and ask questions of myself. After all, I'm still around to think the question 'What has happened to me?'. The answer is that I now exist, only as a brain and without a body. I still exist, but in a reduced state, having just lost a very large part of myself (viz., my body). But if that is the best description, then the conditions of my body's identity over time cannot be the conditions of my identity over time. Consequently, the bodily criterion of personal identity is false.

The brain criterion

In the light of the previous thought-experiments, it would be natural for a defender of the physical criterion to move to the brain criterion: A at t_1 is the same person as B at t_2 if and only if A and B possess the same brain. But is this a well-motivated or plausible criterion of personal identity?

One motivation

According to the brain theory, a person is identical to their brain. The brain theory implies the brain criterion. One motivation for the brain theory lies in taking seriously an alleged analogy between 'person' and

natural kind concepts such as 'gold' or 'water', at least on one contemporary reading of such concepts.[5]

On that reading, to grasp the concept *water,* or to grasp the meaning of the word 'water', one must know a sufficient number of the cluster of superficial characteristics of water—for example, that water is a colourless, odourless liquid, which falls from the sky and fills our lakes and rivers. But, on this view, water has a 'real essence' (its internal structure) which cannot be 'read off' from the concept *water*. Hence, fully understanding the concept 'water' does not reveal the essence of water. The concept contains a 'gap' which is open to empirical completion. This gap is to be filled by discovering whatever it is about the internal structure of water that is causally responsible for its surface features. This gap was filled by the empirical discovery that H_2O molecules causally explain those surface features; hence, we can identify water with H_2O.

If we think of 'person' as a natural kind term on this model, we will identify a person with his brain, since we know that it is the brain which causally sustains the self-conscious mental life distinctive of persons. On this view, the brain stands to personal identity as the molecular structure H_2O stands to water. Brain identity is the 'real essence' of personal identity. I am identical to my brain, just as water is identical to H_2O. From this it follows that the conditions of my identity over time are just the conditions of my brain's identity over time.

However, whether the concept *person* is analogous to the concept *water* depends, obviously enough, on whether *person* is a natural kind concept. It's never an a priori matter whether a given term is a natural kind term; if it is, that is courtesy of the external world. (For example, imagine that samples of what we all call 'water' turned out to have nothing of interest in common at the molecular level. 'Water' would not then denote a natural kind.)

However, a priori considerations can place obstacles in the way of a term's entitlement to be deemed a natural kind term. In particular, if our judgements in certain thought-experiments tell against the brain criterion, we should simply reject the analogy between 'person' and 'water'. So the question to ask is: are my identity conditions simply those of my brain? Could I have existed without my brain?

We can ask an analogous question of water: could water have failed to be H_2O? Can we imagine encountering a colourless, odourless, tasteless liquid, like water in all superficial respects, yet which has a different internal structure (XYZ rather than H_2O)? If we can, and it seems that we can, doesn't this show that water is not essentially H_2O?

The standard reply to this question is to claim that we have imagined something that is like water in all superficial respects, but which isn't water. In the jargon, we have imagined an 'epistemic counterpart' of water which isn't water.[6]

Whatever plausibility this reply has in the case of 'water', it has little plausibility in the case of 'person'. As we shall see, the best description of certain thought-experiments is that I survive without my (biological) brain. A non-biological 'brain' can support my stream of mental life. In which case, the identity of a person over time cannot consist in the identity over time of his brain. The response that in such thought-experiments we do not have a person, but merely an 'epistemic counterpart' of a person has nothing to recommend it. It would be a misplaced scepticism to think that, all behavioural evidence to the contrary, the resulting individual was not really self-conscious, and hence not really a person. The natural conclusion to draw is that the concept *person* is not a natural kind concept like *gold* or *water.*

The implausibility of the brain criterion

If my brain is my essence, or if the conditions of identity over time of my brain 'fix' the conditions of my identity over time, then it will be a necessary truth that I have a brain. However, it does not seem to be a necessary truth that I have a brain, and nor does it seem to be necessary that all possible persons have a central, relatively self-contained control system which we would call a brain. For example, the mental life of some imaginary creatures might be sustained by their entire physical body.

Moreover, even with regard to human persons, there is a thought-experiment which shows that it is possible for us to survive the destruction of our brains. In which case, we are not identical with our brains, and a person's conditions of identity over time are not those of his brain.

Consider the thought-experiment *Bionic Replacement*. There is no reason to think that it is a necessary truth that only biological systems are mental. Imagine that robotics and brain-science have advanced to such a stage that it is possible to construct a silicon brain which supports the very same kind of mental life that is supported by a flesh-and-blood human brain. Imagine also that any part of a human brain can be replaced by silicon chips which subserve the very same mental functions as the original brain tissue.

Suppose that the whole of my brain gradually becomes cancerous. As soon as the surgeons detect a cancerous part, they replace it with silicon chips. My mental life continues as before—the same beliefs,

memories, character, etc., are preserved. Eventually, the surgeons replace all my biological brain with a silicon brain. Since my mental life, physical appearance, and abilities, are all unaffected by this replacement, we would have no hesitation in judging that I have survived the operation. The procedure preserves personal identity. But is this judgement of personal identity consistent with the brain criterion? The answer to this question depends on whether my (later) silicon brain is deemed to be identical to my (earlier) human brain.

It is plausible to suppose that a biological object such as a human brain is a member of a natural kind. If we accept the essentialist theory of natural kinds described above, then it will follow that if an object, such as a heart, brain or liver, is biological, then that object is essentially biological. That is, for example, my flesh-and-blood brain could not have been anything but a biological entity.

This essentialist thesis is consistent with the view that the function of any given biological object (for example, a human heart) could, in principle, be carried out by a non-biological object (a mechanical pump, say). Hence, I am happy to concede that my later silicon brain is a brain. But, given the just mooted essentialist thesis, it cannot be the *same brain* as my earlier human brain. Rather, the effect of all the tissue removals and bionic insertions in my skull is gradually to destroy one brain and replace it with another.

The thought-experiment *Bionic Replacement* is thus a counterexample to the brain criterion. The best description of that experiment is that I survive the operation, but my brain does not. Hence, my brain and myself have different conditions of identity, and so the brain criterion is false.

We can combine *Bionic Replacement* with *Brain Transplant* to construct a single counter-example to both the bodily criterion and the brain criterion. Imagine that following *Bionic Replacement,* my silicon brain is transplanted into a totally bionic body. Call this thought-experiment *Robot*. As in *Bionic Replacement* and *Brain Transplant,* my psychological life is perfectly preserved across all the transformations.

Given our considered judgements about *Brain Transplant* and *Bionic Replacement,* we should judge that I also survive throughout in *Robot*. We now have a single counter-example to both the bodily and brain criteria of personal identity. (A combination of *Bionic Replacement* followed by *Scattered Existence* would also yield such a counter-example.) Since a counter-example to both the bodily and brain criteria is a counter-example to any reasonable version of the animal criterion, we thereby discharge our earlier obligation to reveal the implausibility of the animal criterion of personal identity.

A related case

Our judgements about brain-identity in *Bionic Replacement* and *Robot* can usefully be compared to our judgements of artefact-identity. Consider the thought-experiment, *Ship*. Suppose that a ship is composed of 1,000 wooden planks. As it sails the seas, each plank is removed and replaced. All the changes occur gradually. However, the new planks are made of aluminium. After a few years, all the original planks have been removed and destroyed.

Is the later metal ship identical to the earlier wooden ship? Despite the continuity of the changes, most of us would be inclined to say that it is not the same ship, because the new aluminium planks are too dissimilar from the old wooden planks. Hence, the earlier ship is not identical to the later ship, even though, because of the vagueness of our concept *ship,* there is no determinate first moment of time at which the new ship came into existence.

The example *Ship* shows that the *sort* of matter or stuff with which we replace a ship's old parts can affect the overall identity of that ship, even if continuity of form and function is preserved. The same is true of biological objects. In *Robot,* my (earlier) human brain is not identical to my (later) silicon brain, and, by analogous reasoning, my (earlier) human body is not identical to my (later) bionic body.

Nonetheless, *I* (a person) survived in *Robot*. Thus, the survival conditions for persons differ from those for biological objects such as brains, or those for artefacts such as ships. The explanation of this difference will advert to the importance of psychological continuity in the analysis of personal identity over time. Such continuity is, of course, unique to mental beings such as persons, and has no application to brains or ships.

What is distinctive about persons, unlike ships or brains, is that we can survive total matter replacement even if the new matter is radically different from the old (for example, if the new matter is bionic). Since I can survive with a new brain, as in *Bionic Replacement* and *Robot,* it follows that the brain criterion fails to capture our core intuitions about personal identity over time.

A deeper worry about the brain criterion

There is a deeper worry about the stability of the brain criterion. Why did we move to the brain criterion in response to counter-examples to the bodily criterion? Was it because the human brain is a three-pound pinkish-grey spongy organ that occupies human skulls? No. We moved

to the brain criterion because of what the human brain does, viz., directly supports our mental life. It is because of its mind-supporting function that we were inclined to single out the brain as the seat of personal identity.

Given that motivation, we should not see our identity over time as tied necessarily to the continued existence of the human brain we presently have. What matters more is that our stream of mental life continue to be supported by some physical object (or, indeed, non-physical object, if that were possible), not that it continue to be supported by the very same biological organ. This conclusion is, of course, exactly in line with our dominant response to *Bionic Replacement*.

The psychological criterion

I take *Bionic Replacement* and *Robot* to undermine not just the brain criterion, but also the strong version of the psychological criterion mentioned earlier. I survive with a bionic brain, yet the resulting cause of my psychological continuity is abnormal (the continued existence of a bionic brain).[7]

It might be thought that the combined effect of these conclusions is to push us towards the weak version of the psychological criterion, according to which the identity of a person over time is traced by a line of psychological continuity, *whatever* the cause of the continuity. Indeed some readers may have wondered why we didn't appeal in the first place to a thought-experiment such as *Teletransportation* in order to establish the weak version of the psychological criterion.

In this scenario, a scanner makes a physical and psychological blueprint of me. I am then painlessly destroyed. The blueprint is transmitted to Mars where, out of different matter, a replicator creates an exact physical and psychological copy of me.

It has been claimed that, in this thought-experiment, I am identical to my replica, and hence that physical continuity is not necessary for personal identity over time. However, it would be an illusion to suppose that there is general agreement that the same person persists throughout in cases in which there is psychological continuity but no physical continuity. In fact, any consensus would appear to be in the opposite direction.

Moreover, a natural generalisation from our earlier discussion would be that cases of *Teletransportation* do not preserve personal identity, since the stream of psychological continuity is not supported by a continuously existing physical structure (such as a brain). In which

case, I would claim, I am not identical to my replica in *Teletransportation*.

In general the effect of teletransportation, applied to any object O, is the destruction of O followed by its replication. Nothing justifies saying 'O survived teletransportation' in preference to saying 'O was destroyed and a replica created'. (*Teletransportation* is thus not like the case of a watch that is dismantled and then reassembled. In the latter case, the watch persists because its distinctive parts persist, and are reassembled in the right way. This is not what happens in *Teletransportation,* even if atoms from the original are used to create the replica.)

Thus, intuitions about *Teletransportation* do not support the weak version. In addition, there are two important objections to that version of the psychological criterion.

Two objections to the weak version of the psychological criterion

One objection, due to Bernard Williams, is an objection to all versions of the psychological criterion. It purports to show that psychological continuity is not necessary for personal identity over time. I shall argue that this objection is unconvincing. The other objection (the duplication objection), is aimed only at the weak version of the psychological criterion. It is more successful.

Williams' objection

Williams imagines the following sequence of cases involving a subject, A:

(i) A is subjected to an operation which produces total amnesia;
(ii) amnesia is produced in A, and other interferences lead to certain changes in his character;
(iii) changes in A's character are produced, and at the same time certain illusory 'memory' beliefs are induced in him: they are of a quite fictitious kind and do not fit the life of any actual person;
(iv) the same as in (iii), except that both the character-traits and the memory impressions are designed to be appropriate to those of an actual person, B;
(v) the same as in (iv), except that the result is produced by putting the information into A from the brain of B, by a method which leaves B the same as he was before;

(vi) the same happens to A as in (v), but B is not left the same, since a similar operation is conducted in the reverse direction.[8]

According to Williams, in case (i) it is undeniable that A is the A-body person. He then claims that it is impossible to 'draw the line' somewhere in the sequence (i)–(vi). That is, A survives in case (i), and there is no case, such that, in that case, A is the A-body person, but in the next case, A is not the A-body person. Therefore, A is the A-body person in case (vi) and, consequently, psychological continuity is not a necessary condition of personal identity over time, contrary to all versions of the psychological criterion.

This argument is unconvincing, in large part because the sequence is under-described in important respects. To begin with, I take it that Williams' argument is not simply an instance of the Sorites Paradox. This paradox relies upon some property apparently being preserved across a large sequence of very small changes, resulting in the obviously false conclusion that all members of the sequence have the property in question. A version of the Sorites Paradox runs as follows: A man with no hairs on his head is bald; adding just one hair cannot transform a bald man into a hirsute man; so, no matter how many hairs a man has, he's always bald.

Evidently, this is not a cogent form of argument: the reasoning is paradoxical. However, the changes between adjacent cases in the sequence that Williams envisages are, presumably, too great to constitute a smooth Sorites sequence. In which case, he can avoid the charge of employing Sorites-infected reasoning. Unfortunately, just for that reason, it is far from clear that a line cannot reasonably be drawn.

For all that has been said, a line can defensibly be drawn anywhere between (i) and (vi). To see this, focus on one under-described factor in the story, viz., the extent and type of memory loss. Memories fall into at least three relevant categories: memories of past experiences (for example, remembering eating an ice-cream), factual memories (for example, remembering that Paris is the capital of France), and ability memories (for example, remembering how to ride a bike). If, for example, A loses only his experience-memories, it is plausible to suppose that A survives in case (i), especially if A's character is unaffected.

However, as the psychological changes become more drastic, it is less plausible to think that A survives. (I assume that in cases (iii)–(vi) there is meant to be *no* psychological continuity between A and the resulting A-body person.) So it would be quite defensible to hold that A fails to survive in either case (ii) or case (iii). Either way, Williams'

accusation that we cannot non-arbitrarily 'draw the line' is without foundation. The psychological criterion has not been undermined.

Although the argument based on the above sequence of cases can be evaluated in its own right, it should be noted that case (vi) is meant to be the very same as the case that Williams, earlier in his article, encourages us to describe as 'body-switching'. When case (vi) is presented on its own, we are inclined to judge that A is the B-body person; when it is presented as the terminus of the sequence (i)–(vi), we are inclined to judge that A is the A-body person. The different presentations of the same case are intended to elicit incompatible verdicts. But, since the cases are the same, one of our verdicts has to be revised.

According to Williams, it is the verdict that A is the B-body person which has to go: 'one's fears can extend to future pain whatever psychological changes precede it.'[9] That is, A is the A-body person and not the B-body person. However, given the failure of Williams' argument, and the gist of our discussion so far, we should not conclude that A is the A-body person. But nor should we conclude that A is the B-body person. As in *Teletransportation,* the psychological continuity linking A and the B-body person is not supported by its normal cause or a cause continuous with the normal cause. Rather, in case (vi), A has ceased to exist.

The duplication objection

There is a more telling objection to the weak version of the psychological criterion. The thought-experiment *Branch-Line* is a variant of *Teletransportation*. Imagine that I step into a scanner, which is modified so as not to destroy me upon replication. My psychophysical blueprint is constructed, and sent to Mars, where a replica is created. The scanner duly fails to destroy me. I step out of the scanner on Earth. Unfortunately, the operation of the scanner induces heart-failure, and though I am conscious, I have only a few days to live.

In this case, we have no hesitation in judging that I continue to exist on Earth, and therefore that the replica on Mars is not me. But both me-later and my replica stand to me-earlier in the relation of psychological continuity. If the normality or otherwise of the cause of the psychological continuity is deemed irrelevant to personal identity, as it is in the weak version of the psychological criterion, then it ought to be the case that both later candidates have an equal claim to be me. Yet, as we have seen, we believe that I am identical to the Earthly candidate, who is both physically and psychologically continuous with me. The weak version yields the wrong verdict in this case.

This objection, unlike Williams' objection, seems decisive. Consequently, the weak version of the psychological criterion cannot be correct.

Conclusion

We have seen that neither continuity of body nor brain (nor, by analogous reasoning, the continuity of any other human organ) is a necessary condition for personal identity over time. All versions of the physical criterion, and the strong version of the psychological criterion, are false.

However, we should not conclude from this that the weak version of the psychological criterion is correct. As *Branch-Line* reveals, that version does not accord with our intuitions. The best account of personal identity over time, lying as it does between the strong and weak versions of the psychological criterion, must give weight to *both* physical and psychological lines of continuity.

Given the foregoing, the most consistent and plausible view that can be recovered from our core set of common-sense judgements appears to be the following: a sufficient condition of personal identity over time is not psychological continuity with any cause, but psychological continuity with a cause that is either normal (for example, the continued existence of one's brain) *or* structurally continuous with the normal cause (for example, the gradual replacement of one's brain with a bionic brain); a necessary condition for personal identity over time is psychological continuity, similarly caused.

To claim that psychological continuity is necessary for personal identity over time is to claim that person A is identical to person B only if A and B are psychologically continuous with each other. Why accept this? The best answer to this question will advert to the *point* of our concept of a person, emphasised in Chapter 1, viz., that it delineate a relatively sophisticated kind of mental being. The thesis that personal identity over time requires psychological continuity (suitably caused) consorts well with this underlying point.

There is no reason why we cannot combine the above two conditions into a single necessary and sufficient condition (the intermediate criterion): person A at one time is identical to person B at some later time if and only if A stands to B in the relation of psychological continuity with a cause that is either normal or structurally continuous with the normal cause. This intermediate criterion of personal identity over time is intermediate between the strong and weak readings of the psychological criterion. However, it

gives more weight to psychological continuity than to physical continuity. Personal identity is explicated in terms of psychological continuity, caused in a certain way.

Notice also that, as stated, the intermediate criterion, like the other criteria criticised in this chapter, is strictly circular: persons A and B are referred to on the right-hand side of the criterion, and the sortal concept *person* 'contains' the criteria of identity over time for persons. Such circularity is not objectionable. The intermediate criterion clearly imposes constraints on what it is to be the same person over time. This criterion, after all, conflicts with the other criteria discussed above. The question of whether we can ultimately eliminate reference to persons on the right-hand side of the criterion (and replace such reference with, for example, reference to bodies or 'bundles' of experiences) is just the question of whether some version of reductionism is true. And, as we have seen, the answer to that question is complicated.

If the intermediate criterion is true, then my identity conditions are not those of a human body or brain. Hence I (here and now) cannot be strictly identical to a body or brain. This conclusion undermines the standard or orthodox materialist answers to the nature question (what is a person?). These answers require that a person be identified with some biological entity. However, as we have seen, although a person is made entirely of matter, he is numerically distinct from his body or brain.[10]

A complication

Our discussion thus far has made a certain simplification. The thought-experiment *Branch-Line* exploited the fact that relations of psychological continuity can hold between one person and two later people. In that thought-experiment, only one of the streams of consciousness had its normal cause (that is, the continued existence of my brain). However, we can imagine a case in which a person at one time is psychologically continuous with two later persons, where *both* streams of psychological continuity have what can be deemed their normal cause, that is, the continued existence of each of the brain hemispheres. This is what happens in *Fission*.

Since one person cannot be identical to two distinct people, the sufficient condition for personal identity endorsed above will have to be modified, unless either such branching is impossible or the possibility of branching can be redescribed so that it does not conflict with our sufficient condition. The problems raised by *Fission* are the focus of the next chapter.

4

FISSION

The importance of *Fission*

The thought-experiment *Fission* has been much discussed in recent years, and it is of central importance to a number of issues in personal identity. It shows, as we shall see, that the identity of a person can be extrinsically grounded. That is, the identity of a person may be fixed by the existence of another, causally unrelated, person. This consequence has implications for the form which any adequate criterion of personal identity must take. Further, as we shall see in Chapter 6, *Fission* features in one of the central arguments for the thesis that personal identity is not 'what matters'.

In fission, one thing splits into two or more things of the same kind. Such processes do occur in nature (for example, amoebae). Fission of persons, of course, does not occur—but it might. We can describe a thought-experiment to flesh out this possibility, presented in the first-person.

Like all humans, my mental life crucially depends upon the normal functioning of my brain. Suppose, however, that I have a property which most people don't have, but might have done: each of my brain hemispheres supports the very same mental functions. If one of my hemispheres developed a tumour, that hemisphere could simply be removed and my mental life would continue otherwise unaffected, supported by the remaining hemisphere.

