INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®
SURVIVING IN A DIFFERENT BODY:
AN EXAMINATION AND REJECTION OF
THE LOCKEAN THESIS
THAT CONSCIOUSNESS ALONE MAKES PERSONAL IDENTITY

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By
Alfred C. Lent, M.A.

The Ohio State University
2001

Dissertation Committee
Approved by
Professor William Taschek, Advisor
Professor Donald Hubin
Professor George Pappas

Department of Philosophy
ABSTRACT

The intuition that a person can survive in a different body has a long philosophical pedigree. It finds support not only in substance dualist views like Descartes' but also in psychological views of personal identity such as those of John Locke and Derek Parfit. The defense of this intuition by the psychological theorists relies on a claim first stated by Locke that "consciousness alone makes personal identity." I consider four ways of understanding this thesis: the one traditionally attributed to Locke; a different one I argue is actually found in Locke; an account traditionally attributed to Hume; and the one defended by Parfit. I argue that each should be rejected based upon Joseph Butler's influential—and, I argue, widely misunderstood—challenge to psychological accounts of personal identity of the sort offered by Locke and his contemporary Henry Dodwell.

I show that Butler's most important argument against his contemporaries relies not on assumptions about immaterial substances, as is usually supposed, but on the empirical distinction between self-love and benevolence, a distinction that forms the basis of his enduring refutation of psychological egoism. I argue that the distinction between self-love and benevolence is one that must be respected by a correct metaphysics of persons and that psychological views cannot do so. The problem is not, as is usually supposed, that self-love requires the existence of a simple subject of experience, but, rather, that both self-love and benevolence are mental states whose content includes bodies with a particular history. In other words, psychological views
fail because the person-centered mental events prized by the theory include bodily content to which psychological views are, by their very nature, not entitled. I believe that this shows not only that no purely psychological theory can succeed, but also that the correct metaphysics of persons must include the body as an essential feature of persons.
To Christine and Christian
ACKNOWLEDGMENTS

I wish to thank the three tenacious philosophers who served on my dissertation committee: William Taschek, who oversaw this project from the beginning and never tired of following arguments to their end; Don Hubin, for his penetrating insights and for the many hours of attention he gave my work; George Pappas, whose work and comments on my work helped me to see the beauty and value of exegesis.

I wish to thank Dan Farrell, Stewart Shapiro, Robert Kraut, Diana Raffman, Kathleen Schmidt, Justin D’Arms.

I also wish to thank my fellow graduate students: Jeff Koperski, Joe Salerno, Jon Cogburn, Mark Silcox.

I also wish to thank my friends, the Mortons and the O’Briens, who gave to me ceaselessly.

I wish to acknowledge a deep debt of gratitude to my parents, Flo and Charlie Lent, whose example as hard workers has set a standard to which I continually aspire.

Finally, I want to thank my wife, Christine, who fought for me, loved me, and sacrificed for me; and my son, Christian, who was just a child when I began, and is now my friend and confidant.
VITA

March 11, 1959 ............................................................ Born - Bayshore, New York

1991 .............................................................................. B.A. Philosophy, Liberty University.

1991-1998 ..................................................................... Graduate Teaching Associate,
The Ohio State University

1998-1999 ...................................................................... Lecturer,
Columbus College of Art and Design

1998-1999 ...................................................................... Lecturer,
Marietta College

1999-present .................................................................... Visiting Lecturer,
Wittenberg University

2000 ............................................................................... Lecturer,
College of Wooster

2000 ............................................................................... Lecturer,
Capital University