Suppose that my body develops cancer. The surgeons cannot save my body, but they can do the following. They can remove my brain, and transplant both hemispheres into two brainless bodies, cloned from my body many years ago. This operation is successfully carried out. We now have two people—call them Lefty and Righty—both of whom are psychologically continuous with me (same character, beliefs, apparent memories of a shared past, etc.). They are also physically similar to me,

and each contains a hemisphere from my brain. Thus there is both physical and psychological continuity linking me with Lefty and with Righty. Suppose also that Lefty and Righty are in different rooms in the hospital, and exercise no causal influence on each other. They never meet, and each is unaware of the existence of the other.

How should this case be described—who is who?

Six responses to *Fission*

(1) *The case is not really possible, so we can say nothing about it and learn nothing from it*

This view is implausible. Hemisphere transplants may be technologically impossible, but they are not logically impossible. Hemisphere transplants, like other organ transplants, are surely nomologically possible (that is, consistent with the laws of nature). To imagine a brain or hemisphere transplant is not to imagine something counter-nomic, such as travelling faster than the speed of light. And since such transplants are nomologically possible, they are also logically possible.

Some animalists (see Chapter 2) would advocate response (1). Their argument would proceed as follows. It's true that fission of persons is not a priori or conceptually impossible. But it may be a posteriori impossible. If animalism is true, then certain laws of evolutionary development—which exclude the possibility of fission—may be a posteriori essential to persons, in much the way that, for example, having atomic number seventy-nine is now thought to be a posteriori essential to gold. The fission of persons will then be 'deeply' or metaphysically impossible.

However, there are three replies to this animalist argument. First, as argued in Chapters 2 and 3, the animalist's theory of persons is unmotivated and open to plausible counter-examples. Second, even if the animalist's theory were correct, there is the following disanalogy between the natural kinds *gold* and *human being* which casts doubt on the just mooted line of reasoning. The laws of development governing members of biological natural kinds such as human beings do not play a role analogous to that of the atomic properties of a non-biological natural kind such as gold. How an organism develops depends, *inter alia,* upon its environment, and radical changes in the environment may produce genetic changes over generations which allow future persons to divide like amoebae. So, even if persons were essentially biological, amoebae-like fission need not be impossible.

Third, as noted above, it does seem that fission by hemisphere transplant is nomologically possible for present-day human beings. Even if amoeba-style splitting is impossible for all known persons, fission by surgical transplant is not. It is only technically impossible. For these reasons, response (1) is not a serious contender.

(2) *I survive the operation, and I am one or the other of Lefty or Righty*

Immediately after fission, Lefty and Righty are physically and psychologically indistinguishable. Both of them stand to me in the same physical and psychological relations. They both believe that they are me. According to response (2), one is right and the other wrong.

Response (2) is implausible for two reasons. First, since Lefty and Righty are symmetrically related to me in respect of physical and psychological continuities, the claim that, for example, I am Lefty, can only be sustained on something like the Cartesian view of persons.[1] If we think of a person as an immaterial ego that typically underlies a stream of psychological life, we can suppose that my ego pops into, for example, the left-hand stream of consciousness, leaving the right-hand stream ego-less or even with a new ego. As noted in Chapter 1, this view of persons is bizarre.

Second, the metaphysical implausibility of the Cartesian view has an epistemic counterpart. According to response (2), when I divide, I survive in one of the two streams. So either I am Lefty or I am Righty. But how can we know which? From the third-person point of view, we have no reason to make one identification rather than the other. Nor is appeal to the first-person perspective of any help: both Lefty and Righty take themselves to be me. Nothing in either stream of consciousness will reveal to its bearer that he is me. So if I am Lefty, this truth will be absolutely unknowable. There may be no incoherence in the idea of unknowable truths, but we should nonetheless be suspicious of any theory of personal identity which implies that truths about who is who can be, in principle, unknowable. For these reasons, we should reject response (2).

(3) *I survive fission as both Lefty and Righty*

There are three ways in which we can understand this response. According to the first way, I am identical to both Lefty and Righty (hence, Lefty is Righty). According to the second way, Lefty and Righty are sub-personal constituents of a single person. According to

the third way, Lefty and Righty are persons who together compose me (so that two persons are parts of one larger person, just as two countries can be parts of one larger country).

These views are hard to believe. It seems uncontroversial that Lefty and Righty are numerically distinct, and that they are persons (not sub-personal entities). Lefty and Righty both satisfy the normal physical and psychological criteria for personhood. They qualify as persons. And they are two. They may be exactly alike immediately after fission, but exact similarity does not imply numerical identity. (Two red billiard balls may be exactly similar, yet numerically distinct.) Further, they will soon begin to differ, mentally and physically, so that it would be intolerable to regard them as anything but distinct persons. These considerations undermine the first two ways of understanding response (3).

According to the remaining version of response (3), I exist after fission composed of Lefty and Righty, now regarded as persons in their own right. This is hard to understand. The postulation of my existence in this circumstance (in addition to that of Lefty and Righty) does no work whatsoever. It is completely idle.

Further, can we really make any sense of the idea that one person might be composed of two separate persons? How could one person be composed of two bodies and two minds? Yet, supposedly, after fission I am permanently composed of two unconnected, self-contained spheres of consciousness. How could they possibly constitute a single person? If Lefty believes that Gore will win the next election, and Righty believes that he won't, do I believe that Gore will both win and lose the next election? Such problems multiply. It seems that all ways of understanding response (3) skewer our concept of a person. Response (3) cannot be an adequate description of fission.

(4) *The case of* Fission *has been misdescribed. Lefty and Righty exist prior to fission, but only become spatially separate after fission*

This 'multiple occupancy' theory also has different versions. Some philosophers think that only Lefty and Righty occupy the pre-fission body. Others think that three people (Lefty, Righty, and me) occupy the pre-fission body, but that only Lefty and Righty survive fission. The differences between these versions of the theory will not concern us.

One motivation for the 'multiple occupancy' account is to reconcile two plausible and apparently incompatible theses: the thesis that what

matters in personal survival over time is psychological continuity and connectedness, and the thesis that what matters is identity.[2] On the standard ('single occupancy') account of personal fission, these theses are in tension. Derek Parfit, famously, argued that we should relinquish the second thesis.[3] The 'multiple occupancy' account allows us to retain both theses consistently.

Despite its reasonable-seeming motivation, this account is hard to believe. It involves a tremendous distortion of our concept of a person to suppose that more than one person occupies the pre-fission body. Surely to one body and a unified mind, there corresponds only one person? However, the strangeness of response (4) may depend, in part, on one's general metaphysics. In particular, the degree of strangeness may depend on whether we accept a three-dimensional or four-dimensional view of continuants such as persons.

On the three-dimensional view, persons are 'wholly present' at all times at which they exist (much as a universal, such as redness, is said to be 'wholly present' in each of its instantiations). On this view, persons are extended only in space, not in time, and have no temporal parts. On the four-dimensional view, persons are four-dimensional entities spread out in space and time. Persons have temporal parts as well as spatial ones. Hence, at any given time, say 1993, only a part of me is in existence, just as only a part of me exists in the spatial region presently demarcated by my right foot.

On the three-dimensional view of persons, response (4) is not just strange but barely intelligible. Consider a time just prior to fission. On this view, two 'wholly present' persons (entities of the same kind) occupy exactly the same space at the same time. This ought to be as hard to understand as the claim that there are two instantiations of redness in some uniformly coloured red billiard ball.

On the four-dimensional view, however, Lefty and Righty are distinct persons who, prior to fission, share a common temporal part. It ought to be no more remarkable for two persons to share a common temporal part than for two persons (such as Siamese twins) to share a common spatial part. And, on this view, in contrast to the final two responses to be considered below, the existence of each of Lefty and Righty is independent of the existence of the other. Hence, for example, provided that the left-hemisphere transplant is successful, Lefty exists whether or not Righty also exists.

However, the four-dimensional view is open to objection. For example, Mozart died when he was 35, but he could have lived longer or died earlier. But if Mozart is identical with some four-dimensional

object, then Mozart could not have lived longer or died earlier, since no four-dimensional object (or mereological sum) could have had a temporal extent different from its actual extent. Given that Mozart has a property, not possessed by any four-dimensional object, it follows that Mozart is not a four-dimensional object. This conclusion obviously generalises to other people and other kinds of object.[4]

Second, the original objection still holds: response (4) is counter-intuitive. It is implausible to hold that two persons (Lefty and Righty) share a common temporal segment in the absence of any psychological disunity. We should be loath to give up the principle that to each psychologically unified temporal segment there corresponds just one person.[5] Third, it is a consequence of response (4) that whether two people now occupy a particular body depends upon whether that body undergoes a successful brain-division at any time in the future. Such dependency is odd.

Third, there is the problem of how we are to account for the apparent coherence and unity of the 'I' -thoughts associated with the locus of reflective mental life that occupies the pre-fission body. How can there be such unity if two persons occupy that body?

These objections to response (4) show that the 'multiple occupancy' view is problematic, and we should avoid it if we can.

(5) *When I divide into Lefty and Righty, I cease to exist. Lefty and Righty then come into existence, and are numerically distinct, though initially very similar, persons*

This is the response that will be defended in this chapter. When I divide, there are two equally good candidates for identity with me. Since they are equally good, and since one thing cannot be two things, I am identical to neither. And since there is no one else with whom I could plausibly be identified, I no longer exist after fission. This response respects the logic of identity, and does not violate our concept of a person by supposing either that two persons compose one large, scattered post-fission person or that more than one person occupies the pre-fission body.

However, it is important to realise that, in embracing response (5), we are committing ourselves to a quite particular conception of the identity over time of persons. On this view, for example, I am not Lefty. Why is this true? The reason is not: because Lefty and I do not have the same body, or because Lefty and I do not have the same whole brain. (These would anyway be bad reasons—see Chapter 3.) The reason is that one thing cannot be two.

Whether I continue to exist depends upon whether I have one continuer or two. If only one continuer survives, I survive; if both survive, I do not survive. Since Lefty and Righty are causally isolated from each other, this implies that the identity of a person can be determined by extrinsic factors. That is, paradoxical though it may seem, Lefty can truly say: 'Thank goodness the right hemisphere was not destroyed, otherwise *I* wouldn't have existed'. Lefty's existence depends upon extrinsic factors.

Theories that allow for the extrinsicness of existence-dependence are sometimes called 'best candidate' theories of personal identity. According to these theories, B at t_2 is the same person as A at t_1 only if there is no better or equally good candidate at t_2 for identity with A at t1. If there are two equally good candidates, neither is A. We will look in more detail at such theories below.

(6) *It is vague or indeterminate whether I am Lefty and vague or indeterminate whether I am Righty. There is simply no fact of the matter as to who I am after fission*

This is a possible response to *Fission*. Unfortunately, it has four flaws. (i) It lacks any motivation. (ii) It is counter-intuitive. (iii) It is simply a non-standard version of the best candidate theory. (iv) It fails to apply to scenarios which don't involve persons, but which are relevantly similar to *Fission*. In particular, response (6) fails to apply to the story of the *Ship of Theseus*.

(i) The lack of motivation can be brought out in the following way. It is determinate that Lefty, Righty and myself are all persons. It is perfectly determinate which relations of physical and psychological continuity we stand in to each other. Where is the logical space for indeterminacy? Cases of alleged indeterminacy in identity over time typically arise when something is missing or diminished (as in the thought-experiment *Indeterminacy*). In *Fission,* everything is present, twice-over. There is no room for indeterminacy.

(ii) The response is counter-intuitive. I exist prior to fission; Lefty and Righty exist after fission. If it is indeterminate whether I am Lefty, then it is indeterminate whether Lefty exists prior to fission. The same is true of Righty. In which case, it is indeterminate how many persons exist prior to fission. But, as our discussion of response (4) brought out, our common-sense intuition is that there is, determinately, one and only one person who occupies the pre-fission body.

(iii) The present response holds that it is indeterminate whether I am Lefty, and indeterminate whether I am Righty. That is, it is neither true

that I am Lefty, nor true that I am Righty. The reason for the indeterminacy must lie in some feature of the duplication (though as noted above, it is hard to see how duplication could induce indeterminacy).

However, if Lefty had not existed, I would survive in the right-hand branch. There would then be no indeterminacy in my identity with the sole survivor. This implies that whether or not it is true that I survive depends upon whether I have one off-shoot or two. Thus response (6) is simply an unorthodox or non-standard version of the best candidate theory (response (5)). In which case, why not simply opt for the standard version, and avoid the problems with the present response?

(iv) The story of the *Ship of Theseus* runs as follows. The wooden ship of Theseus regularly sails the seas, and is in constant need of repair. Over time, planks are removed and replaced. In fact, the repairs are so extensive that, after a number of years, none of the original planks remains. Call the resulting ship 'the continuously repaired ship'. Suppose that the discarded planks are retained and used to build another ship, exactly similar to Theseus' ship. Call this 'the re-constituted ship'.

The classical source for the story can be found in Plutarch:

> The vessel in which Theseus sailed and returned safe with these young men went with thirty oars. It was preserved by the Athenians up to the times of Demetrius Phalerus; being so refitted and newly fashioned with strong plank, that it afforded an example to the philosophers in their disputations concerning the identity of things that are changed by addition, some contending that it was the same, and others that it was not.[6]

The question that Plutarch took the philosophers to dispute was: Can a ship survive total replacement of its parts? This is a good question, even if the best answer to it is affirmative (provided that the new planks are of the same type as the old planks).[7]

However, Plutarch's question involved no reference to the ship built from the discarded planks—the re-constituted ship. It is the presence of this latter ship which has been thought to create the deeper puzzle. Thomas Hobbes, the modern source of the puzzle, refers to both ships in his account:

> if…that ship of Theseus, concerning the difference whereof made by continued reparation in taking out the

> old planks and putting in new, the sophisters of Athens were wont to dispute, were, after all the planks were changed, the same numerical ship it was at the beginning; and if some man had kept the old planks as they were taken out, and by putting them afterwards together in the same order, had again made a ship out of them, this, without doubt, had also been the same numerical ship with that which was at the beginning; and so there would have been two ships numerically the same, which is absurd.[8]

Although Hobbes refers to the re-constituted ship in his account, it has seemed to most modern commentators that the central issue is not how we are to avoid the conclusion that both later ships are identical to Theseus' ship. That conclusion is not one that we are intuitively forced to draw, as the following diagnosis makes clear.

The dominant reaction to the *Ship of Theseus* is not that Theseus' ship is identical to both later ships. It is that Theseus' ship is the continuously repaired ship and not the re-constituted ship. Despite the total replacement of parts, the former ship has the best title to be deemed the ship of Theseus. The explanation of this is that in the case of artefacts such as ships, we operate with two criteria of identity—the continuity-under-a-sortal criterion and the identity-of-original-parts criterion—and the former outweighs the latter.

According to the continuity-under-a-sortal criterion, A (at t_1) is the same ship as B (at t_2) if we can trace a continuous path, under the sortal 'ship', through space and time from A to B. Any exchange of planks, if they are to preserve ship-identity, must occur in the normal working life of the ship. We should also require that the new planks are of roughly the same size and material type as the originals. (We can assume that these and other relevant qualifications are all met in the *Ship of Theseus*.) According to the simpler identity-of-original parts criterion, A (at t_1) is the same ship as B (at t_2) if they are made of the very same planks.

In any scenario in which only one of the criteria is applicable, that criterion is a sufficient condition for ship-identity over time. But, as our ordinary judgements make clear, when both criteria are applicable (as in the *Ship of Theseus*), then the continuity-under-a-sortal criterion outweighs the identity-of-original-parts criterion. The former criterion is dominant with respect to the latter. Put differently: in our scenario, there are two candidates for identity with the ship of Theseus, and the continuously repaired ship has the better claim to be the ship of Theseus.

However, now consider the following counterfactual: had the removed planks not been replaced, the ship then re-constituted would have been the ship of Theseus. This counterfactual seems hard to deny. For, in the situation described in the antecedent, all that happens, in effect, is that a ship is dismantled, transported, and then reassembled. This sequence of events is one that generally preserves the survival of a single ship. It is not relevantly different from the familiar case of a tent that is frequently dismantled and reassembled, or of a watch that is dismantled, repaired, and reassembled. It may be unclear (at least in the case of the watch) whether we should say that the watch still exists while dismantled, but what is not in doubt is that the earlier watch is the later watch. The same is true of the ship of Theseus in the just imagined scenario in which none of the removed planks is replaced.

If we combine our dominant reaction to the Theseus story with adherence to the above counterfactual, we are committed (like response (5)) to the extrinsicness of existence-dependency. We acknowledge that Theseus' ship is the continuously repaired ship and not the re-constituted ship (call the latter 'RC1'). We accept that, had the continuously repaired ship not been replenished, the ship then reconstituted (call it 'RC2') would have been the ship of Theseus. But *that* ship (RC2) is not the same ship as the re-constituted ship in the original story (RC1). Since the ship of Theseus is identical to RC2, and distinct from RC1, RC1 cannot be identical to RC2.[9] Since RC1 is not RC2, we can truly say the following: 'Had this ship (pointing to the continuously repaired ship) not been replenished, that very ship (pointing to RC1) wouldn't have existed.'

Thus, in the *Ship of Theseus,* as in *Fission,* we are committed to the extrinsicness of existence-dependence. But the indeterminacy response cannot apply since, in the original scenario, the ship of Theseus is definitely the continuously repaired ship and definitely not the re-constituted ship. There is no room for the indeterminacy response.[10] Yet given the structural similarities between *Fission* and *Ship of Theseus,* it would be good if one response could cover them both. This is precisely what is offered by response (5).

The best candidate theory of personal identity

Are best candidate theories, and hence response (5), ultimately acceptable? Don't best candidate theories violate the widely accepted semantic thesis that certain identity sentences are, if true, necessarily true and, if false, necessarily false? Isn't the upshot of response (5) precisely that I am not Lefty, but that had Righty not existed (for

example, had the surgeon accidentally dropped the right hemisphere), I would have been Lefty?

No. The widely accepted thesis is that identity sentences *containing only 'rigid' singular terms* (that is, terms which do not shift their reference across possible worlds) are, if true, necessarily true and, if false, necessarily false. This use of the term 'rigid' is due to Saul Kripke.[11] Kripke contrasted proper names and natural kind terms with non-rigid terms, such as familiar uses of definite descriptions. Typical uses of, for example, the definite description 'the tallest man in the world' are non-rigid. In this world it may pick out Smith, but in other possible worlds it picks out Jones. Its reference shifts across possible worlds. Kripke's intuition, now widely shared, is that proper names, pronouns, demonstratives, and natural kind terms do not shift their reference across possible worlds. They are rigid designators.

We can read the term 'Lefty' as rigid or as non-rigid.

(i) If it is non-rigid (perhaps abbreviating the definite description 'the person who happens to occupy the left-hand branch'), then it is true that, had Righty not existed, I would have been Lefty. So the sentence 'I am Lefty' is contingent. It is false in the fission world, and true in nearby worlds in which Righty doesn't exist. But since 'Lefty' is non-rigid, this contingency is consistent with the necessity of identity sentences containing only rigid singular terms. Thus understood, the contingency of 'I am Lefty' ought to be no more worrying than the contingency of 'I am my father's only son'.

(ii) If 'Lefty' is rigid, then the best candidate theorist, if he is to respect the necessity of identity sentences containing only rigid singular terms, must deny that I would have been Lefty if Righty had not existed. If Righty had not existed, I would then have occupied the left-hand branch, but that person (namely, me) is not Lefty. Lefty doesn't exist in the nearest world in which Righty doesn't exist, though an exact duplicate of Lefty—Twin Lefty—exists there.

Thus best candidate theories do not violate the necessity of identity sentences which contain only rigid terms. However, they do have consequences that might be thought objectionable. Consider again the world in which I divide into Lefty and Righty. According to the best candidate theory, Lefty can truly say 'Thank goodness Righty exists, otherwise *I* wouldn't have existed.' Given that Lefty and Righty exert no causal influence on each other, such dependence is apt to seem mysterious. Clearly, counterfactual dependence is not mysterious where there are appropriate causal connections. For example, it is not paradoxical that I would not have existed if my mother had not existed. But in the present case causal connection is precisely what is missing.

However, these consequences are not objectionable. They simply illustrate the fact that properties like *being occupied by Lefty* (where 'Lefty' is understood to be rigid) are extrinsic properties of bodies. That is, whether the left-hand body has the property of being occupied by Lefty, rather than by Twin Lefty, is fixed by an extrinsic factor (viz., the existence or non-existence of Righty). But this is not counter-intuitive. The property *being occupied by Lefty* is not a causal property of a body. In contrast with properties of shape and weight, etc., this identity-involving property does not contribute to the causal powers of any body in which it inheres. (The causal powers of the left-hand body are unaffected by whether Lefty or Twin Lefty is its occupant.)

It is typical of a non-causal property that its possession by an object may depend upon what happens to other objects which exercise no causal influence on it. For example, the property of being a war widow is not a causal property and, unsurprisingly, whether a woman is a war widow typically depends upon what happens to someone who, at the relevant time, exercises no causal influence on her. Response (5) teaches us that identity-involving properties (like *being occupied by Lefty*) are also extrinsic. This is not a counterexample, merely a consequence.

Some comments on the best candidate theory

Three points about the best candidate theory need to be emphasised. First, as noted above, our preferred description of *Fission* does not imply any violation of the semantic thesis that an identity sentence containing only rigid designators is, if true, necessarily true and, if false, necessarily false. It is precisely because we respect these theses, that we acknowledge the consequence that Lefty is not Twin Lefty.[12] Further, the fact that Lefty is not identical to Twin Lefty, though they occupy the very same body, confirms the conclusion reached in earlier chapters that a person and his body are numerically distinct.

Second, and related, it is not a consequence of our account that the relation of identity is extrinsic. The extrinsic determination of the truth-value of certain identity sentences no more implies the extrinsicness of identity than the existence of contingent identity sentences (such as 'Smith is the tallest man in the world') implies that identity is contingent.

The analogy is worth pursuing. The thesis of the necessity of identity is the thesis that $(\forall x)(\forall y)(x=y \rightarrow ?\ (x=y))$. This is a thesis in metaphysics, and has nothing to do with natural language identity sentences or singular terms. Even if natural languages contained no rigid terms, the thesis would still be true. It so happens that the

metaphysical thesis has a semantic corollary, but that is accidental. Similarly, the fact that the truth-value of certain identity sentences can be fixed by extrinsic factors does not imply that identity itself is extrinsic. To suppose that the best description of *Fission* and related scenarios shows identity to be extrinsic is to confuse facts about language (sentences, names, descriptions, etc.,) with facts about ontology (in this case, the relation of identity).

The following thesis captures the intrinsicness of identity: (∀x) (∀y) (whether x is y does not turn on any fact concerning anything other than x or y). This thesis is *not* violated by best candidate theories. In particular, the fact that I am not Lefty does not depend on any contingent fact concerning objects other than me and Lefty. If it did, the fact that I am Lefty (where 'Lefty' is rigid) would be a contingent fact, which it is not. What is true is that Lefty's existence depends (in part) on events which exercise no causal influence on the events that constitute the career of Lefty, and the truth-value of the sentence 'I am the person occupying the left-hand branch' is determined by whether or not Righty exists. Neither of these results implies that the identity relation itself is extrinsic.

Third, we should therefore properly describe the best candidate theory, not as committed to the extrinsicness of identity, but as committed to the *extrinsicness of existence-dependence*. Whether Lefty exists depends upon whether Righty exists, where Lefty and Righty exert no causal influence on each other.

The lesson of *Fission*

The best candidate theory provides the most satisfying response to the case of fission. It also reveals something important about our concept of a person. It shows that our concept of a person is the concept of an entity whose existence conditions can be determined by extrinsic factors. This result may be surprising, but it is not objectionable.

If we combine this result with the criterion of personal identity endorsed in Chapter 3, we arrive at the full formulation of our preferred intermediate criterion of personal identity over time. Person A at an earlier time is identical to person B at some later time if and only if A stands to B in the relation of psychological continuity with a cause that is either normal or structurally continuous with the normal cause, *and* there is no better or equally good candidate at the later time for identity with A. The thought-experiment *Fission* revealed the need for this second conjunct. The implications of this thought-experiment for value theory will be discussed in the Chapter 6.

5

IDENTITY AND VAGUENESS

The commitment to vagueness

According to the intermediate criterion of personal identity over time endorsed at the end of the previous chapter, a person at an earlier time is identical to a person at some later time if and only if the earlier person stands to the later person in the relation of psychological continuity with a cause that is either normal or structurally continuous with the normal cause, *and* there is no better or equally good candidate at the later time for identity with the earlier person.

A consequence of this criterion of personal identity, and indeed of any criterion that understands personal identity, *inter alia,* in terms of relations of physical and/or psychological continuity, is that it can be *vague* or *indeterminate* whether a person at one time is identical with a person at some later time.

What is vagueness?

What do the terms 'vague' and 'indeterminate' mean? English contains many vague terms: for example, predicates like 'bald', 'red', and 'heap', and quantifiers like 'many'. On one standard view, to say that 'bald' is vague is to say that the term has no sharp boundaries. That is, a man who is bald cannot cease to be bald by the addition of a single hair.

Further, because the predicate 'bald' lacks sharp boundaries, the predicate will have a 'grey' area of application sandwiched between cases in which the predicate clearly applies and cases where it clearly fails to apply. That is, 'bald' will admit of a large number of borderline cases, in which the predicate neither definitely applies nor definitely fails to apply. On the standard view, if Fred's pate is a borderline case

of baldness, then the sentence 'Fred is bald' will be indeterminate in truth-value (that is, neither true nor false).[1]

Can indeterminate sentences include not just subject-predicate sentences like 'Fred is bald', but also identity sentences, including sentences of personal identity? If there can be such vague identity sentences, two questions come to the fore. First, is the source of such vagueness linguistic/conceptual (as many believe it is with 'bald')? Second, what is the connection between the notion of vague identity and the suggestion that the world itself might be vague?

We certainly have good grounds for holding that there are vague sentences of personal identity. Consider the thought-experiment, *Indeterminacy*. An alteration machine selectively re-arranges my brain-matter, the result of which is that 50 per cent of my mental states (memories, beliefs, desires, character traits, etc.) are extirpated and replaced with new mental states.

In this thought-experiment, not enough mental connections are retained in order to justify us saying that I am psychologically continuous with the resulting person. But nor are so few retained to justify us saying that I am not psychologically continuous with the later person. It is vague or indeterminate whether I am psychologically continuous with the later person. There is simply no fact of the matter.

In such a case, it is indeterminate whether I am psychologically continuous with the later person, the cause of the psychological continuity is normal (it is carried by the—albeit interfered with—brain), and there is no other candidate for identity with me. According to our preferred criterion of personal identity over time, therefore, the indeterminacy in psychological continuity implies that it is indeterminate whether I am identical to the later person. In response to the question 'Am I the later person?', we should simply shrug our shoulders. The sentence 'I am the later person' is indeterminate in truth-value.

It might be thought that the indeterminacy of the sentence 'I am the later person' conflicts with what many have taken to be a quite general proof to the contrary first presented in a much discussed one-page article by Gareth Evans.[2] However, in this chapter I shall argue that any tension with Evans' argument, properly understood, is illusory. The purpose of Evans' argument is not to establish the impossibility of vague identity sentences or the statements they express.

Evans' article is in two parts. Its first paragraph attempts to gloss the idea of vague objects. Its second and third paragraphs present a *reductio* proof which purports to show that the identity relation always either determinately obtains or determinately fails to obtain.

Evans intended his proof to undermine the coherence of the idea that the world might contain vague objects. Although I will discuss the connection between vague identity and vague objects, my central focus in this chapter will be on Evans' proof of the determinacy of identity.

Evans' proof

Evans wrote:

> It is sometimes said that the world might itself *be* vague. Rather than vagueness being a deficiency in our mode of describing the world, it would then be a necessary feature of any true description of it. It is also said that amongst the statements which may not have a determinate truth value as a result of their vagueness are identity statements. Combining these two views we would arrive at the idea that the world might contain certain objects about which it is a *fact* that they have fuzzy boundaries. But is this idea coherent?
>
> Let 'a' and 'b' be singular terms such that the sentence 'a =b' is of indeterminate truth value, and let us allow for the expression of the idea of indeterminacy by the sentential operator '∇'.
>
> Then we have:
>
> (1) $\nabla(a=b)$.
>
> (1) reports a fact about b which we may express by ascribing to it the property '$x[\nabla(x=a)]$':
>
> (2) $x[\nabla(x=a)]b$.
>
> But we have:
>
> (3) $\sim\nabla(a=a)$
>
> and hence:
>
> (4) $\sim x[\nabla(x=a)]a$.
>
> But by Leibniz's Law, we may derive from (2) and (4):
>
> (5) $\sim(a=b)$
>
> contradicting the assumption, with which we began, that the identity statement 'a=b' is of indeterminate truth value.

If 'Indefinitely' and its dual 'Definitely' ('Δ') generate a modal logic as strong as S5, (1)–(4) and, presumably, Leibniz's Law, may each be strengthened with a 'Definitely' prefix, enabling us to derive

(5') Δ~(a=b)

which is straightforwardly inconsistent with (1).[3]

Evans' proof examined

Evans begins by contrasting two views of vagueness: the view that the world itself might be vague and the view that vagueness is 'a deficiency in our mode of describing the world'. His next thought, presumably, is that there are vague objects ('objects about which it is a *fact* that they have fuzzy boundaries') only if there are vague identity statements, the vagueness of which is not due to any vagueness, or referential indeterminacy, in the relevant singular terms (that is, not due to a 'deficiency' in our mode of description). The point of the proof is to show that there cannot be such statements.

Evans' proof, thus construed, is consistent with the undeniable fact that there can be vague identity sentences. Here is an example. Imagine that a series of pens is arranged in such a way that the first is red and the last is orange, and that adjacent pens match imperceptibly in colour. According to one standard view of vagueness, in such a smooth sequence, there is no first orange pen, and the definite description 'the first orange pen' has no determinate reference. (That is, it is vague which pen it picks out; not: it picks out something vague.) Thus, if I own a particular pen in the reddish-orange region and say 'my pen is the first orange pen', the statement expressed by that identity sentence is vague, yet the vagueness is due to the referential indeterminacy of the definite description, 'the first orange pen'.[4]

Evans' proof is consistent with the vagueness of 'my pen is the first orange pen'. What Evans intended to show is that there cannot be a vague identity sentence, '∇(A=B)', which implies '(∃x)(∃y) ∇(x=y)'. This would be a pure case of indeterminacy of identity. Our example of a vague identity sentence ('my pen is the first orange pen') does not imply '(∃x)(∃y) ∇(x=y)'. The singular term 'the first orange pen' does not determinately single out some pen which is such that it's vague whether it is my pen. Analogously, it would be wrong to think that the truth of 'Bill is the world's tallest man, but he might not have been' implies '(∃x)(∃y)((x=y) & ◊~(x=y))'. The necessity of identity is

consistent with the contingency of identity sentences such as 'Bill is the world's tallest man'. Similarly, the determinacy of identity is consistent with the indeterminacy of 'my pen is the first orange pen'.

Evans' proof

The structure of the proof is clear: (1) entails (2), and (3) entails (4); (2) and (4), by Leibniz's Law, entail (5); (1)–(5) can all be strengthened with a 'Definitely' prefix ('Δ'), yielding (5'), which is 'straightforwardly inconsistent' with (1).

As the pen example makes evident, some restrictions are needed on the singular terms that can figure in the proof. We should take 'a' and 'b' to be constants—denoting terms of logic that have no descriptive content. However, the following question now comes to the fore. Why did Evans rely on the move from (1) to (2), rather than simply take (2) to be the premise for *reductio*? This is a good question for two reasons.

First, analogous moves elsewhere are often thought invalid. For example, the move from the *de dicto* (i) 'John believes that the tallest spy is a spy' to the *de re* (ii) 'the tallest spy is such that John believes him to be a spy' is invalid. Sentence (i) may be true simply because John, from his armchair, assumes (rightly, let's suppose) that there is a tallest spy, and he is aware of the truism that the tallest spy is a spy. But sentence (ii) will be false if John is like most of us, since its truth requires John to be 'acquainted' with the man who is in fact the tallest spy. The move from (i) to (ii) is fallacious. The transition from (1) to (2) may similarly be invalid.

Second, and more important, it is (2) rather than (1) which properly captures the idea that identity (that extra-linguistic item) might fail to hold determinately. According to (2), a given object (*b*) is such that it's vague whether it is *a.* It is vague whether the relation of identity holds between *a* and *b*. Hence, if we cannot validly reach (2), or if (2) is not coherent, then Evans will have won at the outset. There would then be no stable position for him to argue against.

For these reasons, we can re-state Evans' proof as follows:

[1] x[∇(x=a)]b	(Supp.)
[2] ~x[∇(x=a)]a	(Truism)
[3] ~(a=b)	([1], [2], LL)
[4] Δ~(a=b)	(Strengthened [1] and [2], LL)
[5] ~x[∇(x=a)]b	([4], contradicting [1])

We can now ask the following three questions: (i) Does 'x[∇(x= a)]' denote a property? (ii) Is the proof valid? (iii) Is [2] really truistic?

Question (i)

It's obviously essential to the validity of Evans' proof that 'x[∇(x= a)]' is not analogous to, for example, '——is so-called because of his size'.[5] The argument:

Giorgione was so-called because of his size;
Barbarelli was not so-called because of his size; so

Giorgione is not Barbarelli

is famously invalid. On one view, the invalidity of this inference is linked to the fact that the predicate '——is so-called because of his size' does not denote a genuine property of Giorgione. The predicate fails to denote a genuine property of Giorgione because whether it can be truly ascribed depends on *how* we refer to its intended object (for example, whether as 'Giorgione' or as 'Barbarelli').[6]

Is 'x[∇(x=a)]' analogous to '——is so-called because of his size'? If it is, then it doesn't denote a genuine property, and for that reason we should not believe in the possibility of vague identity. (Evans wins by default.) To keep the debate going, let's assume that it does denote a property. We can now proceed to questions (ii) and (iii).

Question (ii)

There are two places at which the validity of the proof might be questioned—the step from [1] and [2] to [3], and the step from [3] to [4].

(a) The step from [1] and [2] to [3]

It might be thought that the predicate 'x[∇(x=a)]' denotes *different* properties in [1] and [2]. That is, the reference of the predicate shifts when appended to a different subject-term. In contrast to the no-denotation view described above, it's possible to take such a view of the predicate '——is so-called because of his size'. On this second view, the predicate stands for the property *being called 'Giorgione' because of his size* when attached to 'Giorgione', and the property *being called 'Barbarelli' because of his size* when attached to

'Barbarelli'. This ambiguity explains the invalidity of the earlier argument.

However, there is no reason to think that any such reference-shift occurs in the move from [1] to [2]. The reference of 'x[∇(x=a)]' is surely not determined by the terms to which it is attached.

(b) The step from [3] to [4]

The step from [3] to [4] is more problematic. In the final paragraph, Evans writes: 'If "Indefinitely" and its dual, "Definitely" ("Δ") generate a modal logic as strong as S5, (1)–(4), and, presumably, Leibniz's Law, may each be strengthened with a "Definitely" prefix, enabling us to derive (5')…'. In our version of the proof, this reduces to the claim that [1] and [2] may both be prefixed with 'Δ'. Is this right?

To start with, Evans' assertion that '∇' and 'Δ' are duals is true only if 'Δ' is read as non-factive (that is, if Δ*p* does not imply *p*).[7] If 'Δ' is read as 'it is definite whether', then ~∇*p i*s equivalent to Δ~*p,* just as ~◊*p* is equivalent to ?~*p*. Second, I assume that Evans is not just endorsing a (trivial) conditional in his final paragraph. He must believe that the antecedent of the conditional has some plausibility. That is, he must believe that [1] and [2] may both be prefixed with 'Δ'.

The principle which would justify these strengthenings of [1] and [2] is: ∇*p* → Δ ∇*p*. (This principle is the analogue of the axiom distinctive of the modal system S5, ◊*p* → ? ◊*p*.) But this principle is not valid, because of considerations to do with higher-order vagueness. Consider the case where ∇*p* is itself indeterminate (a case of second-order vagueness). In that case, the conditional ∇*p* → Δ ∇*p* has an indeterminate antecedent. Its consequent should then be counted as false (if q is indeterminate, Δq is false). Plausibly, a conditional with an indeterminate antecedent and false consequent should not receive the value *true,* rather it should be counted indeterminate. Consequently, ∇*p* → Δ ∇*p* cannot be an axiom of vague logic. No plausible logic of vagueness will be as strong as the modal system S5.

However, all is not lost. A believer in vague identities presumably believes that some vague identities are definitely vague. (Just as in our example of the sequence of pens, the reddish-orange pens in the middle of the sequence are definite borderline cases of redness.) In the case of such identities, we can prefix [1] with 'Δ'. And, if [2] is true, we can also prefix [2] with 'Δ'. We can then validly infer [4] and conclude that there cannot be any definite cases of vague identity.

A defender of vague identities may claim that this shows only that all vague identities are indefinite. But there are two reasons why this is not a comfortable position to occupy. First, as noted, it is plausible that if there can be vague identities, there can be definite cases of vague identity. (There can be definite borderline cases of all other vague properties and relations. Why should identity be special in this regard?) Second, such a reply forces the defender of vague identities down an infinite regress: at no point can he allow a 'Δ' operator to appear in front of any vague identity; so he will be forced down an endless stream of higher orders of vagueness. This is not just uncomfortable, it is barely coherent.

Hence, despite the fact that $< \nabla, \Delta >$ does not generate a logic as strong as S5, Evans' proof is a valid *reductio* of cases which ought to be central to a friend of vague identities.

It's worth noting that it's not entirely clear why Evans thought his final paragraph necessary to his argument (that is, in my presentation, it's not clear why premise [4] is required). For [3] contradicts [1] in a perfectly straightforward way: [3] is true if and only if 'a=b' is false; and the falsity of 'a=b' is incompatible with [1]. The distinction between strong and weak negation, which might be thought to create a problem for this reply, is irrelevant here. (Negation is strong just if '~p' is true if and only if 'p' is false; weak just if '~p' is true if and only if 'p' is either false or indeterminate.) The only negation used in the derivation of [3] appears in [2], and that is surely an instance of strong negation. In which case, there is no reason to think that the negation in [3] is anything other than strong negation. So the final stage of Evans' argument appears to be dispensable.

Question (iii)

If the argument is valid, everything hinges on the answer to the question: Is premise [2] true? Evans assumed that *a* does not have the property of being such that it's vague whether it is identical to *a*. This is a plausible assumption. If we have successfully singled out an object, we cannot sensibly go on to ask whether that object is only vaguely identical to itself.

It might be objected that we are conflating two theses. One is the thesis that *a* is definitely self-identical (identical to itself). The other is the thesis that *a* is definitely identical to *a.* The truth of the first thesis, it may be urged, does not entail the truth of the second. And it is the second that Evans requires. However, even if this is right,

we still have no reason to question premise [2], which is plausible in its own right.

Since I take it that [2] is true, I conclude that Evans' proof is cogent. Identity is everywhere determinate. Hence, whenever an identity sentence is vague, this is because one or both of its singular terms is indeterminate in reference (like 'the first orange pen' in our example), and not because of any vagueness in the identity relation.

Premise [2] and vague objects

Should a believer in vague objects accept [2]? Perhaps a vague object *x* is precisely an object which is such that it's vague whether it is identical to *x*. So, if the purpose of Evans' proof is to undermine the possibility of vague objects, his proof is question-begging. David Wiggins has replied to this objection. He writes:

> even if…*a* were a vague object, we still ought to be able to obtain a (so to speak) perfect case of identity, provided we were careful to mate *a* with exactly the right object. And surely *a* is exactly the right object to mate with *a*. There is a complete correspondence. All their vagueness matches exactly.[8]

But there is a danger in this reply. If it is conceded that the truth of [2] is consistent with the view that *a* is a vague object, then surely the truth of [5] ought to be consistent with the view that *b* is a vague object. But in that case Evans cannot use his proof to undermine the possibility of vague objects.

So: either [2] is inconsistent with the possibility of vague objects or it is not. If it is, then Evans' proof, interpreted as a proof that there cannot be vague objects, is question-begging. (Indeed, we might then wonder why the proof is needed at all: why didn't Evans just write the following, much shorter, article: '~x[(∇(x=a)]a; so there cannot be any vague objects'?) If it is not, then [5] is also not inconsistent with the possibility of vague objects, so the conclusion of the proof fails to establish the impossibility of vague objects.

Fortunately, however, we can distinguish the soundness or otherwise of a proof from the uses to which its conclusion might or might not be put. It may be that, although Evans' proof is suasive, we cannot use its conclusion to argue against the possibility of vague objects.

We have already been presented with one reason for thinking that [5] does not imply the impossibility of vague objects. In the remainder of this section, I present another.

Vague identity and vague objects

We have not yet made explicit the connection between Evans' first paragraph and the remaining two, that is, between the incoherence in the idea of vague objects and the proof of the determinacy of identity. Evans presumably had a certain connection in mind viz., if $\sim\Diamond(\exists x)(\exists y)\ \nabla(x=y)$, there cannot be vague objects. Or, put differently, if it is always determinate whether or not the identity relation obtains, there cannot be vague objects. This conditional straightforwardly links the conclusion of the proof with the impossibility of vague objects.