Major Field: Philosophy
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Abstract</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td></td>
</tr>
<tr>
<td>The Body-swap Intuition</td>
<td>1</td>
</tr>
<tr>
<td>2. Locke's CAT</td>
<td>11</td>
</tr>
<tr>
<td>2.1 Identity and Substance</td>
<td>12</td>
</tr>
<tr>
<td>2.2 Personal Identity</td>
<td>14</td>
</tr>
<tr>
<td>2.3 “Person” as a “Forensic Term”</td>
<td>22</td>
</tr>
<tr>
<td>2.4 A Conflict between CAT and our Norms</td>
<td>24</td>
</tr>
<tr>
<td>3 Dodwell's CAT and its Problems</td>
<td>30</td>
</tr>
<tr>
<td>3.1 Dodwell</td>
<td>30</td>
</tr>
<tr>
<td>3.2 Clarke's criticisms of CAT</td>
<td>31</td>
</tr>
<tr>
<td>3.3 Butler and Reid</td>
<td>42</td>
</tr>
<tr>
<td>3.4 Butler: circularity, interest, and overlap</td>
<td>44</td>
</tr>
<tr>
<td>3.5 Summary</td>
<td>58</td>
</tr>
<tr>
<td>4. Hume and CAT</td>
<td>62</td>
</tr>
<tr>
<td>4.1 The traditional interpretation</td>
<td>63</td>
</tr>
<tr>
<td>4.2 A problem for the traditional interpretation</td>
<td>66</td>
</tr>
<tr>
<td>4.3 An alternative interpretation</td>
<td>71</td>
</tr>
<tr>
<td>4.4 Rationality and the pride of the philosophers</td>
<td>76</td>
</tr>
<tr>
<td>4.5 Summary and conclusions</td>
<td>78</td>
</tr>
</tbody>
</table>
CHAPTER 1
THE BODY-SWAP INTUITION

There is a common intuition that a person can exist in a different body if somehow or other the mental life of a person is transferred from one body to another. Imagine that I wake one morning and, as is my routine, I stare at the ceiling, reflect on yesterday's events, make my plans for the day, and finally, leave my bed to carry out my plans for the day. Imagine this routine is interrupted when I glance at the mirror and behold a different face staring back at me. I examine the other limbs and parts that I have been associated with since I awoke this morning and this confirms my fears. There can be no question: This is not the body I thought I remembered going to bed with.

It seems to make sense, at least at first, for me to think of myself existing in a different body. If we are supposing that I have the mental life of the one whose past I seem to remember, then I would know things that the one whose past I seem to remember knew. I could try to convince my loved ones that I am their loved one on that basis. I would know details about the past that no one else, it would seem, could conceivably know. It would seem that, were I to find myself in this situation, I could make at least a prima facie case that it is I, the loved one of my loved ones, existing in a different body. Call the intuition described above the Body-swap Intuition:

The Body-swap Intuition: A person can exist in a different body if somehow or other the mental life of a person is transferred to a different body.
As I have stated it, the Body-swap Intuition is a claim that emerges from the hurly-burly of commonsense. It is not a philosophical view but, rather, it is an inchoate belief that many either have or, if prompted by a clever yarn, will readily assent to.

The Body-swap Intuition can be broken down into two claims that have found wide acceptance among philosophers. First, since it claims that a person can exist in a different body, it says that sameness of body is not a necessary condition for the existence of a person at different times. Second, since it identifies the mental life as the locus of personhood it says that sameness of mental life is a sufficient condition for the existence of a person.

Locke, for example, accepts both of these claims. He imagines that the soul of a prince carrying with it the prince’s “consciousness”—Locke’s term for one’s mental life—departs the body of the prince and enters the body of a cobbler. He says that the resulting person would be “the same Person with the Prince,” who now exists in a different body.¹ Locke says that what makes it the case that the resulting person is the same person as the prince is that the same consciousness has been preserved at different times. Sameness of consciousness is therefore sufficient for the existence of a person at different times. Locke, therefore, can be said to accept the Body-swap Intuition.

So too can a traditional dualist view such as Descartes’. On a dualist view, a person’s existence through time consists in the numerical identity of a simple immaterial substance at different times. On this view, body and soul are distinct substances that are only contingently related. Descartes says, "I am truly distinct from my body, and that I can exist without it."² He believes he could exist without his body,


². *Meditations on First Philosophy*, 78. Dualist views directed specifically at the
which means that he doesn't take his body to be necessary for his continued existence. If one could continue to exist without one's body, it is no great leap to imagine that one could continue to exist even if one were associated with a different body. If I awoke in a different body because my soul had transmigrated, the dualist view implies that I am the same person as the one whose mental life I now have.

Some have defended the claim that a person's existence through time consists in the continued existence of the same functioning brain. This too implies the Body-swap Intuition. If I awoke in a different body because my brain had been transplanted into a different body, this view implies I am the same person as the one whose past I seem to remember.

Both the dualist view and the brain-alone view imply the Body-swap Intuition, but not without qualification. On both of these views, sameness of consciousness is sufficient for the existence of a person at different times, but sameness of consciousness requires sameness of a substance, an immaterial substance for the dualist and a brain for the brain-alone theorist.