There are two problems with this conditional. First, as noted above, if [2] (and hence [5]) are consistent with the existence of vague objects, then the conditional is false. Second, the conditional is open to a quite general doubt. There are other criteria for the existence of vague objects which are consistent with the determinacy of identity. For example, there is the criterion that an object is vague if it lacks precise spatial boundaries, and the criterion that an object is vague if it is vague whether it has such-and-such as a part.

These criteria may seem fairly anodyne. They are apt to provoke the deflationary response: 'If that's what you mean by "vague objects", then of course there are vague objects!' However, if these criteria are unsatisfactory, this is no thanks to the Evans conditional. Hence, the conclusion of Evans' proof, in the absence of further argument, is independent of issues concerning vague objects (in particular, the issues of what it means to say 'x is a vague object', and whether there are any vague objects).

Evans' proof and Kripke's proof

Evans' proof that identity cannot be vague is structurally similar to Saul Kripke's proof that identity cannot be contingent.[9] Just as Kripke proved $(\forall x)(\forall y)(x=y \rightarrow \Box(x=y))$, so Evans proved that $(x)(y)(x=y \rightarrow \Delta(x=y))$. But there is a point of disanalogy. As discussed in Chapter 4, Kripke noticed that the metaphysical thesis of the necessity of identity happens to have a semantic corollary. He noticed that, in the case of, for example, proper names, the following substitution is valid:

(i) A=B; so

(ii) □(A=B).

Terms for which such a substitution is valid, he called 'rigid designators'.

What is striking is that there appears to be no specifiable class of singular terms in natural languages (for example, proper names, or demonstratives) such that for any two members of that class, 'A' and 'B':

(iii) Δ(A=B)

is guaranteed to be true (where '?' is read as 'it is definite whether…'). Natural languages do not contain a non-gerrymandered class of 'precise' designators (that is, a class of terms such that any non-empty member of that class is guaranteed to have determinate reference). Singular terms from *any* semantic category (definite descriptions, proper names, demonstratives, etc.) can fail to have determinate reference.[10]

Conclusion

Evans' proof, properly understood, is sound. It shows that the identity relation cannot be vague or indeterminate. Hence, in particular, personal identity cannot be vague. But it would be a mistake to think that this result conflicts with our preferred criterion of personal identity over time. That criterion implies that there can be vague sentences of personal identity. It is vague whether I am the resulting person in *Indeterminacy*. But, in that case, the source of the vagueness is not the identity relation, it is the referential indeterminacy of our singular terms.

This conclusion may help mitigate Bernard Williams' observation that it is hard to know how I should react, upon hearing of the unfortunate fate of some future person, where I know that it is vague whether I am that person.[11] As he points out, we have no model for such expectation. It is not like the anticipation I might feel when told that I will die sometime in the next ten years, or that one of us in this room will shortly be killed, or that some unknown horror will befall me. In these cases, something unfortunate will either happen to me or it won't. There is no vagueness of the sort that has concerned us in this chapter. However, this situation may be rendered less problematic once we appreciate that the vagueness is linguistic and not ontological.

Thus, the soundness of Evans' proof in no way conflicts with consequences of our preferred criterion of personal identity over time. That criterion implies that certain personal identity sentences are vague. This is consistent with Evans' proof of the determinacy of identity.

6

PARFIT AND 'WHAT MATTERS'

Persons and value theory

In this chapter, I want to investigate whether the metaphysics of personal identity has any implications for value theory (theories in ethics and rationality). One contemporary philosopher, Derek Parfit, is the best known advocate of such implications.[1] He argues that the most plausible metaphysics of persons yields radical conclusions for ethics and rationality.

It is, of course, uncontentious that there is some connection between theories of persons and value theory. For example, a Roman Catholic's belief that we are immortal souls may bear on his view of the morality of abortion and euthanasia. He may hold that foetuses and the severely brain-damaged have souls, and that is why it is wrong to kill them.

However, in the case of a debate between a Catholic and an atheist about the morality of abortion, the value of persons is not called into question. What is in question is the range of entities which should be considered persons, and awarded the corresponding rights and/or obligations. In particular, should the extension of the concept *person* include foetuses and the severely brain-damaged? This is a question about the extension of the concept *person*.

The intent of Parfit's project, however, is far more subversive. The purpose of that project is to undermine the significance we currently attach to personal identity and distinctness, even amongst beings (for example, normal adult humans) that are uncontroversially persons. This conclusion, in turn, is used to motivate an impersonal or utilitarian ethical theory according to which the *quality* of experiences matters more than *who* has them. This is a novel line to take on a familiar dispute (utilitarianism versus absolutist moral theories) that has seemed to many to terminate in stalemate. Whether or not this project is

ultimately successful (and I shall argue that it is not), it is important to recognise its form or shape, and the character of the arguments that have been put forward on its behalf.

A new value theory?

The central feature of Parfit's value theory is the thesis that personal identity is not, in itself, an important relation. Identity is not what matters. It is a relation that is of no moral, rational, or practical importance. The thesis that personal identity is not what matters has two strands. According to one strand, personal identity *over time* is unimportant. According to the other strand, personal identity *at a time* is unimportant.

The unimportance of personal identity over time

On Parfit's value theory, personal identity over time is not what matters. Rather, it is various psychological relations which matter, relations which are concomitants of personal identity over time in the normal case, but not in an abnormal case such as fission. In *Fission,* I am psychologically continuous with both Lefty and Righty, but identical to neither. The relations of psychological continuity and personal identity come apart. On Parfit's theory, it would be irrational to regard what will happen to me in *Fission* as being as bad as ordinary death, or even strongly to prefer my own continued existence to fission. My relations to Lefty and Righty contain all that matters, even in the absence of identity.

Similarly, on Parfit's theory, it would be irrational to fear what will happen to me in *Teletransportation*. If I know that I will be psychologically continuous with my replica, then I know that all that matters is preserved, and I should not fear what will happen to me. It is irrelevant whether we judge that I am identical to my replica, or that I am distinct from my replica, or even that it is vague whether I am identical to my replica. Facts about personal identity are irrelevant to what matters.

More dramatically, it is a consequence of Parfit's view that I should not fear what will happen to me in *Branch-Line,* where I will shortly die, while my replica lives. My relation to my replica contains all that matters, even though, on any plausible theory of personal identity, my replica is not me. The fact that I am a different person from my replica in *Branch-Line* ought in no way to diminish my reason for 'special concern' about his future.

Parfit finds these conclusions liberating, and they reduce his fear of death. He writes:

> After my death, there will be no one living who will be me. I can now redescribe this fact. Though there will later be many experiences, none of these experiences will be connected to my present experiences by chains of such direct connections as those involved in experience-memory, or in the carrying out of an earlier intention…. Now that I have seen this, my death seems to me less bad.[2]

Other implications

The thesis that the identity of a person over time is unimportant has also been taken to undermine the self-interest theory of rationality, and has implications for the tenability of transtemporal moral notions such as compensation, punishment, responsibility, and personal commitment.

The unimportance of personal identity over time implies that pure self-interested concern is irrational. That is, it is irrational for me to be especially concerned about the fate of some future person just because that person is me. It follows that the self-interest theory of rationality is false. According to this theory, which has dominated so much thinking about rationality, there is only one future person that it is supremely rational for me to benefit: the future person identical to me. Since the self-interest theory places immense weight on a relation which has no rational significance, this theory cannot be correct.

Further, if we do not believe that personal identity over time is important, this may change our attitude to punishment, compensation, and commitment. Consider a case where there are only weak psychological connections between different stages of the same life. For example, suppose that a one-time criminal is now completely reformed. On the present view, the grounds are thereby diminished for holding the later self responsible for the crimes of the earlier self, or for compensating the later self for burdens imposed on the earlier self, or for regarding earlier commitments as binding on the later self. The uncontroversial truth that the earlier person *is* the later person is deemed too superficial or unimportant to support these moral claims.

If identity over time is not what matters, then it may be rational for me to care less about my psychologically distant self in the future. This may seem to restrict the scope of my concern. However, Parfit has suggested that even if it is rational for me to care less about my psychologically distant future self, I may be *morally* obliged to care as

much about my future self (that is, myself in the future) as I do about other people. What was previously thought to lie in the domain of rationality or prudence may, in fact, lie in the domain of morality.[3]

The unimportance of personal identity at a time

The thesis that the identity and distinctness of persons at a time is unimportant has been thought to lend support to utilitarianism. The thesis has been taken to imply that the fact of the 'separateness of persons' is not 'deep', and that less weight should be assigned to distributive principles. It thus supports (in part) the utilitarian doctrine that no weight should be assigned to distributive principles. On this view, we should simply aim to maximise the net sum of benefits over burdens, whatever their distribution. It is irrelevant *who* receives the benefits and burdens.

Self-concern and special concern

We can take the claim that identity is not what matters to imply the following normative claim: self-concern is irrational. Thus, my concern that *I* continue to exist, or that *I* not be in pain, is without justification. How radical a claim this is depends on the answer to a contrasting question: what is it rational to care about?

It might be thought that Parfit is saying the following: suppose I believe that a severe toothache will befall someone tomorrow. I'm justifiably concerned: suffering, after all, is a bad thing. Suppose someone tells me: 'the toothache will happen to you'. I am now much more concerned: *I* am going to be in pain.

Perhaps Parfit's claim is that the extra concern I have when I receive the additional piece of information is irrational. That is, although I have reason to be concerned about pain-tomorrow, I have no reason to be especially concerned about my-pain-tomorrow. This interpretation would fit with the utilitarian view that what matters is the quality of experiences, not who has them.

However, this is not Parfit's view. He thinks that, although identity does not matter, other relations (psychological continuity and/or connectedness) do matter. That is, it's perfectly rational for me to be more concerned about the fact that someone psychologically continuous with me will suffer toothache tomorrow than it is to be concerned about the fact that just anyone will suffer toothache tomorrow. That this is Parfit's view is clear from his choice of

examples: they all raise the question of what attitude I should have to future people who are strongly psychologically connected to me. In *Fission, Teletransportation* and *Branch-Line* I stand to some future person in just the strong psychological relations in which, for example, I stand to myself tomorrow, but I am not identical to that person. I have a psychological continuer who is not me.

Thus suppose that I shall divide tomorrow, and I know that Lefty will suffer toothache. Parfit thinks that I have reason for 'special concern' about Lefty's toothache. Moreover, Parfit thinks that I ought to be indifferent between the outcome of my being in pain tomorrow and the outcome of Lefty being in pain tomorrow. In this sense, identity does not matter.

What is this 'special concern' I have for Lefty? Plainly, it cannot be self-concern, since Lefty is not me. However, Parfit will take this to be a merely verbal point. It is simply a verbal fact that we cannot call my concern for Lefty 'self-concern', just as it is a verbal fact that we cannot call a married man a 'bachelor'.

That is, just as 'bachelor' is a composite concept built up out of more basic components, so Parfit will think that 'self-concern' is a composite concept, built up out of two components: *identity* and *concern*. My concern for Lefty's toothache is essentially the same as my concern for my own toothache, though only in the latter case can we speak of 'self-concern'.

Hence, the claim that self-concern is irrational implies that it is irrational for me to have any more concern for my own fate than I have for Lefty's fate. The identity component in 'self-concern' can only be a source of irrationality.

This line of thought will be untenable if 'self-concern' is not a composite concept or if it is not merely a verbal fact that we cannot call my concern for Lefty 'self-concern'.

This dispute over the concept of self-concern is analogous to a dispute over the concept of memory (see Chapter 2). On one version of reductionism, the concept *person* can be 'reduced' to other mental concepts, all of which can be understood without reference to the concepts of *person* or *personal identity*. But, as Locke emphasised, *memory* is a key constituent of our idea of a person, yet it seems resistant to identity-free description. If I remember yesterday's toothache, I must remember *my* toothache yesterday. That fact seems analytic of the concept of memory. So reduction is impossible—personal identity is built into the concept of memory.

The reductionist's reply is that we can think of *memory* as a composite concept, built up out of the concept of *identity* and the (invented) concept of *quasi-memory*. The latter concept is explicitly defined to be like *memory* in all relevant causal and phenomenological respects, yet is stipulated to be identity-free: I can q-remember someone else's experiences (for example, Lefty can q-remember my experiences). According to the opposing view, *memory* is a unitary concept, and *q-memory* is not really identity-free. Q-memory is an illusion of memory, and so presupposes the concept of memory.

The dispute about the concept of memory is structurally similar to our present dispute about the concept of self-concern. Just as the reductionist thinks of memories as q-memories of one's own experiences, so Parfit must think of self-concern as 'q-concern' ('special concern') for one's own future. And just as I can have q-memories of someone else's experiences, so I can have q-concern for someone else's future (for example, Lefty's future).

However, if *self-concern* is a unitary concept, the attempt to forge a concept of concern ('special concern') that lies between universal concern (equal concern for every person's pain) and self-concern will fail. The thesis that identity is not what matters would then have to imply: only universal concern is rational.

Four arguments for the new value theory

The new value theory is radical, and revisionary of ordinary ways of thinking. It is underwritten by the thesis that personal identity is not what matters. What are the arguments for this thesis? I can discern four such arguments in Parfit's work. Three are arguments for the thesis that personal identity over time is unimportant (the argument from analysis, the radical argument from analysis, and the argument from fission), and one is an argument for the thesis that personal identity at a time is unimportant (the argument from reductionism). I am critical of all four arguments.

The argument from analysis

What I call the 'argument from analysis' relies on the following general principle: for all relations, X and Y, if X is identical to (or 'consists in') Y, and Y doesn't matter, then X doesn't matter. Parfit illustrates the plausibility of this principle (which he calls 'reductionism about significance') with the following example.[4]

Suppose we accept, as a matter of definition, that someone is alive if and only if his heart is beating. Hence, someone who is brain-damaged and irreversibly unconscious will be alive provided his heart is still beating.

Is being alive, in itself, what matters to us? The condition of my being alive consists in my heart still beating. Yet there can be situations in which my heart is beating (hence I'm alive) which don't contain what matters: for example, if I were to become irreversibly unconscious. That outcome is as bad as death for me. What this shows is that being alive is not, in itself, what matters. What matters is rather a state that normally accompanies life, and for which life is a necessary condition—consciousness.

Thus: being alive consists in my heart beating; heart beating is not what matters (since there are situations in which my heart beats, yet what matters is absent); so being alive is not what matters.

I have no quarrel with the above reasoning. However, Parfit wants to apply the very same reasoning to the case of personal identity. This application is controversial. Parfit argues as follows. Suppose we accept that the relation of personal identity 'consists in' or is identical to some other relation (for example, the relation of non-branching psychological continuity). Then since the latter relation is not important (why should it matter to me whether I am psychologically continuous with one future person rather than with two?), so the relation of personal identity is not important either. What matters are relations of psychological continuity and connectedness that accompany personal identity in the normal case.

Parfit's general principle is, I think, unobjectionable. That is to say, the following argument is valid:

(i) X=Y;
Y is not an important relation;
——— so, X is not an important relation.

However, the following argument is also valid:

(ii) X=Y;
X is an important relation;
——— so, Y is an important relation

Parfit's principle is consistent with the validity of both types of argument. Thus his principle, by itself, cannot show that personal identity is not what matters. That conclusion follows only if an

argument of type (i), in the personal identity case, is sound. And arguments of type (i) are not always plausible.

Consider a case concerning not a relation, but a type of event: pain. If we were to become materialists about the mind, for example, identifying pain with a type of brain-state, it would be absurd to conclude that pains don't matter on the grounds that neural events are, in themselves, of no importance. Rather we would reason in accord with an argument of type (ii).

This case highlights a difficulty in applying Parfit's principle. We are supposed to ask whether Y matters, intrinsically or 'in itself'. That is, Parfit assumes that we can assess the significance of Y independently of the fact that Y is (or constitutes) X. But this will be impossible if such facts can imbue Y with significance (as in the pain example). So the question arises: why could we not use an argument of type (ii) in the case of personal identity?

This question would be answered if there is some reason to think that the relation of non-branching psychological continuity is unimportant. Two reasons have been given. Whether this relation obtains can depend on extrinsic factors (as in *Fission*). And whether it obtains can sometimes depend, in a purely causal sense, on a relatively trivial fact (say, whether or not a nurse drops the right hemisphere). A relation that depends on extrinsic facts or trivial facts cannot be important. Hence, in the case of personal identity, we should reason in accord with an argument of type (i).

Unfortunately, what matters can sometimes be determined extrinsically, and can sometimes depend on a trivial fact. For example, what matters can sometimes depend on extrinsic factors such as lack of equally good competitors (for example, it might matter a great deal to a scientist that they be the first person to cure AIDS). What matters can also depend on trivial factors. Important things can depend on trivial ones. The exact position of a bullet may be trivial, but a person's life may depend on it.

The latter point highlights another unclarity: what does 'trivial' mean in this context? Clearly, any 'important' event will depend causally on other events, and those latter events can always be described in such a way that, under that description, we're inclined to say that they're trivial. But, on this sense of 'trivial', it will be virtually tautological to claim that what matters can depend on a trivial fact. The claim that what matters cannot depend on a trivial fact will be obviously false.

Consequently, Parfit has not shown that acceptance of, for example, the intermediate criterion (or any criterion which incorporates a non-branching component) need lead to any reassessment of the importance of personal identity.

The radical argument from analysis

We mentioned above the possibility that changing our view of personal identity may alter our attitude to the justification of moral practices such as punishment and compensation. If direct psychological connections, and not personal identity, are what matters, a now-reformed criminal may deserve less or no punishment for the crimes of his earlier delinquent self. This can be regarded as a moderate claim, in that it does not undermine the retributivist justification for punishment in cases where there are strong psychological connections between a criminal and his later, unreformed, self.

However, Parfit has recently tried to argue for the much more radical conclusion that no one *ever* deserves to be punished for anything they did, and that it is impossible to compensate someone for suffering they endured earlier.[5] His argument is essentially the same in both cases. For simplicity I will focus on the case of punishment. The argument runs as follows.

Consider again the thought-experiment *Branch-Line,* and call my replica, 'Backup'. It is generally agreed that I am not Backup, even though I am fully psychologically continuous with him. Does Backup deserve to be punished for my crimes? Parfit writes: 'Backup is not me only because…these [psychological] continuities do not have their normal cause: the continued existence of my brain. Is it the absence of this normal cause which makes Backup innocent? Most of us would answer "no". We would think him innocent because he is not me.'[6]

According to Parfit, this reply would show that we take personal identity over time to be a 'further fact' over and above the obtaining of various physical and psychological continuities, normally caused. Only this 'further fact' could justify punishment. Hence, once we reject the 'further fact' view, we ought to conclude that '[n]o one ever deserves to be punished for anything they did'.[7]

There are two problems with this argument. First, the fact (if it is one) that most of us would answer 'no' to Parfit's question is only of relevance if it is the result of rational reflection. It is of no interest if it is merely an 'off-the-cuff' reaction. And, given the arguments of earlier chapters, I think that the answer 'no' would be wrong.

Second, there is a gap in the argument. Parfit assumes that if we believe (A) the 'further fact' view of personal identity, and we accept the truism that guilt requires personal identity (that is, X can be guilty of Y's crimes only if X is the same person as Y), then we will believe that guilt requires the 'further fact'. This conditional is correct. But Parfit suggests that if we were to give up (A) and embrace instead (B) the view that personal identity 'consists in' psychological continuity, normally caused, we ought to conclude that '[n]o one ever deserves to be punished for anything they did.' In other words, we will continue to believe that guilt requires the 'further fact', even after we have given up the 'further fact' view, and so will conclude that no one is ever guilty, and, hence, that no one ever deserves to be punished.

This cannot be right. The only reason we believed that guilt required the 'further fact' was because we believed (A). There is no reason whatever to think that the former belief will remain in place once belief in (A) has been given up. On the contrary, provided that we continue to believe that guilt requires personal identity, then once we believe (B) instead of (A), we will believe that guilt requires psychological continuity, normally caused. We will deem a person guilty only if that condition is satisfied. People will sometimes deserve to be punished.

More generally, it is plausible to hold that guilt requires personal identity, *whatever* the correct account of personal identity turns out to be. This independence of theories of value from theories of personal identity runs counter to the whole thrust of arguments from analysis.

We thus have no reason whatever to endorse the more radical conclusion that no one ever deserves to be punished.

The argument from fission

Recall our earlier thought-experiment, *Fission.* I argued that the most plausible description of *Fission* is that I am numerically distinct from both Lefty and Righty, though psychologically continuous with both. This description provides the first premise of the argument from fission: (1) I am not identical to either Lefty or Righty. The second premise is this: (2) fission is not as bad as ordinary death. This premise is taken to imply a third: (3) my relation to Lefty and Righty contains what matters. The first and third premises jointly imply: (4) personal identity is not what matters.

This is an interesting argument, which has had many adherents in recent years (most notably Parfit).[8] However, there is a problem with it. The problem concerns the move from the second to the third premise.

The second premise is certainly true: the prospect of fission is not as bad as that of ordinary death. But what grounds this premise, and what exhausts its true content, is simply that presented with a choice between those two options, virtually everyone would choose fission. Such a choice is both explicable and reasonable. After fission, unlike after ordinary death, there will be people around who can complete my public projects (that is, those projects of mine which others can, in principle, complete—finishing my book, looking after my family, etc.).

However, if the third premise is grounded in the second, the claim that my relation to Lefty and Righty contains what matters merely reflects the innocuous truth that fission is preferable to ordinary death. That is, even in the absence of identity, other things can matter. But whoever thought otherwise?

The argument from fission thus possesses no radical import. Its conclusion ((4)) does nothing to show, for example, that it's rational to be indifferent between the prospects of fission and continued existence, let alone that it's irrational strongly to prefer continued existence to fission. In particular, the argument from fission does nothing to undermine the self-interest theory of rationality.[9]

The argument from reductionism

The argument from reductionism attempts to establish the unimportance of personal identity and distinctness at a time. The argument presupposes a version of reductionism according to which a description of reality which refers to bodies and experiences, but omits reference to persons, can be complete. It would leave nothing out. (See the discussion of the entailment and epistemic models of reductionism in Chapter 2.) The argument from reductionism attempts to show that, if this version of reductionism is true, the fact of the 'separateness of persons' (for example, the fact that you and I are distinct persons) is not 'deep' or 'significant', and hence that less weight should be assigned to distributive principles.

The argument can be presented as follows. Suppose that reality can be completely described without reference to persons. If such a complete and impersonal description is possible, how can the boundaries between persons be important? Failing an answer to this question, the argument from reductionism concludes that the boundaries between persons are not morally significant.

The validity of this argument turns on the truth of the general principle that if reality can be completely described without referring to Fs, then the boundaries between Fs cannot be of any importance.

However, this principle is ambiguous. On one reading it says: if reality can be completely described without referring to Fs *explicitly,* then the boundaries between Fs cannot be of any importance. But this is false. A description of reality that fails explicitly to refer to water, but does explicitly refer to H_2O, may be complete. Yet water and its boundaries can be very important.

On its second reading, the principle says: if reality can be completely described without referring to Fs *implicitly,* then the boundaries between Fs cannot be of any importance. Now the danger is of triviality. It would seem that the antecedent of this conditional can be true only if there are no Fs, in which case the question of the importance of the boundaries between Fs cannot even arise. If we are to avoid eliminativism, the general principle cannot sustain the value conclusion.

Hence, on either reading the general principle is powerless to motivate an impersonal value theory.

Finally, even if the argument from reductionism were valid, there would still be legitimate worries, canvassed in Chapter 2, about the tenability of the version of reductionism which it presupposes.

Conclusion

The four central arguments for the thesis that identity is not what matters are all open to criticism. The failure of these arguments emphasises how difficult it is to undermine the moral and prudential importance I attach to the fact that such-and-such a person tomorrow is me, and to the fact that you are not me. Unless other arguments are forthcoming, we can continue to believe that personal identity is important, and to endorse the traditional views in ethics and rationality which that belief supports.

7

ANSCOMBE ON 'I'

Introduction

So far in the book, our concern has been with the metaphysics of personal identity, and (in the previous chapter) with the alleged consequences of particular metaphysical positions for issues in value theory. However, there are two other important dimensions to the topic of personal identity. One dimension is semantic. Here the following question comes to the fore: in what ways does the first-person singular differ from other personal pronouns and personal proper names?

The other, not unrelated, dimension is epistemological. In this case, the central question concerns the nature of the evident connection that exists between the ability to engage in first-person judgements (judgements of the form 'I am F') and self-consciousness. What is it about first-person judgements which makes them expressions of self-consciousness? Conversely, why is it that third-person judgements ('he is F', 'Garrett is F', 'the G is F', etc.), where the contained singular terms refer to the judger, are not expressions of self-consciousness?

A competent utterance of the form 'I am F' always and everywhere manifests self-consciousness. Yet utterances of 'he's F', 'Garrett is F', 'the G is F', where the singular terms refer to the utterer, do not manifest self-consciousness. Thus, if I know that I am F, I possess self-conscious self-knowledge. If I know that the G is F, where I am the G, I do not thereby possess self-conscious self-knowledge (though, in some sense, I know a truth about myself). Why the asymmetry?

If we can answer this question, we can illuminate the connection between first-person thought and self-consciousness. And if we can become clearer about what it is to be self-conscious, we can better understand what it is to be a person.

I will be directly concerned with the epistemological dimension of first-person judgement in the next chapter. In this chapter, my concern is to criticise one answer to the semantic question. This answer is due to Elizabeth Anscombe.[1] Anscombe argues that, despite all syntactic and semantic appearances to the contrary, the first-person singular pronoun is not a device of reference.

By this Anscombe does not mean that utterances of 'I' fail to refer in the way that empty names like 'Pegasus' or 'Odysseus' fail to refer. She means that 'I' does not even belong to the logical/syntactic category of referring terms (singular terms). In this respect, 'I' differs from other personal pronouns and from personal proper names. No one doubts that tokens of, for example, 'you' or 'she' or 'Nixon' belong to the category of referring terms.

Anscombe's thesis suggests one answer to the question mooted above concerning the link between 'I' -judgements and self-consciousness. What is special about 'I' -judgements, and what explains why they are expressions of self-consciousness, is the fact that, unlike 'he', 'Nixon', 'the inventor of bifocals', etc., 'I' is not a referring term. By appeal to this fact, it might be thought, we can explain what is special about first-person judgement in relation to self-consciousness.

Anscombe doesn't discuss this natural way of construing the wider significance of her thesis, and perhaps for good reason. The proposal is very unclear. Whatever it is about 'I' -judgements that makes them expressions of self-consciousness, it surely has very little to do with whether or not 'I' is a referring term. After all, a token of 'it' in 'it's raining' is generally agreed to be non-referential (a 'dummy' singular term), but this has nothing to do with self-consciousness.

It might be questioned whether 'I' -judgements are always and everywhere expressions of self-consciousness. On one view, 'I' -judgements are not, in themselves, expressions of self-consciousness. If an 'I' -user is self-conscious, that is in virtue of a further, psychological fact about him, viz., that he has self-referential intentions. That is, self-consciousness consists in the presence of an intention to self-refer, and not all competent 'I' -users need be deemed to possess such intentions. So, for example, it might be thought that a fairly sophisticated computer could count as a competent 'I' -user, whilst lacking self-referential intentions, and hence lacking self-consciousness.

However, it is not clear how something could count as a competent 'I' -user unless it possessed the intention to self-refer. Language use is an intentional activity. In the absence of linguistic intentions, there is no language use in the relevant sense. This is why parrots and speak-

your-weight machines are not language users, however much they may reproduce familiar sounds. Why think the computer is any different? Of course, I do not deny the possibility of a self-conscious computer (like HAL in the film *2001*), but such a computer, since also a person, is not a counter-example to the identification of 'I' -users with persons.

In this chapter I argue that 'I' is a referring term. Anscombe's arguments for the conclusion that 'I' is not referential are flawed, and there are positive reasons why we should regard 'I' as a referring term. Consequently, if 'I', like other personal pronouns and personal proper names, is a device of reference, we will need to appeal to further features of 'I' and 'I' -judgements in order to explain why uttering or thinking such judgements is our most distinctive manifestation of self-consciousness.

In order to illuminate those aspects of the epistemology of 'I' -judgements which make them apt for the expression of self-consciousness, we will have to invoke the 'as subject'/'as object' distinction drawn by Wittgenstein in the *Blue Book*. This distinction is the central concern of the next chapter, but the groundwork for it will be laid in the present chapter.

The common-sense view of 'I'

In contrast to Anscombe's view, there is a natural view of the first-person singular which I'll call the *common-sense view of 'I'*. This view has two components: the referential view and the indexical view. According to the referential view, 'I' is a referring term, just as much as names like 'Clinton' and 'Nixon'. According to the indexical view, the reference of a particular utterance of 'I' gets fixed by virtue of the following self-reference rule: a given token of 'I' refers to whoever produced it. So the common-sense view states that 'I' is a singular term, and explains how the reference of particular tokens of 'I' gets fixed.[2]

Note that the indexical view is not the only possible view of how the reference of tokens of 'I' gets fixed. It could be held that a token of 'I' gets its reference fixed by a definite description such as 'the thinker of these thoughts'. Alternatively, it might be held that the reference-fixer for 'I' is an 'inner' demonstrative such as 'this self'.

However, there are problems with both these views. The descriptive view is in danger of collapsing into the indexical view. Note that uses of 'I' always have sure-fire referential success: if X comprehendingly utters 'I am F', his token of 'I' can neither fail to refer nor refer to anyone other than X.[3] If we are to respect these referential guarantees,

we will have to understand the phrase 'these thoughts' in 'the thinker of these thoughts' as either '*my* thoughts' or 'thoughts to which *I* have special access'. But how are we to understand these latter occurrences of 'my' and 'I'? It seems that here we must implicitly rely on the indexical view. As for the demonstrative view, it is undermined by legitimate and familiar Humean worries about the possibility of any 'inner' demonstration of 'the self'.[4]

In addition to its superiority over rival views, two considerations favour the common-sense view. First, it explains why Smith's utterance of 'I am F' is true if and only if an utterance of 'Smith is F' is true. On the common-sense view, this biconditional is platitudinous because Smith's token of 'I' is referential, and it refers to the person referred to by 'Smith'. Second, the common-sense view does not require that a competent 'I' -user be in possession of any specifiable stock of true beliefs about himself. The 'austerity' of the self-reference rule thus explains that and why one's competent use of 'I' can survive both loss of many 'objective' beliefs about oneself (for example, beliefs about one's nature, history, and spatio-temporal location), and the acquisition of massively false beliefs about oneself (for example, 'I am Napoleon').

Of course, the fact that 'I' is referential cannot be the whole story about 'I' or 'I' -judgements. As noted, acknowledging the referentiality of 'I' fails to explain why 'I' -judgements are expressions of self-consciousness. However, the common-sense view claims not merely that 'I' is referential, but that it is governed by the self-reference rule. That is, 'I' is used as a device of *criterionless* self-reference: uses of 'I' are not grounded in any act of identification. As we shall see, this is intimately connected to the fact that 'I' -judgements manifest self-consciousness.

Two arguments against the common-sense view

In her paper 'The First Person', Anscombe attempts to undermine the common-sense view. She writes that: '"I" is neither a name nor another kind of referring expression whose logical role is to make a reference, *at all*.'[5] In this, she echoes Lichtenberg's anti-Cartesian recommendation that instead of 'I think' one ought to say 'it's thinking', on analogy with forms of words such as 'it's raining' and 'it's snowing'.

Anscombe's paper contains two main arguments for the conclusion that 'I' is non-referential. Her first argument attacks the referential view by attacking the indexical view. Her second argument attempts to

counter the referential view directly. These arguments fail, but their failure is instructive. Moreover, there is much else in Anscombe's article with which to agree. But the sober and true things she has to say can be accommodated without forfeiting the referential status of 'I'.

The first argument, which I shall call Anscombe's challenge, alleges that the indexical view fails to explain what is special about 'I', namely, that its competent use in judgement manifests self-consciousness.

The second argument, which I shall call the tank argument, concludes that '*if* "I" is a referring expression, then Descartes was right about what the referent was.'[6] That is, if 'I' refers, it refers to an immaterial Cartesian ego. This is taken to be a *reductio* of the view that 'I' is a referring term.

Anscombe's challenge

Anscombe invites us to:

> Imagine a society in which everyone is labelled with two names. One appears on their backs and at the top of their chests, and these names, which their bearers cannot see, are various: 'B' to 'Z' let us say. The other, 'A', is stamped on the inside of their wrists, and is the same for everyone. In making reports on people's actions everyone uses the name on their chests or backs if he can see these names or is used to seeing them. Everyone also learns to respond to utterances of the name on his own chest and back in the way and sort of circumstances in which we tend to respond to utterances of our names. Reports on one's own actions, which one gives straight off from observation, are made using the name on the wrist.[7]

Each person in this imaginary community has two proper names: one that is unique, and one that is shared ('A'). These names are the only devices of 'self'-reference in this community. Reports on one's own actions are made on the basis of observation (using the name 'A' on one's wrist), and on the basis of inference, including inference from the testimony of others.

It is difficult to over-estimate the extent of the differences between the 'A' -users and ourselves. When an 'A' -user says 'A is F' his judgement is always based on third-person or publicly accessible grounds, for example, observation of his behaviour or bodily condition, or inference from the testimony of others (on hearing 'B is F', B can infer 'A is F' given that he accepts 'A is B'). Even the 'A' -users' 'self'-

ascriptions of pain will have to be based on behavioural data. In short, as Anscombe observes, 'our description does not include self-consciousness on the part of people who use the name "A"'.[8]

The 'A' -users can be supposed to suffer quite generally from the 'lapse of self-consciousness'[9] displayed by William James' character, Baldy. James writes: 'We were driving…in a wagonette; the door flew open and X, alias "Baldy", fell out on the road. We pulled up at once, and then he said "Did anyone fall out?"…When told that Baldy fell out he said "Did Baldy fall out? Poor Baldy!".'[10]

In Anscombe's thought-experiment, we have described a singular term ('A') which, it seems, we can substitute for 'I' in the self-reference rule, salva veritate. In the case of our imagined community, 'A' is the name that each person uses to refer to himself. From this Anscombe infers that 'A' is governed by the self-reference rule. Yet, ex hypothesi, uses of 'A' in judgement fail to manifest self-consciousness. Hence, the indexical view can make no space for the incontrovertible fact that our 'I' -judgements manifest self-consciousness.

However, even if we were to grant Anscombe her premise that 'A' is governed by the self-reference rule, it would not follow that the indexical view is untenable. We must distinguish two conclusions: the weak conclusion that the indexical view *fails to explain* the evident fact that 'I' -judgements manifest self-consciousness, and the strong conclusion that the indexical view is *incompatible* with that evident fact. Anscombe needs the strong conclusion in order to refute the indexical view, but her argument, if successful, only supports the weak conclusion. The weak conclusion would tell against the indexical view only on the additional assumption that that view, if true, must explain the link between first-person judgement and self-consciousness. But why assume that the indexical view incurs this explanatory obligation?

More importantly, Anscombe's key premise is false: 'A' is not governed by the self-reference rule. Unlike 'I', competent use of 'A' is based on criteria. Certain observational conditions must be satisfied in order for an 'A' -user to refer using 'A'. That is, 'A' is not used *simply* as a device of indexical self-reference. Consequently, Anscombe's first argument is unsuccessful.

Some contrasts between 'I' and 'A'

In attempting to undermine the common-sense view, Anscombe shifts, not entirely consistently, from emphasising the supposed analogy between 'I' and 'A' to emphasising some of the supposed disanalogies. We have seen that the pronoun 'I' and the name 'A' are not semantically

equivalent. Do these and related contrasts between 'I' and 'A' give us reason to deny that 'I' is a name or any other type of referring expression?

What exactly are the contrasts? In discussing the various 'guarantees' to which a word 'X' might be subject, Anscombe distinguishes the following three: (i) the user of 'X' must exist, otherwise they would not be using 'X'; (ii) using 'X' implies the '…guaranteed existence of the object *meant* by the user'.[11] Thus, 'if I know someone called "X" and I call something "X" with the intention of referring to that person, a guarantee of reference in this sense would be a guarantee that there is such a thing as X.'[12] and (iii) in addition to the guarantee specified in (ii), 'what I take to be X *is* X'.[13]

It is not clear how the second guarantee is meant to differ from the third. It could be that Anscombe intends (ii) to correspond to immunity to reference-failure and (iii) to correspond to immunity to misreference. However, as stated, (iii) is an epistemic constraint rather than a semantic one, and failure to satisfy (iii) is actually compatible with immunity to misreference, as we shall see.

Anscombe concedes that 'I' and 'A' are both, trivially, subject to guarantee (i). She claims that 'A' is subject to guarantee (ii) but not to guarantee (iii), whereas 'I' is subject to all three guarantees. However, uses of 'A' are not subject to constraint (ii) since reference-failure and misreference are possible using 'A'. (An 'A' -user may hallucinate an 'A' -inscribed wrist or may mistake someone else's wrist for his own.)[14] Moreover, although all uses of 'I' are immune to both reference-failure and misreference, some uses of 'I' are not subject to constraint (iii).

Anscombe thinks it impossible that an 'I' -user might 'take the wrong object to be the object he means by "I".'[15] And, in parenthesis, she writes: 'The bishop may take the lady's knee for his, but could he take the lady herself to be himself?'[16] The answer to this question, unfortunately, is 'yes'. In perhaps slightly bizarre circumstances, the bishop could indeed take the lady to be himself, just as one might misidentify oneself in a mirror or photograph. (I might single someone out in a photograph and falsely judge that person to be me.) Not all uses of 'I' are subject to constraint (iii).

The central contrasts between uses of 'I' and 'A' appear then to be the following. Most fundamentally, uses of 'A' involve the satisfaction of observational criteria, and require an identification and reidentification of their object. Uses of 'I' are criterionless in virtue of being governed by the self-reference rule. As a consequence, uses of 'I'

are immune to reference-failure and to misreference, whilst uses of 'A' are not so immune.

Are these contrasts prejudicial to the referential status of 'I'? Anscombe writes that: 'Getting hold of the wrong object *is* excluded, and that makes us think that getting hold of the right object is guaranteed. But the reason is that there is no getting hold of an object at all.'[17] Elsewhere she describes as nonsensical the idea that utterances of 'I' have guaranteed reference since 'it would be a question what guaranteed that one got hold of the right self'.[18]

These remarks are far from persuasive. Given the failure of Anscombe's challenge, there is no reason why we cannot explain the guarantee of sure-fire reference by citing the self-reference rule which governs 'I': a token of 'I' refers to whoever produced it. If Anscombe is implying that the very idea of guaranteed reference is some sort of oxymoron, then argument to that effect is required.[19]

Anscombe cites the criterionless application of 'I' as a reason to count it as a non-referring expression. She suggests that 'I' is not a referring expression since, if it were, 'a repeated use of "I" in connection with the same self would have to involve a reidentification of that self…but this is not any part of the role of "I". The corresponding reidentification *was* involved in the use of "A", and that makes an additional difference between them'.[20]

As noted above, it is true that uses of 'I' are criterionless or unmediated. When I come to believe, in the normal way, that I am in pain, I do not first judge that something which satisfies a certain criterion or condition is in pain. Similarly, my use of 'I' in first-person judgements of memory or intention does not involve any reidentification (or 'keeping track') of their subject.

Why does Anscombe think that, if 'I' were a referring term, its continued use by the same individual would have to involve a reidentification of its object? She must assume that the continued use of any referring term involves a reidentification of its object. In particular, since an 'A' -user's continued use of 'A' involves a reidentification of its object, the same must be true of 'I' if it is a referring term. But why grant the assumption? There is no reason why 'I' cannot be a referring term even though its continued use does not involve any reidentification of its object.

In sum, the contrasts between 'I' and 'A' give us no reason to think that the semantic features of 'I' militate against its claim to be a referring term.

Some residual worries

Anscombe has two residual worries: (i) Is there any appropriate sense-giving sortal which might 'cover' 'I' if it really is a referring term? (ii) Is there any extant referential category in which the pronoun 'I' might usefully be placed?

(i) According to Anscombe, if 'I' is a referring term, we will be 'driven to look for something that, for each "I" -user, will be the conception related to the supposed name "I".'[21] Here Anscombe assumes, with Frege, that every referring expression has a sense, or conception of its object, associated with it. However, even given this assumption, why can't we regard for example, *human being* as the sortal governing 'I', in just the way that *city* covers 'Sydney' and *kangaroo* covers 'Skippy'?

Anscombe has two objections to this suggestion.[22] First, if a token of 'I' refers to the human being who uttered it, then the relation of reference together with the sortal concept *human being* should enable us to explain the puzzling properties of 'I' (in particular, the link between 'I' -judgements and self-consciousness). But they don't: the name 'Richard Nixon', for example, refers to an individual who falls under the sortal *human being,* but this fact does nothing to illuminate the phenomenon of self-consciousness.

However, why would anyone think that the interesting features of 'I' have to be explained solely in terms of the sortal *human being* and the relation of reference? Obviously they cannot. We will have to appeal to those epistemic features of 'I' -judgements which Wittgenstein attempted to mark with his 'as subject'/'as object' distinction.

Second, Anscombe thinks that *human being* cannot serve as a covering sortal for 'I' since, if 'I' refers, it can only refer to an immaterial ego and not to a human being. The only sortal that could cover 'I', if it were a referring expression, would be *ego*. This is the conclusion of her tank argument, which I will discuss shortly.