Suppose, however, that I awoke in a different body, not because my soul had transmigrated, nor because my brain had been transplanted, but because the brain of the


one whose past I seem to remember was scanned by a device that recorded its states and then reproduced these states in a different brain. On neither of these views would I be the same person as the one whose past I seem to remember. I shall say that these two views imply the Qualified Body-swap Intuition because, while they imply the two claims of the Body-swap Intuition, they also imply that sameness of consciousness requires sameness of a substance. The Qualified Body-swap Intuition, therefore, says the following:

   The Qualified Body-swap Intuition:

   1. Sameness of body is not a necessary condition for the existence of a person at different times.

   2. Sameness of consciousness is sufficient for the existence of a person at different times.

   3. Sameness of substance is necessary for sameness of consciousness.

Some have accepted the two claims of the Body-swap Intuition, while denying that sameness of substance is necessary for sameness of consciousness. I shall say that these views imply the Unqualified Body-swap Intuition, which says the following:

   The Unqualified Body-swap Intuition

   1. Sameness of body is not a necessary condition for the existence of a person at different times.

   2. Sameness of consciousness is sufficient for the existence of a person at different times.

   3. Sameness of substance is not necessary for sameness of consciousness.
The difference between the Qualified Body-swap Intuition and the Unqualified is the third condition. According to the Qualified Body-swap Intuition, sameness of substance is necessary for sameness of consciousness; according to the Unqualified Body-swap Intuition, it is not.

If the Unqualified Body-swap Intuition is true, what should we say if I awoke with a different body in virtue of the device I described earlier? What should we say of a case in which no body and nothing that we can conceivably call a substance is the same, but only a mental life or consciousness alone?

Henry Dodwell, a contemporary of Locke, and an early proponent of Locke’s view of personal identity, would say I am the same person as the one whose past I seem to remember. He accepts the three claims that constitute the Unqualified Body-swap Intuition, and he believes the existence of a person through time to be best describable in terms of sameness of person or personal identity. He regards sameness of consciousness alone as necessary and sufficient for personal identity. He would therefore take the present case as one in which it is correct to say that I am the same person as the one whose past I seem to remember, even if neither the brain, the body, nor the soul are the same.

Derek Parfit, a contemporary philosopher, also accepts the three claims of the Unqualified Body-swap Intuition. He believes that a person’s existence at different times consists in relations among mental events, which is his way of describing what I have been calling consciousness or a mental life.

There is, however, an important difference between Dodwell’s view and Parfit’s. While Dodwell believes that consciousness alone is necessary and sufficient
for personal identity, Parfit claims only that consciousness is necessary and sufficient for survival, a relation that has only some of the properties of identity. Let me explain one difference between identity and what Parfit calls survival.

We can imagine two variations on the case in which consciousness may be thought to be preserved but neither brain, body, nor soul is the same. We can imagine that the device I described earlier does its magic either by destroying the original brain or by leaving it in tact, functioning as it had been in the original body. On the first variation, the relation between the one whose past I seem to remember and me is one-one: I am the only one who has the same consciousness as the one whose past I seem to remember. On the second, consciousness has taken a branching form: There are two of us who seem to remember one person’s past. Parfit says that in this second case, we cannot say that the one whose past I seem to remember is identical to both of us, because identity is a relation that is, among other things, a one-one relation. According to Parfit, although we cannot say that the one person is identical with the two of us, we can say that he has survived as both of us.

Both Dodwell and Parfit accept the three claims of the Unqualified Body-swap Intuition. An important difference between Dodwell’s view and Parfit’s is that Dodwell describes the existence of a person at different times in terms of personal identity or sameness of person, while Parfit believes the existence of a person at different times is best described in terms of survival.4

Locke, Dodwell, and Parfit can all be said to hold a thesis I call the Consciousness Alone Thesis (CAT), which says that a person’s continued existence consists in consciousness alone. I have already mentioned that Dodwell describes a person’s existence through time as personal identity. He defends his version of CAT as

4. I discuss the significance of this distinction in chapter 5.
a claim about personal identity. He argues that personal identity consists in consciousness alone. Parfit defends the claim that a person's existence through time, best described as survival, not identity, consists in relations among mental events. Either version implies that the Unqualified Body-swap Intuition is true.