(ii) Anscombe discusses the two models of self-reference criticised earlier: the descriptive model and the demonstrative model.

Anscombe is critical of both models of self-reference. But even if these models are inadequate, it does not follow that 'I' is not a referring term, since there is available a perfectly tenable model of self-reference: the indexical view of 'I'. We should therefore say that 'I' is an indexical term, a category it shares with 'here' and 'now'. This is the referential category in which 'I' can be situated.

Of course, Anscombe thinks that her arguments have undermined the indexical view. But it has emerged unscathed from her criticisms.

And, as we shall see, the availability of the indexical model of self-reference enables us to undermine the central presupposition underlying her second argument.

The tank argument

Anscombe's second argument starts out with the following question: 'Let us waive the question about the sense of "I" and ask *only* how reference to the right object could be guaranteed.'[23] '[T]his reference could only be sure-fire if the referent of "I" were both freshly defined with each use of "I", and also remained in view so long as something was taken to be *I*…. [I]t seems to follow that what "I" stands for must be a Cartesian [e]go.'[24]

This line of reasoning is underwritten by the tank argument:

> [I]magine that I get into a state of 'sensory deprivation', [no input from the senses and no bodily feeling]…I tell myself 'I won't let this happen again!'. If the object meant by 'I' is this body, this human being, then in these circumstances it won't be present to my senses; and how else can it be 'present to' me? Am I reduced to, as it were, 'referring in absence'? I have not lost my 'self-consciousness'; nor can what I mean by 'I' be an object no longer present to me.[25]

In other words: if 'I' refers, what I mean by 'I' is an object that is always 'present to' me. In a sensorily deprived state, no material object (for example, human body or human being) is 'present to' me. Since I remain a competent 'I' -user whilst sensorily deprived, what is 'present to' me must be something immaterial, a Cartesian ego. But the Cartesian view is absurd. Hence, we should reject the assumption that led to this result, and conclude that 'I' is not a referring expression.

Anscombe makes two questionable assumptions in the course of this argument. First, she assumes that if 'I' were to refer to my body, then the referent of 'I' would have to 'present' itself to me as a body. However, why assume that if the self were something bodily, and were perceived introspectively, it would have to be perceived *as* something bodily? An analogy suggests otherwise: even if pains are neural events, it doesn't follow that when I feel a pain, I feel it *as* a neural event.

Second, Anscombe assumes, more generally, that if 'I' refers, its object must be 'present to' its subject. But the thesis that self-reference requires self-presentation has little to recommend it. It is quite consistent to endorse the referential view of 'I' and to concede, with

Hume, that there is no distinctive introspective phenomenology of the self.[26] Indeed, elsewhere Anscombe rightly insists that self-consciousness is best understood, not as consciousness of a self (where this is understood as 'inner perception' of an object), but as 'consciousness that such-and-such holds of oneself'.[27]

The two assumptions which motivate the tank argument have a deeper source. The underlying presupposition is that if 'I' were a referring term, it would function as an 'inner' demonstrative (for example, 'this self'). It would not be classified as an indexical term like 'here' or 'now'.

It is plausible to suppose that the demonstrative 'this' differs from the indexical 'here' in the following respect. It is necessary for a subject to individuate an object demonstratively that he actually have information deriving from that object. For example, if I am to succeed in individuating someone using the present-tense sentence 'that man is bald', I must currently be in receipt of appropriate visual information deriving from him.

Arguably, this constraint is not required in order for a subject to think of a place simply as 'here'. A subject need only be *disposed* to have his 'here' -thinking controlled by the appropriate information. For example, I may successfully refer to a place using 'here' even if I am receiving no sensory input which enables me to individuate that place. (I may be blindfolded and driven around in the boot of a car.) It is enough that I would receive the appropriate information if certain barriers were removed. Hence, the reference-fixing rule for 'here' ('an utterance of "here" refers to the place at which it was uttered'), enables one to refer to a place as 'here', even in the absence of any place-individuating information.

Anscombe's guiding thought in the tank argument is that if 'I' refers, it must be a device of demonstrative reference, rather than of indexical reference. If 'I' were a referring term, it would function as an 'inner' demonstrative (for example, 'this self'). As such, any token 'I' -thought would require that its thinker be in receipt of information deriving from the object demonstrated. In the tank, only an immaterial ego could serve as the source of such information. Hence, if 'I' refers in the tank, it refers to an ego, and so much the worse for the view that 'I' refers.

However, since we have no reason to accept Anscombe's guiding thought, the tank argument collapses. Indeed, Anscombe's thought-experiment may give us an additional reason to reject the demonstrative view. How could 'I' function as a demonstrative when its sensorily deprived subject is in receipt of such minimal information?

Anscombe's positive view

What is Anscombe's positive view of 'I', given that she thinks its function is not to refer? This is sketched in the last few pages of her article. Neither 'I am BG' nor 'I am this thing here' is a proposition of identity. Rather '"I am this thing here" means: this thing here is the thing, the person (in the "offences against the person" sense) of whose action *this* idea of action is an idea, of whose movements *these* ideas of movements are ideas,…'[28] '"The person" is a living human body.'[29] '[T]his body is my body' means 'my idea that I am standing up is verified by this body, if it is standing up.'[30] 'These "I" -thoughts [such as "I am sitting"] are examples of reflective consciousness of states, actions, motions, etc., not of an object I mean by "I", but of this body. These "I" -thoughts…are unmediated conceptions…of states…of this object here [EA].'[31] She goes on to say: 'the "I" -thoughts I've been considering have been only those relating to actions, postures, movements and intentions. Not, for example, such thoughts as "I have a headache", "I am thinking about thinking", "I see a variety of colours", "I hope, fear, envy, desire", and so on. My way is the opposite of Descartes'. These are the very propositions he would have considered, and the others were a difficulty for him. But what were most difficult for him are most easy for me.'[32]

What should be said about this positive view? We can agree that a person can have 'unmediated' access to a range of states that he is in, such as sitting. The realisation that one can have 'unmediated' knowledge of physical self-ascriptions such as 'I am sitting' prefigures Gareth Evans' insight that such 'unmediated' or 'as subject' self-knowledge is not confined to mental self-ascriptions.[33] However, we can retain this insight consistently with regarding 'I' as a bona fide referring term.

Moreover, we need not agree with Anscombe that 'objective' or bodily 'I' -judgements are somehow more basic than 'subjective' or mental 'I' -judgements. Nor, with Descartes, need we assume the opposite. Rather, mental self-ascriptions and physical self-ascriptions should be seen as equally basic to first-person thinking.

Supporting the referential view

The referential view of 'I' has emerged unscathed. Further, as Anscombe is aware, two considerations strongly favour this view. First, 'I' has the same 'syntactical place'[34] as a referring expression. Second, an occurrence of 'I' in a sentence 'I am F', uttered by X, can be

replaced salva veritate by the name 'X'. Both considerations make a powerful case for the referentiality of 'I'.

Anscombe is not convinced. She objects to the first consideration on the grounds that it is 'absurd' to argue from syntax to reference— 'no one thinks that "it is raining" contains a referring expression, "it"'.[35] But the analogy is lame. The non-referential character of such uses of 'it' is manifested in other ways. For example, we cannot infer 'Something is raining' from 'it is raining'. But we can infer 'Someone is in pain' from 'I am in pain'. The possibility of parenthetical qualification of the subject is a further mark of referentiality.[36] Such qualification is possible in the case of 'I' (as in, for example, 'I, the person speaking to you now, am Scottish'). But no such parenthetical qualification makes sense in the case of feature-placing uses of 'it' (for example, 'it, the sky above you, is raining').

Anscombe objects to the second consideration on the grounds that although the biconditional 'If X asserts something with "I" as subject, his assertion will be true if and only if what he asserts is true of X' is perfectly correct, it is not a 'sufficient account' of 'I', since it does not distinguish between 'I' and 'A'.[37] However, even if this were right, it would not imply that 'I' is not referential. Second, as we have seen, 'I' and 'A' can be distinguished—in particular, 'A' is not governed by the self-reference rule. Moreover, the above biconditional is not true with 'A' in place of 'I'. If an 'A' -user (say, B) were to mistake C's wrist for his own, he might truly assert 'A is F', yet fail to assert something true of himself. In such a case, B's use of 'A' would refer to C.

Conclusion

Anscombe ends her article with some comments on the story of Baldy, her prototype 'A' -user. She says of Baldy that:

> …his thought of the happening, falling out of the carriage, was one for which he looked for a subject, his grasp of it one which required a subject…. He did not have what I call 'unmediated agent-or-patient conceptions of actions, happenings and states'. These conceptions are subjectless. That is, they do not involve the connection of what is understood by a predicate with a distinctly conceived subject. The (deeply rooted) grammatical illusion of a subject is what generates all the errors we have been considering.[38]

Here, as elsewhere in Anscombe's article, there is much that is right. On the occasion in question, Baldy does lack 'unmediated' conceptions of his states and actions. But the assertion that 'unmediated' conceptions are 'subjectless' runs together two distinct thoughts, one good, the other bad. The good thought is that when I sincerely judge, in an 'unmediated' or criterionless way, that I am in pain, my judgement is not the upshot of any identification or observation.

This thought should be distinguished from the bad thought according to which 'unmediated' conceptions are subjectless in the sense that there is no subject to which 'I' refers. There is indeed no 'distinctly conceived' subject in unmediated conceptions of one's states and doings. There is no object that one first identifies and then judges to be oneself. But this does not imply that there is no subject. The subject of first-person reference exists. It is not a 'grammatical illusion', or shadow, cast by the intriguing features of the singular pronoun 'I'.[39]

8

WITTGENSTEIN ON 'I'

Introduction

The concern of the previous chapter was with the semantics of 'I'. The concern of this chapter is with the epistemology of 'I' -judgements. The central question in this latter topic is: What explains the link between 'I' -judgements and self-consciousness? That is, what is it about 'I' -judgements which makes their comprehending utterance apt for the expression of self-consciousness?

Another question concerns whether we can be said to know, or have any cognitive attitude towards, an important subset of our own first-person judgements. This subset, known often as 'avowals', is composed paradigmatically, but not exclusively, of present-tense first-person psychological judgements. Further, whether or not we can be said to know what we avow, how are we to account for the evident authority with which we credit an avower regarding the veracity of his avowals? In addressing all these questions, I will draw heavily on some ideas of the middle and later Wittgenstein.[1]

In the previous chapter, I argued against Anscombe that 'I' is a singular term. As noted there, the mere fact that 'I' is a referring term cannot account for the connection between 'I' -judgements and self-consciousness. My own proper name is also a singular term which I use to refer to myself, but it does not feature in judgements which manifest self-consciousness. I can know that Garrett is F without knowing that I am F, if, for example, I am amnesiac and have forgotten that I am Garrett. My knowledge that Garrett is F is not self-conscious self-knowledge. Whereas the knowledge I express by uttering or thinking 'I am F' is always self-conscious self-knowledge. This asymmetry stands in need of explanation.

In addition to maintaining that 'I' is a referring term, I also claimed that 'I' is governed by the indexical self-reference rule according to which a given token of 'I' refers to whoever produced it, and that, as such, 'I' is used as a device of criterionless self-reference. Any such device will feature in judgements which are apt for the expression of self-consciousness. The use of such a referential device will require the presence of an intention to self-refer, and such intentions can be possessed only by self-conscious beings.

First-person judgements are our most distinctive expression of self-consciousness. Nonetheless, we can and should draw an important distinction within the class of 'I' -judgements—the distinction Wittgenstein labelled as that between 'as subject' and 'as object' uses of 'I'.[2] Although all 'I' -judgements express self-consciousness, the 'as subject' use of 'I' is fundamental to an understanding of self-consciousness, and this use is prior to the 'as object' use.

Self-consciousness

Our hope then is that understanding Wittgenstein's distinction will help to illuminate our concept of self-consciousness. What is that concept? Locke's definition of personhood captures the core idea of a self-conscious being as: 'a thinking, intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places.'[3] But what is it to consider 'oneself as oneself, in the sense that Locke intends?

I agree with Anscombe that one answer can immediately be discounted. The capacity to consider 'oneself as oneself' is not tied to any 'inner perception'. Self-consciousness should not be understood as 'inner perception of oneself'.[4] As Hume pointed out, when I look 'inside' myself I never catch myself; I only perceive some particular thought, or feeling, or memory image. (Indeed, it is unclear what it would be to perceive oneself introspectively.)[5] Wittgenstein's distinction holds out the promise of an alternative, non-perceptual account of what it is to consider 'oneself as oneself, that is, of what it is to be self-conscious.

It should be noted that, in attempting to illuminate the concept of self-consciousness, there need be no prospect of a 'reduction' of the concept of self-consciousness. In fact, the general drift of these last two chapters is towards the conclusion that the concept of self-consciousness is a basic concept.

That is, it is plausible to suppose that the concept of self-consciousness cannot be decomposed into more basic concepts

('consciousness' plus something else), such that the 'something else' can be fully understood without reference to the concept of self-consciousness.[6] Of course, the concept of consciousness can be so understood (many animals are conscious, but not self-conscious). What is contentious is the idea that the conceptual residue (the 'something else') can be fully understood without reference to the concept of self-consciousness.

One reason for thinking that *self-consciousness* is a basic concept is that it does not seem possible to add some ingredient to the conceptual repertoire of Anscombe's 'A' -users, so that the 'A' -users would then become self-conscious, where this new ingredient can be fully understood without reference to the concept of self-consciousness. (For discussion of the 'A' -users, see the previous chapter.)

Wittgenstein and the 'as subject' use of 'I'

Consider the following brief but well-known passages from Wittgenstein's *Blue Book:*

> Now the idea that the real I lives in my body is connected with the peculiar grammar of the word 'I', and the misunderstandings this grammar is liable to give rise to. There are two different cases in the use of the word 'I' (or 'my') which I might call 'the use as object' and 'the use as subject'. Examples of the first kind are these: 'My arm is broken', 'I have grown six inches', 'I have a bump on my forehead', 'The wind blows my hair about'. Examples of the second kind are: 'I see so-and-so', 'I hear so-and-so', 'I try to lift my arm', 'I think it will rain', 'I have toothache'. One can point to the difference between these two categories by saying: The cases of the first category involve the recognition of a particular person, and there is in these cases the possibility of an error, or as I should rather put it: The possibility of an error has been provided for…. It is possible that, say in an accident, I should feel a pain in my arm, see a broken arm at my side, and think it is mine, when really it is my neighbour's. And I could, looking into a mirror, mistake a bump on his forehead for one on mine. On the other hand, there is no question of recognizing a person when I say I have a toothache. To ask 'are you sure it's *you* who have pains?' would be nonsensical…. And now this way of stating our idea suggests itself: that it is impossible that in making the statement 'I have a toothache' I should have mistaken another person for myself, as it is to moan

> with pain by mistake, having mistaken someone else for me. To say 'I have pain' is no more *about* a particular person than moaning is.[7]
>
> We feel then that in cases in which 'I' is used 'as subject', we don't use it because we recognise a particular person by his bodily characteristics; and this creates the illusion that we use this word to refer to something bodiless, which, however, has its seat in our body. In fact this seems to be the real ego, the one of which it was said, 'Cogito, ergo sum'…[8]

For present purposes, we can extract the following five claims from these remarks, (i) There are two different uses of the pronoun 'I', 'the use as object' and 'the use as subject', (ii) Given Wittgenstein's examples, we are evidently meant to infer that 'as subject' uses feature only in mental self-ascriptions.[9] (iii) All and only 'as object' uses 'involve the recognition of a particular person', (iv) Only in such uses has 'the possibility of an error been provided for' viz., the error of mistaking another person for oneself, (v) It is a misreading of the 'grammar' of 'as subject' uses of 'I' which fuels the illusion of a Cartesian subject.

However, Wittgenstein's characterisation of the 'as object'/'as subject' distinction is flawed. The problem is that (ii) is inconsistent with (iii) and with (iv). On natural ways of unpacking (iii) and (iv), 'as subject' uses of 'I' can feature in physical (as well as mental) self-ascriptions.

The tension between (ii) *and* (iii)

What does Wittgenstein mean when he says that 'as object' uses of 'I' 'involve the recognition of a particular person'? What is meant by 'recognition'? There is a trivial sense in which all singular thought involves recognition of its object: when I think that x is F, I must identify the referent of 'x' in thought. Presumably, this is not what Wittgenstein had in mind.

We do better to suppose that a judgement 'I am F' involves the recognition of its subject if and only if it is the result of an inference from premises 'X is F' and 'I am X' *or* it is the upshot of a non-inferential act of 'outer' perception (for example, a demonstrative identification).[10] In short, a judgement 'I am F' involves the recognition of its subject if and only if it is *criterially based*.

With regard to the first disjunct, it is important that the inference utilise a first-person identity premise ('I am X'). Not all inferred self-

ascriptions need involve recognition of their subject, for example, where 'I am F' is inferred from 'I am F and I am G' (where neither conjunct is criterially based).

Thus: an occurrence of 'I' in 'I am F' is 'as object' just if it is criterially based. A typical example of this is when a person judges 'I am F' after identifying himself with someone in a mirror or photograph.

However, (ii) and (iii) are now in conflict. First, many occurrences of 'I' in physical self-ascriptions will count as 'as subject'. My judgement 'I am sitting', known in the normal way (through proprioception), is not criterially based. In fact, self-ascriptions which are the result of explicit identity-involving inferences or acts of 'outer' perception are fairly rare.

Second, some occurrences of 'I' in non-dispositional mental self-ascriptions will count as 'as object'. Imagine a futuristic brain-scanner which emits a certain signal whenever the person connected to it is in pain. If I am connected to the brain-scanner, I might judge that I am in pain on the grounds that, according to the emitted signal, someone is in pain, and I believe that person to be me. (This fanciful example is perhaps rendered more palatable if we are anyway inclined to reject the self-intimation thesis that, necessarily, if I am in pain, I thereby know, by introspection, that I am in pain.)

The tension between (ii) *and* (iv)

According to (iv), in 'as object' uses of 'I' there is always the possibility of mistaking another person for oneself. In contrast, self-ascriptions incorporating 'as subject' occurrences of 'I' are immune to error through misidentification of the subject. How should we understand this immunity? If we read 'misidentification' as 'misrecognition', (iv) merely becomes a variant of (iii). If there is no recognition or identification of the subject in 'as subject' uses of 'I', there can be no possibility of misrecognition or misidentification.

However, there is another way of understanding the impossibility of misidentification. We can say that a self-ascription 'I am F', known in a certain way W, is immune to error through misidentification of the subject just if the question 'Someone is F, but is it me?' makes no sense, where the existential component is also known in way W. Thus, if I feel a pain in my foot, the question 'Someone is in pain, but is it me?' will make no sense to me, provided that the existential component is known in the same way

(viz., feeling). (The qualification 'known in the same way' is necessary, otherwise the 'Someone…?' question would always make sense.)

One consequence of this definition is that if the justification for a putative 'as subject' judgement (say, 'I remember watching sunrise yesterday') were found to be wanting, that justification could not simply transfer over and act as a justification for 'Well, at least someone saw sunrise yesterday'.[11] Rather, the conclusion to draw would be that one was the victim of some kind of illusion.

On this reading, (iv) does not carve up 'as object'/'as subject' uses in the way implied by (ii). In particular, an occurrence of 'I' in the physical self-ascription 'I am sitting', known in the normal way (that is, through proprioception), is immune to error through misidentification of the subject. In such a case, the question 'Someone is sitting, but is it me?' makes no sense, where the existential component is known through proprioception. (This should not be taken to imply that 'I am sitting' is incorrigible. I can, of course, be mistaken about my bodily position.)

Further, an occurrence of 'I' in the non-dispositional mental self-ascription 'I am in pain', known via the brain-scanner's emissions, is not immune to error through misidentification of the subject. Since my belief that I am hooked up to the scanner may be false, the question 'Someone is in pain, but is it me?' will make sense—at least if it is only via the brain-scanner that I know that I'm in pain.

Running repairs to the 'as subject'/'as object' distinction

Thus (ii) and (iii), and (ii) and (iv), are inconsistent pairs. The culprit is (ii). Wittgenstein thought that whether an occurrence of 'I' is 'as object' or 'as subject' depended only upon the *type* of proposition in which it figured. That is, he thought that all occurrences of 'I' in tokens of 'I am in pain' are 'as subject', and all occurrences of 'I' in tokens of 'I am sitting' are 'as object'. This was a mistake. As (iii) and (iv) make clear, the 'as object'/'as subject' distinction is epistemological. It is a contrast between different ways of knowing truths about oneself, and it is a contrast which cuts across the distinction between physical and mental self-ascriptions.

The last point is important. Gareth Evans cites the fact that 'as subject' uses of 'I' feature in both physical and mental self-ascriptions in support of an anti-Cartesian view of persons according to which mental properties are no more basic to personhood than physical

properties.[12] This view is a close cousin of P.F.Strawson's claim that the concept of a person is 'primitive'.[13] That is, it is the concept of a subject to which both mental and physical predicates can be applied.

Moreover, it is misleading to talk, as Wittgenstein does, of different *uses* of the pronoun 'I'. This wording suggests that the 'as subject'/'as object' distinction is a semantic one, and that 'I' is ambiguous in meaning. As we have seen, the distinction is an epistemic one. Further, it is a distinction between complete 'I' -judgements and not between uses of the word 'I'. (Since the terminology is so ingrained, I will continue to talk of the 'as object' and 'as subject' uses of 'I', but this should always be understood as an epistemic distinction within the class of 'I' -judgements.)

A new definition

From our reconstruction of Wittgenstein's remarks we have generated two definitions of the 'as subject' use of 'I'. *First,* a token judgement 'I am F' is an 'as subject' judgement just if it is known in a way that does not involve any inference from premises 'X is F' and 'I am X', or any act of 'outer' perception. Call an 'I' -judgement known in a way that is not thus criterially based, 'ungrounded'. (The emphasis on knowledge is important: an 'I' -judgement that just pops into one's head at random, though not criterially based, does not count as 'as subject'.) *Second,* a token judgement 'I am F', known in a certain way W, is an 'as subject' judgement just if it would make no sense to ask 'Someone is F, but is it me?', where knowledge of the existential component is gained by way W.

However, it is plausible to suppose that the second definition is true just in case the epistemic route W is ungrounded. In which case, the first definition should be regarded as fundamental, the second merely a consequence of it.

There are also problems with the second definition. To start with, if I *know* that I am F, the question 'Someone is F, but is it me?' presumably will not make much sense. But the reason for this has nothing to do with first-person knowledge. It is simply a consequence of the fact that the singular proposition in question is known. If, sitting in a café, I watch rain fall on the Boulevard Saint-Michel, the question, should it occur to me to ask it, 'It's raining somewhere, but is it raining in Paris?', presumably won't make much sense. But this is simply because I know it's raining in Paris.

There is a further worry. Suppose I know that I was F, on the basis of memory. Would it make sense for me to say 'Someone was F, but

was it me?', where knowledge of the existential component was gleaned by memory? As we saw from our discussion of memory and q-memory in Chapter 2, memory-states are identity-involving. It is impossible to have an experience-memory with the impersonal or identity-free content: 'Someone was F'. Memory-knowledge of the existential component must therefore be inferred from the prior knowledge that I was F. But then, in such a case, the question 'Someone was F, but was it me?' obviously cannot make any sense. The doubt it expresses in that context is not coherent.

However, this result is secured by the triviality that experience-memory implies identity. That is, we can't call a mental state 'experience-memory' unless the rememberer is the same person as the person who had the remembered experience. So, in the case of identity-involving concepts such as memory and intention, the second definition will again be satisfied too easily, and for the wrong reason.

We should therefore take the first definition to be our canonical definition of the 'as subject' use of 'I'. The hope is that the epistemic phenomenon of ungroundedness, which that definition encapsulates, may help to elucidate those features of 'I' -judgements which make their competent use expressions of self-consciousness.

A possible tension?

It might be wondered whether there is any tension between the 'as subject' or ungrounded aspect of certain, for example, memory-based, 'I' -judgements and the fact that judgements of personal identity have grounds or criteria (viz., those elucidated in Chapters 3 and 4). Thus, X's *ungrounded* memory-judgement ('I was F', uttered on the basis of memory) implies a truth of personal identity ('X is identical to the person who was F'). All such truths are criterially *grounded*? Is there a contradiction?

There is no contradiction. The ungroundedness of certain first-person self-ascriptions has no metaphysical implications. In particular, that phenomenon does not conflict with our preferred criterion of personal identity over time (which, in itself, is neither a first-person nor a third-person criterion, though it can be presented in either of these ways).

We should distinguish the question of how we arrive at various personal identity judgements (including first-person judgements), from the question of what makes such judgements true. These are different questions. It is quite consistent to suppose that access to certain first-person judgements is ungrounded or criterionless, even though the

truth-maker for such judgements involves essential reference to physical and psychological continuities (that is, criteria of personal identity).

The status of the 'as subject' use

In what sense, if any, is the 'as subject' use of 'I' basic? We can distinguish two senses in which this use might be basic. First, the 'as subject' use might be more basic or fundamental than the 'as object' use. Second, understanding the 'as subject' use might help to elucidate the sense in which 'I' -judgements are expressions of self-consciousness. I shall suggest that the 'as subject' use is basic in both senses.

The 'as subject' use is more basic than the 'as object' use

Here is one way in which the 'as subject' use of 'I' is more basic than the 'as object' use. Shoemaker writes: 'where "F" is a [material] predicate, to say that I am F is to say that my body is F. And if asked what it means to call a body "my body" I could say something like this: "My body is the body from whose eyes I see, the body whose mouth emits sounds when I speak, the body whose arm goes up when I raise my arm, the body that has something pressing against it when I feel pressure, and so on." All the uses of "I" that occur in this explanation of the meaning of the phrase "my body"…are…"as subject".'[14]

This argument, though correct in essentials, is slightly misleading. It is no part of the definition of the 'as object' use that all such occurrences of 'I' refer to the body. This is clear once we appreciate that 'as object' uses of 'I' can occur in mental self-ascriptions. But even in the case of physical self-ascriptions, it is controversial to suppose that the referent of 'I' is the body. If an utterance of 'I' referred to the body, it would follow that a person is identical to his body. Yet it was argued in Chapter 3 that such an identification is false. It is thus a substantial, and in my view false, metaphysical claim that a token of 'I' in, for example, 'I am six feet tall' refers to my body, rather than to a psychophysical substance which is spatially coincident with my body. (Think of analogous debates about whether or not a gold statue is numerically identical to the lump of gold which constitutes it.)

Shoemaker may simply have assumed that since we 'can't substitute for "I" a description of the body'[15] in 'I feel pain', so in physical self-ascriptions we *can* make such a substitution. That assumption should

be questioned. However, it is not required for the success of Shoemaker's argument, which does establish one way in which the 'as subject' use of 'I' is more basic than the 'as object' use.

Self-consciousness and the 'as subject' use of 'I'

What is the link between the 'as subject' use of 'I' and the concept of self-consciousness? It might seem excessive to claim that our concept of self-consciousness can be exhaustively characterised in terms of the 'as subject' features of 'I'. There is more to self-consciousness than that: a self-conscious being must be able to engage in reason and reflection, and consider itself as a being in time, with a past accessible in experience-memory and a future accessible in intention.

However, these features are possible only because of the 'as subject' features of 'I'. A being whose devices of 'self'-reference were not governed by the self-reference rule, and all of whose 'self'-ascriptions were grounded, could only think of himself as one object amongst others. His own features would be accessible only by inference or 'outer' perception. Such a being would be incapable of the sort of direct or ungrounded access to his past, present, and future that is constitutive of self-conscious thought.

This is confirmed when we reflect that the best explanation of why the 'A' -users are not self-conscious will make reference to the fact that all 'A' -judgements are grounded. The 'A' -users lack ungrounded self-knowledge. This observation, in turn, allows us to specify what it is for us to possess a first-person perspective on ourselves.

We now have an elucidation of the key phrase in Locke's definition of self-consciousness. Locke wrote that a self-conscious being 'can consider itself as itself, the same thinking thing, at different times and places'.[16] What is it for a creature to consider 'itself as itself'? It is for that creature to be a competent 'I' -user, and therefore to use 'I' in its 'as subject' use.

Interpreting Wittgenstein on avowals: reference, knowledge and authority

Reference

Nothing in Wittgenstein's remarks in the *Blue Book,* nor in my reconstruction of those remarks, implies that occurrences of 'I' in 'as subject' judgements (or avowals) are not referential. First, facts about recognition or identification are epistemological, and imply nothing

about the referentiality or non-referentiality of 'I'. Second, many features of the 'as subject' use of 'I' are shared by typical utterances of 'here' and 'now', yet such tokens are referential.

Hence, Anscombe's non-referential view of 'I' receives no support whatever from our reconstruction of Wittgenstein's 'as object'/'as subject' distinction. We are not forced to regard 'as subject' occurrences of 'I' as non-referential.

Indeed, such a conclusion would be at odds with Wittgenstein's official methodology. The first sentence of our quote from the *Blue Book* clearly states Wittgenstein's aim: not to criticise ordinary usage, but to combat misunderstandings to which the 'peculiar' grammar of the word 'I' is liable to give rise. One of these misunderstandings is the belief that certain uses of 'I' refer to an immaterial ego ('the one of which it was said, "Cogito, ergo sum"…').[17] Wittgenstein's antidote is to draw our attention to the possibility of these misunderstandings, and to make us see them as misunderstandings. His response is *not* to claim that a word which, by all relevant syntactic and semantic criteria, counts as referential, is really not referential.

It is also unjustified to attribute to Wittgenstein the so-called 'non-assertoric' thesis of avowals: the thesis that a typical utterance of, for example, 'I am in pain' is not an assertion.[18] Arguably, when Wittgenstein compares 'I have toothache' to a moan, his point is not that an utterance of that sentence fails to make an assertion. (Indeed, Wittgenstein says: 'it is impossible that in making the *statement* "I have a toothache"' (my emphasis).) His point is that in both cases (avowal and moan) it is impossible…to mistake another person for oneself.

Knowledge

Certain passages from Wittgenstein's *Philosophical Investigations* are often cited in support of the attribution to Wittgenstein of the so-called 'non-cognitive' thesis of avowals: the thesis that one cannot be said to know judgements of the form 'I am F', where the use of 'I' is 'as subject'. For example, Wittgenstein writes that 'I don't know whether I'm in pain' is not a 'significant proposition'[19]; that '[i]t can't be said of me at all (except perhaps as a joke) that I *know* I am in pain. What is it supposed to mean except perhaps that I *am* in pain.'[20]; and '[i]t is correct to say 'I know what you are thinking', and wrong to say 'I know what I am thinking'.[21] Surely, it might be thought, these passages justify the attribution to Wittgenstein of the non-cognitive thesis.

But matters are not so straightforward. One can interpret Wittgenstein's claim that an utterance of 'I know I'm in pain' is not

'significant' and that it is 'wrong to say "I know what I am thinking"', as drawing attention to the fact that such knowledge-claims have no practical use, that they make no move in the 'language-game' of reporting one's sensations or beliefs. Because they have no use, they are not 'significant'. Because they make no move in the language-game, yet—if uttered—present themselves as making such a move, it is misleading to say them.

In the quote from *Investigations,* Wittgenstein makes this clear: to be told, of some third party, 'he knows that he's in pain', where this is not intended as a joke, provides one with no more information than 'he's in pain'.[22] That is, no additional move is made by saying 'he knows that he's in pain' than is made by saying 'he's in pain'. One could make the same point about utterances such as 'I know I'm here', 'I know I exist', 'I know the time is now'. In all these cases, knowledge is not any sort of 'cognitive achievement'.[23] In such cases, the truth-conditions of the self-ascription ('I am F') and the truth-conditions for the corresponding knowledge-claim ('I know I am F') coincide. It is pointless to assert the knowledge-claim in preference to the self-ascription, and that is why the knowledge-claim sounds perverse.

It is plausible to suppose that Wittgenstein is advancing the view that the knowledge-claim presented in, for example, 'I know I'm in pain', does not involve any sort of 'cognitive achievement'. Thus his point is not that an utterance of 'I know I'm in pain' is a category mistake, or grammatically ill-formed, or meaningless. Such a view would be implausible since, to give just one reason, in suitable circumstances I can validly derive 'I know I'm in pain' from uncontroversial premises. For example, suppose I know that Jones knows that I am in pain; then, I know that it is known that I'm in pain. Applying the uncontroversial principle that 'I know that it is known that p' implies 'I know that p', it follows that I know I'm in pain.

Moreover, this non-cognitive view would sit ill with Wittgenstein's insistence that '[p]hilosophy may in no way interfere with the actual use of language; it can in the end only describe it…. It leaves everything as it is'.[24] Wittgenstein's point is that 'I know I'm in pain' is idle, not that it is illegitimate. After commenting that it is 'wrong to say "I know what I am thinking"', Wittgenstein comments in parenthesis: 'A whole cloud of philosophy condensed into a drop of grammar.'[25] The moral to be drawn from this remark is that we should defuse philosophical misunderstandings of the 'grammar' of first-person knowledge claims (for example, the idea that such knowledge-claims constitute cognitive achievements), and not criticise the linguistic construction itself.

Authority

We have attempted to characterise the 'as subject' use of 'I', and its importance for the concept of self-consciousness. Some philosophers think that a deep philosophical question remains about the explanation of the *authority* with which we credit utterances containing 'as subject' occurrences of 'I'. What explains the authority of an avower over the presence and character of, for example, some immediate mental state, an authority which one person cannot have over the mental state of another?

The Cartesian, of course, has a quasi-perceptual story to tell, involving infallible access to a private, inner realm. Such access constitutes a cognitive achievement. The expressivist also has a story to tell. According to this story, avowals are not descriptive of an 'inner' realm; consequently, the subject's authority over his avowals is not cognitive. But this is not to deny first-person authority. A normal spontaneous utterance of, for example, 'I have a headache' is the *expression* (not description) of a mental state, whose appropriateness—absent any obvious defeating conditions—is no more to be questioned than a child's cry.

Wittgenstein is often enlisted into the expressivist cause, and often on the basis the quoted passages from the *Blue Book* and §246, §408 and II ix of the *Investigations*. However, such conscription would be a mistake. Wittgenstein is advancing the view that the knowledge implicated in, for example, 'I know I'm in pain', is not any sort of 'cognitive achievement'. But it is still knowledge. It would be true, if pointless, to say 'I know I'm in pain'. On this interpretation, Wittgenstein is not an expressivist.

However, should we accept the aforementioned claim that the authority of avowals stands in need of substantial philosophical explanation? Is it not enough to observe that, presumably for evolutionary reasons, there is a remarkably reliable and effortless connection between, for example, a subject's being in pain and his belief that he is in pain? The authority we credit to avowals merely reflects the reliability of that connection. In contrast, the link between your pain and my belief that you're in pain is nowhere near as reliable: stoicism and pretence are always possible.

Conclusion

I have suggested that there is an important link between the fact that 'I' is a device of criterionless self-reference and the fact that 'I'-

judgements are manifestations of self-consciousness. It transpired that our use of 'I' in judgement requires that we have ways of knowing about ourselves which are 'as subject'. The best understanding of the 'as subject' way of knowing truths about oneself is in terms of ungroundedness.

I argued that nothing in Wittgenstein's remarks (or in my reconstruction of them) implies that 'as subject' uses of 'I' are non-referential or that judgements incorporating such uses are not genuine assertions. Further, Wittgenstein was right to deny that knowledge of one's avowals constitutes any sort of 'cognitive achievement'. Finally, I suggested that crediting people with first-person authority, to the extent that we do, reflects our acknowledgement of certain well-established intra-personal empirical regularities.

NOTES

1 THE PROBLEM AND ITS PLACE IN PHILOSOPHY

1 It is sometimes said that a question such as 'under what conditions is A identical to B?' ought to be as odd as the question 'under what conditions is A identical to A?'. But any oddness here is solely a result of the way the question is posed. We could have asked 'under what conditions does A continue to exist?', which is a perfectly sensible question. This is a question about identity, even though it does not use the word 'identity' (it is equivalent to 'under what conditions will there be something identical with A?').

2 D.Parfit, *Reasons and Persons,* Oxford, Oxford University Press, 1984, Part III.

3 J.Locke, *Essay Concerning Human Understanding* (ed.) W.Carroll, Bristol, Thoemmes, 1990, II xxvii 9.

4 See, for example, R.Descartes, *Meditations,* London, Penguin, 1978.

5 For a modern defence of dualism, see, for example, W.Hart, *The Engines of the Soul,* Cambridge, Cambridge University Press, 1988.

6 This happened when surgeons severed the *corpus callosum* in order to cure epilepsy. Communication between the two upper hemispheres of the brain was disrupted in order to minimise the frequency and intensity of epileptic fits. An unexpected side-effect was the fragmentation of consciousness within a single human being. (For a useful discussion, see T.Nagel, 'Brain Bisection and the Unity of Consciousness' in Nagel, *Mortal Questions,* Cambridge, Cambridge University Press, 1979.)

7 The principle at work here is Leibniz's Law, according to which x and y are identical only if they have all their properties in common. Since the statue and the lump do not have all their properties in common, it follows that they are not identical. (Note that the distinctness of the statue and the lump does not depend on there actually being a future meltdown. The statue and the lump differ over possession of the modal property *possibly existing in a melted state,* and this is enough to render them distinct.)

8 T.Nagel, *The View From Nowhere,* New York, Oxford University Press, 1986, p. 40.

9 ibid., p. 40.

10 D.Hume, 'Of Personal Identity', in *A Treatise of Human Nature* (ed.) L.A.Selby-Bigge, Oxford, Oxford University Press, 1978, and D.Parfit, *Reasons and Persons.*
11 Interestingly, the converse principle is not true. That is, it is not true that, if a person cannot survive or continue to exist without possessing a certain property, then possession of that property is essential to that person. If I was born on a Tuesday, then it will always be true of me that I was born on a Tuesday. My continuing to have this property is not merely accidental: I could not lose this property without ceasing to exist. Yet the property of having being born on a Tuesday is not part of my essence. It is contingent that I acquired this property; but, once acquired, I cannot then lose it.
12 J.Locke, *Essay* II xxvii 9.
13 L.Wittgenstein, *The Blue and Brown Books,* Oxford, Basil Blackwell, 1972, p. 62.
14 This example is due to Sydney Shoemaker, *Self-Knowledge and Self-Identity,* New York, Cornell University Press, 1963, ch. 1.
15 This thought-experiment is due to Daniel Dennett. For more detail, see his entertaining article 'Where Am I?', in D.Hofstadter and D.Dennett, *The Mind's I,* London, Penguin, 1981, pp. 217–30.
16 The examples of *Teletransportation* and *Branch-Line* are due to Derek Parfit. See his *Reasons and Persons,* Part III.
17 David Wiggins makes use of this thought-experiment as part of his defence of animalism. See *Sameness and Substance,* Oxford, Basil Blackwell, 1980, ch. 6, section 9.
18 This thought-experiment first appeared in the philosophical literature in David Wiggins' first book *Identity and Spatio-Temporal Continuity,* Oxford, Basil Blackwell, 1967.
19 D.Parfit, *Reasons and Persons.*
20 D.Parfit, 'Comments', *Ethics* vol. 96.4, 1986.

2 ANIMALISM AND REDUCTIONISM

1 David Wiggins, *Sameness and Substance,* p. 171.
2 As Wiggins concedes, ibid., pp. 171–2.
3 ibid, pp. 176–9.
4 ibid., pp. 176–9.
5 D.Hume, *Treatise,* p. 252.
6 See S.Kripke, *Naming and Necessity,* Oxford, Basil Blackwell, 1980, p. 50, for a discussion of this idea in the case of nations.
7 D.Hume, *Treatise,* for example, Section XIV.
8 D.Parfit, *Reasons and Persons,* p. 223.
9 See, for example, D.Pears, *The False Prison,* Oxford, Clarendon Press, 1988, p. 240, and D.Parfit, *Reasons and Persons,* p. 226.
10 See, for example, S.Shoemaker, 'Persons and their Pasts' in his *Identity, Cause and Mind,* Cambridge, Cambridge University Press, 1984, and D.Parfit, *Reasons and Persons,* ch. 11.
11 See J.McDowell, 'Reductionism and the First Person', in J.Dancy (ed.) *Reading Parfit,* Oxford, Basil Blackwell, 1997.

12 A further worry concerns whether q-memory is sufficiently memory-like to serve as a surrogate for memory. Is q-memory an acceptable surrogate, or just a delusion? Does the debate over q-memory illicitly presuppose an overly imagistic conception of 'personal' memory? (See Section IV of Arthur Collins' 'Personal Identity and the Coherence of Q-memory', *Philosophical Quarterly* vol. 47, 1997, pp. 73–80.)
13 See D.Wiggins, *Sameness and Substance,* pp. 24–7 and pp. 63–6, respectively.
14 D.Parfit, *Reasons and Persons.*
15 S.Shoemaker's 'Critical Notice of *Reasons and Persons', Mind* vol. 94, 1985, pp. 449–50.
16 D.Parfit, *Reasons and Persons,* p. 210.
17 ibid., p. 212.
18 ibid., p. 213.
19 D.Parfit, 'The Unimportance of Identity' in H.Harris (ed.) *Identity,* Oxford, Oxford University Press, 1995, pp. 36–7.
20 D.Parfit, *Reasons and Persons,* p. 260.

3 CRITERIA OF PERSONAL IDENTITY

1 J.Locke, *Essay* II xxvii 9, and II xxvii 17.
2 T.Reid, 'On Mr Locke's Account of Personal Identity' in J.Perry (ed.) *Personal Identity,* California, California University Press, 1975, p. 114.
3 S.Shoemaker, *Self-Knowledge and Self-Identity,* pp. 23–4.
4 See D.Dennett, 'Where Am I?', in *The Mind's I.*
5 See, for example, S.Kripke, *Naming and Necessity,* Lecture III, and H. Putnam 'Meaning and Reference', *Journal of Philosophy,* vol. 70, 1973, pp. 699–711.
6 See S.Kripke, op. cit., pp. 140–55.
7 Note, however, that the brain criterion and the strong version of the psychological criterion are not extensionally equivalent. They come apart, for example, in their descriptions of the thought-experiment presented in the opening pages of Bernard Williams' 'The Self and the Future' (in *Problems of the Self,* Cambridge, Cambridge University Press, 1982). We are to imagine a fabulous machine which can instantly change a person's whole psychology. Imagine that two people, A and B, 'swap' mental lives, so that A's stream of consciousness, memories, desires, etc. continues in the B-body, and B's stream of consciousness continues in the A-body. According to the brain criterion, A is the resulting A-body person. According to the strong version of the psychological criterion, however, A is identical to neither resulting person (neither stream of psychological continuity has its normal cause). Consequently, the two criteria are not equivalent.
8 B.Williams, 'The Self and the Future' in *Problems of the Self.*
9 ibid., p. 63.
10 Are persons essentially embodied? As far as the theory of personal identity endorsed here goes, 'No'. However, it may well be metaphysically impossible for our current mental life to be realised in an immaterial substance, for reasons that have nothing to do with the best theory of personal identity.

4 FISSION

1 Some philosophers have endorsed this second account of *Fission,* yet have denied that persons are Cartesian egos (for example, Colin McGinn in *The Character of Mind,* Oxford, Oxford University Press, 1982, ch. 6). On this view, the identity of a person over time is completely ungrounded. It is grounded neither in soul-identity, nor in physical or psychological continuities. However, this view is unintuitive, and alien to our normal practice of individuating and re-identifying persons.

2 See, for example, David Lewis 'Survival and Identity' in his *Philosophical Papers,* vol. 1, Oxford, Clarendon Press, 1983.

3 D.Parfit *Reasons and Persons,* especially chapters 12 and 13.

4 This objection assumes a rigid criterion of trans-world identity for four-dimensional objects, viz., that a given four-dimensional object could not have had more or less temporal parts than it actually has. This may seem implausible.

I agree that events and processes are four dimensional; parties, for example, are four-dimensional entities. And can't we truly say things such as 'Bill's party might have ended an hour earlier'? We can and we do. But if a four-dimensionalist wants to operate with a more relaxed criterion of transworld identity, he must tell us what it is, and what constraints it imposes. In which case, the present objection becomes a challenge.

5 Just as Lewis' account overestimates the number of pre-fission persons, so in other cases the four-dimensional view underestimates the number of entities in existence. Consider the following variant of Statue, *Harmony:* Statue and Bronze come into existence at the same time, and cease to exist at the same time. The four-dimensionalist must say that Statue is Bronze, since they share all and only the same 'temporal stages'. Yet surely this is wrong: Statue and Bronze differ in their modal properties, and so cannot be identical, even given their complete spatio-temporal coincidence. See A.Gibbard, 'Contingent Identity', *Journal of Philosophical Logic,* vol. 4, 1975, for an ingenious (though, I think, unconvincing) defence of the four-dimensionalist's verdict on this kind of example.

6 Plutarch *Lives,* sections 22–3. (Quoted on p. 92, n. 15, of D.Wiggins, *Sameness and Substance.*)

7 Note that even if we thought that the best answer to Plutarch's question was a negative one, the Theseus puzzle (discussed below) might still be with us. Almost everyone accepts that artefacts can survive replacement of *some* of their parts. What guarantee do we have that we cannot construct an appropriate competitor from only a subset of the parts of some larger working artefact, generating a Theseus-type puzzle?

8 T.Hobbes, *De Corpore,* part II, ch. II, in W.Molesworth (ed.) *The English Works of Thomas Hobbes,* London, John Bohn, 1839–15, vol. 1, p. 136. (Quoted on p. 92 of D.Wiggins, *Sameness and Substance.*)

9 'RC1' and 'RC2' are rigid designators. (See below.)

10 Nor is there any room for the four-dimensional 'multiple occupancy' response. That response is motivated in the case of a tie, as in *Fission*.

But in *Ship of Theseus* there is no tie: our dominant response is that Theseus' ship is the continuously repaired ship. On a four-dimensional view, we should simply regard the ship of Theseus and the continuously repaired ship as different stages of a single ship, and regard the re-constituted ship as a wholly distinct ship.

11 S.Kripke, *Naming and Necessity*, Lecture I.

12 We are not forced to regard the non-identity sentence 'Lefty is not Twin Lefty' as a 'bare' or 'ungrounded' non-identity. For all that has been said in the present chapter, we could see the existence of Righty as grounding this non-identity, though the grounding is extrinsic to Lefty.

5 IDENTITY AND VAGUENESS

1 A defender of the epistemic view of vagueness would reject this characterisation. According to that view, vague predicates do not lack sharp boundaries. Fred either is bald or he isn't, it's just that, since he's a borderline case, we can't know which side of the sharp dividing line he falls on. See T.Williamson, *Vagueness,* London, Routledge, 1994, for a thorough and ingenious defence of this unintuitive view.

2 'Can There Be Vague Objects?', *Analysis* vol. 38.4, 1978, p. 208.

3 Ibid., p. 208. ©

4 I assume that 'the first orange pen' is a singular term. This assumption is controversial. Many philosophers think that Russell's Theory of Descriptions showed that definite descriptions are not referring terms. (See B.Russell, 'On Denoting' in *Logic and Knowledge* (ed.) R.C.Marsh, London, George Allen and Unwin, 1956.) If so, then the sentence 'my pen is the first orange pen' would not be a sentence of identity. However, this linguistic question is irrelevant to the success or otherwise of Evans' proof. Kripke's proof of the necessity of identity does not require the assumption that proper names are not disguised descriptions; nor does Evans' proof require the assumption that definite descriptions are referring terms.

5 For a new twist on this predicate, see N.Feit, 'On a famous counterexample to Leibniz's Law', *Proceedings of the Aristotelian Society,* 1996, pp. 381–6. Feit emphasises the indexical nature of this predicate, and points out that there are contexts in which it would be true to say 'Barbarelli was so-called because of his size', or even 'Someone was so-called because of his size'.

6 See W.V.Quine, 'Reference and Modality', in *From a Logical Point of View,* New York, Harper and Row, 1963, p. 145.

7 See Lloyd Humberstone 'The Logic of Non-contingency', *Notre Dame Journal of Formal Logic* vol. 36, 1995, pp. 214–29.

8 D.Wiggins, 'On Singling Out an Object Determinately', in P.Pettit and J.H.McDowell (eds), *Subject, Thought and Context,* Oxford, Oxford University Press, 1986, p. 175.

9 See, for example, S.Kripke, 'Identity and Necessity', in M.K.Munitz (ed.) *Identity and Individuation,* New York, New York University Press, 1971.

10 In this respect, '∇' is analogous to an epistemic operator such as 'X believes that—'. Arguably, there is no class of singular terms such that co-referring members of that class are guaranteed to be substitutable *salva veritate* in a context generated by 'X believes that—'.

11 See the relevant, unduly neglected, passages in Williams' 'The Self and the Future' in *Problems of the Self,* pp. 58–61.

6 PARFIT AND 'WHAT MATTERS'

1 D.Parfit, *Reasons and Persons,* chapters 12 and 13.
2 ibid., p. 281.
3 ibid., ch. 15.
4 D.Parfit, 'Who do you think you are?', *The Times Higher* 1992.
5 D.Parfit, 'Comments', in *Ethics* vol. 96.4, 1986, pp. 832–72.
6 ibid., pp. 838–9.
7 ibid., p. 839.
8 D.Parfit, *Reasons and Persons,* ch. 12.
9 M.Johnston, 'Reasons and Reductionism', *Philosophical Review,* vol. 101.3, 1992, has a reply to the argument from fission. Unfortunately, Johnston's reply rests on the assumption that fission involves indeterminate identity. This claim was criticised in Chapter 4.

7 ANSCOMBE ON 'I'

1 G.E.M.Anscombe, 'The First Person' in *Metaphysics and the Philosophy of Mind,* Collected Papers vol. II, Oxford, Basil Blackwell 1981, pp. 21–36.
2 Is the common-sense view relevant to the more familiar Frege/Russell debate about the nature of content? In the case of first-person judgements, the debate concerns whether the content of X's judgement 'I am F' is the Fregean thought consisting of X's (private) sense of 'I' together with the sense of 'is F' *or* whether it is the Russellian proposition consisting of the person X together with the property of F-ness. (See G. Frege 'On Sense and Reference' and 'Thoughts' in G.Frege, *Collected Papers,* (ed.) B.McGuinness, Oxford, Basil Blackwell, 1984, and B. Russell, 'On the Nature of Acquaintance' and 'The Philosophy of Logical Atomism' in B.Russell, *Logic and Knowledge*.) If the indexical view is true, it's hard to see what room there could be for a private sense corresponding to each person's use of 'I'. For Frege, sense determines reference; yet, on the common-sense view, reference is fixed by the indexical rule.
3 Is this feature unique to 'I'? Surely, if Bill utters 'Bill is F', the second occurrence of the name 'Bill' must refer to Bill? But more than one person can be called 'Bill': the utterer may intend to refer to someone else called 'Bill'.
A further disanalogy is that Bill's knowledge that Bill exists is a posteriori, whereas his knowledge that he exists (the knowledge he would express by saying 'I exist') is a priori. Indeed, the latter piece of knowledge is a clear example of contingent a priori knowledge.
4 D.Hume, *Treatise,* p. 252.
5 G.E.M.Anscombe, 'The First Person', p. 32.
6 ibid., p. 31.
7 ibid., p. 24.
8 ibid., p. 24.

9 ibid., p. 36.
10 W.James, *Principles of Psychology* II London, 1901, p. 273. (Quoted by Anscombe on p. 36.)
11 G.E.M.Anscombe, op. tit, p. 30.
12 ibid., p. 30.
13 ibid., p. 30.
14 That is, it is only a contingent matter that a token of 'A', uttered in conformity with the observational criteria for its correct employment, refers to its utterer. This confirms the fact that 'A' is not governed by the self-reference rule. (Any term governed by that rule is immune to reference-failure or reference to anyone other than the speaker.)
15 ibid., p. 30.
16 ibid., p. 30.
17 ibid., p. 32.
18 ibid., p. 25.
19 It may be that Anscombe is relying on the Wittgenstein-inspired thought: if you can't go wrong, you can't go right either. But this thought is simply too blunt to do any work. A wide range of examples—indexicals ('here', 'now'), some demonstratives ('this red visual impression') and some self-referring terms ('these very words…') —count against the idea that 'guaranteed reference' is a self-contradiction.
20 G.E.M.Anscombe, 'The First Person', p. 27.
21 ibid., p. 26.
22 ibid., p. 27 and pp. 31–2, respectively.
23 ibid., p. 30.
24 ibid., pp. 30–1.
25 ibid., p. 31.
26 D.Hume, *Treatise,* p. 252.
27 G.E.M.Anscombe, 'The First Person', p. 26.
28 ibid., p. 33.
29 ibid., p. 33.
30 ibid., p. 34.
31 ibid., p. 34.
32 ibid., p. 35.
33 G.Evans, *The Varieties of Reference,* Oxford, Oxford University Press, 1983, ch. 7.
34 G.E.M.Anscombe, 'The First Person', p. 29.
35 ibid., p. 30.
36 See J.Katz 'Descartes' *Cogito'* in P.Yourgrau (ed.) *Demonstratives,* New York, Oxford University Press, 1990, esp. pp. 172–81, for some useful criticisms of Anscombe.
37 G.E.M.Anscombe, 'The First Person', p. 32.
38 ibid., p. 36.
39 Is there any illumination to be gleaned from looking at the first-person plural 'we'? There clearly is some analogy between 'I' and 'we'. They are both indexicals. But notice two disanalogies: first, 'we', like 'here' and 'now', can have variable reference ('we philosophers', 'we Australians', 'we humans', etc.); second, 'we' can admit of reference-failure (I say 'we' meaning to refer to all of us in the room, but in fact I

am surrounded by holograms). In these respects, 'we' differs from 'I'. More importantly, 'we' appears to presuppose 'I': a competent 'we' -user must be a competent 'I' -user, but not conversely. In which case, a study of the semantics of 'we' will not yield any independent insights into the concept of self-consciousness.

8 WITTGENSTEIN ON 'I'

1 L.Wittgenstein, *The Blue and Brown Books,* Oxford, Basil Blackwell, 1972, and *Philosophical Investigations,* Oxford, Basil Blackwell, 1978.
2 L.Wittgenstein, *The Blue and Brown Books,* pp. 66–7.
3 J. Locke*, Essay*, II xxvii 9.
4 G.E.M.Anscombe, 'The First Person', pp. 25–6.
5 D.Hume, *Treatise,* p. 252.
6 This conceptual irreducibility should not be taken to imply that the phenomenon of self-consciousness is ontologically irreducible. Even if the concept of self-consciousness is not reducible to some range of physical concepts, there is no reason why purely physical systems cannot be self-conscious.
7 L.Wittgenstein, *Blue Book,* pp. 66–7.
8 ibid., p. 69.
9 It would be wrong to infer that 'as object' occurrences of 'I' can occur only in physical self-ascriptions. Occurrences of 'I' in dispositional psychological self-ascriptions— 'I am courageous', 'I am intelligent', etc. —should be classified 'as object'.
10 Demonstrative knowledge (for example, the knowledge expressed by a typical perceptual demonstrative judgement 'that man is bald') is not the result of inference, yet it evidently involves 'recognition' or 'identification' of its object. Hence the need for the second disjunct.
11 See Andy Hamilton, 'A New Look at Personal Identity', *Philosophical Quarterly* vol. 45, 1995, pp. 332–49. For a reply to some of the claims of that paper, see my 'Hamilton's New Look: a Reply', *Philosophical Quarterly* vol. 46, 1996, pp. 220–6.
12 G.Evans, *The Varieties of Reference,* Oxford, Oxford University Press, 1983, ch. 7.
13 See P.F.Strawson, *Individuals,* London, Methuen, 1959.
14 S.Shoemaker, 'Self-Reference and Self-Awareness' in *Identity, Cause and Mind,* Cambridge, Cambridge University Press, 1984, p. 18.
15 L.Wittgenstein, *Blue Book,* p. 74.
16 J.Locke, *Essay,* II xxvii 9.
17 L.Wittgenstein, *Blue Book,* p. 69.
18 Note that the no-reference thesis provides no automatic support for the non-assertoric thesis. The occurrence of 'it' in 'it's raining' is non-referential, yet an utterance of 'it's raining' is a genuine (i.e., truth-evaluable) assertion. For a useful discussion of these and related issues, see P.M.S.Hacker, *Insight and Illusion,* Oxford, Oxford University Press, 1972, ch. IX.
19 L.Wittgenstein, *Philosophical Investigations,* §408.
20 ibid., §246.
21 ibid., II xi.

22 ibid., §246.
23 See P.Boghossian, 'Content and Self-Knowledge', *Philosophical Topics* vol. XVII.1, 1989, pp. 5–27
24 *Philosophical Investigations,* §24.
25 ibid., §124.

BIBLIOGRAPHY

Adams, R.M. (1990) 'Should Ethics Be More Impersonal?', *Philosophical Review* vol. 98.

Anscombe, G.E.M. (1981) 'The First Person', in *Metaphysics and the Philosophy of Mind,* Oxford, Basil Blackwell.

Boghossian, P. (1989) 'Content and Self-Knowledge', *Philosophical Topics,* vol. XVII. 1.

Cambell, J. (1994) *Past, Space and Self,* Cambridge, MIT Press.

Collins, A. (1997) 'Personal Identity and the Coherence of Q-memory', *Philosophical Quarterly* vol. 47.

Dancy, J. (ed.) (1997) *Reading Parfit,* Oxford, Basil Blackwell.

Descartes, R. (1978) *Discourse on Method and the Meditations,* London, Penguin.

Evans, G. (1978) 'Can There Be Vague Objects?', *Analysis* vol. 38. 4.

——(1983) *The Varieties of Reference,* Oxford, Oxford University Press.

Feit, N. (1996) 'On a famous counterexample to Leibniz's Law', *Proceedings of the Aristotelian Society.*

Gibbard, A. (1975) 'Contingent Identity', *Journal of Philosophical Logic* vol. 4.

Hacker, P.M.S. (1972) *Insight and Illusion,* Oxford, Oxford University Press.

Hamilton, A. (1995) 'A New Look at Personal Identity', *Philosophical Quarterly,* vol. 45, pp. 332–49.

Harris, H. (ed.) (1995) *Identity,* Oxford, Oxford University Press.

Hofstadter, D. and Dennett, D. (eds) (1981) *The Mind's I,* London, Penguin, 1981.

Hume, D. (1978) *A Treatise of Human Nature,* (ed.) L.A.Selby-Bigge, Oxford, Oxford University Press.

Johnston, M. (1992) 'Reasons and Reductionism', *Philosophical Review,* vol. 101. 3.

Katz, J. (1990) 'Descartes' *Cogito',* in P.Yourgrau (ed.) *Demonstratives,* New York, Oxford University Press.

Kripke, S. (1980) *Naming and Necessity,* Oxford, Basil Blackwell.

Lewis, D. (1983) 'Survival and Identity' in *Philosophical Papers,* vol. 1, Oxford, Oxford University Press.

Locke, J. (1965) *Essay Concerning Human Understanding,* (ed.) J.W.Yolton, London.

McGinn, C. (1982) *The Character of Mind,* Oxford, Oxford University Press.

Nagel, T. (1979) *Mortal Questions,* Cambridge, Cambridge University Press.

——(1986) *The View from Nowhere,* Oxford, Oxford University Press.
Noonan, H. (1989) *Personal Identity,* London, Routledge.
Nozick, R. (1981) *Philosophical Explanations,* Cambridge, Harvard University Press.
O'Brien, L. (1994) 'Anscombe and the Self-Reference Rule', *Analysis,* vol. 54. 4, pp. 277–81.
Parfit, D. (1984) *Reasons and Persons,* Oxford, Oxford University Press.
——(1986) 'Comments', *Ethics* vol. 96. 4.
——(1992) 'Who do you think you are?', *The Times Higher.*
——(1995) 'The Unimportance of Identity', in H.Harris (ed.) *Identity,* Oxford, Oxford University Press.
Perry, J. (1977) 'Frege on Demonstratives', *Philosophical Review,* vol. lxxxvi, pp. 474–97.
——(1979) 'The Essential Indexical', *Nous* vol. xiii, pp. 3–21.
Pettit, P. and McDowell, J.H. (eds) (1986) *Subject, Thought and Context,* Oxford, Oxford University Press.
Reid, T. (1975) 'On Mr Locke's Account of Personal Identity', in J.Perry (ed.) *Personal Identity,* California, California University Press.
Shoemaker, S. (1963) *Self-Knowledge and Self-Identity,* New York, Cornell University Press.
——(1984a) 'Self-Reference and Self-Awareness', in his *Identity, Cause and Mind,* Cambridge, Cambridge University Press.
——(1984b) 'Persons and their Pasts', in his *Identity, Cause and Mind,* Cambridge, Cambridge University Press.
——(1985) 'Critical Notice of *Reasons and Persons', Mind,* vol. XCIV.
Snowdon, P. (1990) 'Persons, Animals and Ourselves', in C.Gill (ed.) *The Person and the Human Mind,* Oxford, Clarendon Press.
——(1991) 'Personal Identity and Brain Transplants', in D.Cockburn (ed.) *Human Beings,* Cambridge, Cambridge University Press.
Strawson, P.F. (1959) *Individuals,* London, Methuen.
Unger, P. (1990) *Identity, Consciousness and Value,* Oxford, Oxford University Press.
Wiggins, D. (1967) *Identity and Spatio-Temporal Continuity,* Oxford, Basil Blackwell.
——(1980) *Sameness and Substance,* Oxford, Basil Blackwell.
Williams, B. (1982) *Problems of the Self,* Cambridge, Cambridge University Press.
Williamson, T. (1994) *Vagueness,* London, Routledge.
Wittgenstein, L. (1972) *The Blue and Brown Books,* Oxford, Basil Blackwell.
——(1978) *Philosophical Investigations,* Oxford, Basil Blackwell.

INDEX