There are other ways CAT has been understood. In chapters 2 and 4, I consider two versions of CAT that do not imply the Unqualified Body-swap Intuition. In chapter 2, I consider Locke's claim that personal identity consists in consciousness alone. I show that while Locke does accept CAT, he withholds on the claim that distinguishes the Qualified Body-swap Intuition from the Unqualified Body-swap Intuition. Locke's version of CAT, I argue, does not imply the Unqualified Body-swap Intuition, but is consistent with either version.

In chapter 4, I consider two interpretations of Hume. On the traditional interpretation of Hume, it is always a mistake to say that a person is the same at different times, that ascriptions of personal identity are always false. On a second interpretation, Hume provides a corrected account of personal identity attributions. On either interpretation, Hume defends a version of CAT, but the traditional interpretation of Hume's CAT, I argue, implies that the Unqualified Body-swap Intuition is false.

Each of the versions of CAT I consider can be shown to conflict in an important way with quite ordinary commonsense claims about praiseworthiness and blameworthiness. Let me explain.

The main writers I consider believe that ordinary commonsense claims about praiseworthiness and blameworthiness imply claims about personal identity. For example, commonsense holds that if Hauptmann is guilty of murdering the Lindbergh baby, then Hauptmann and the murderer are the same person. This implication seems to hold whether we are talking about moral guilt or legal guilt. If Hauptmann is not the
same person as the murderer, it would surely be a miscarriage of justice to punish Hauptmann for murdering the Lindbergh baby, either with the force of law or with the force of public opinion. Not only do claims about moral guilt and legal guilt have this result, claims about desert quite generally have this result. If Lindbergh is praiseworthy for having crossed the Atlantic, it seems to follow that Lindbergh and the one who crossed the Atlantic are the same person.

The details vary from writer to writer. Locke, for example, believes that X is the same person as Y if and only if X is justly rewarded or punished for Y's actions. Hume believes that claims about pride and shame, among other claims, imply claims of personal identity.

On each version of CAT I consider, there is a prima facie incompatibility between CAT and some commonsense notion of praiseworthiness and blameworthiness recognized by the one promoting CAT. For example, on the traditional interpretation of Hume's discussion of personal identity, claims of the form "X is the same person as Y" are always false. Hume believes that claims of the form "X is proud" and "X is ashamed" imply a claim about personal identity. Together, these two commitments imply that claims of the form "X is proud" and "X is ashamed" are always false, which is something that Hume does not accept.

These writers all accept that some commonsense normative claims such as "X is guilty," "X is praiseworthy," "X is proud," or "X is ashamed" imply claims of personal identity. I claim that CAT conflicts in some important way with quite ordinary normative judgments, in many cases judgments made by the proponent of CAT.

To show that there is a conflict, however, even an essential conflict, does not, in itself show that we should reject CAT. Suppose, for example, that Hume holds the view that the traditional interpretation attributes to him. Suppose further that it is right
to say that by claiming that Hauptmann is guilty of the murder, we are thereby making an ascription of personal identity. If all such ascriptions are false, however, so is the claim that Hauptmann is guilty of the murders. The trouble is, not only is Hauptmann off the hook, so is everyone else. No statement of the form “X is guilty for having done some act A” is true if such claims imply a claim about personal identity. Claims of the form “X is praiseworthy for having done act A,” which also seem to imply a claim about personal identity would also be false. X is praiseworthy for having done action A only if X and the one who did A are the same person. If personal identity attributions are always false, so are claims about praise worthiness. Well then, if claims of personal identity are always false, it seems that we are neither guilty of anything nor praiseworthy for anything.

During Hume’s time such a conflict would have been seen as a conclusive reason to reject Hume’s view of personal identity. In our time, Derek Parfit exploits such conflicts to argue that we ought to revise our moral theories and our theories about rationality. He argues that we have false beliefs about ourselves and about our existence through time, and that when we see the truth, we should revise our moral theories and our theories about rationality.\(^5\)

In chapter 5, I consider the contemporary CAT. I raise two arguments urging that we reject the contemporary CAT. The first is that the contemporary CAT implies that sympathy is justified and self-love is not. I argue that this is absurd, which implies that the contemporary CAT is false. The second is that the contemporary CAT implies that neither self-love nor benevolence are justified. I argue that because self-love and

---

5. Parfit, x.
benevolence constitute our paradigm for what rational action is, self-love and benevolence are justified if anything is. Therefore, the contemporary CAT ought to be rejected.

The contemporary CAT says that a person's continued existence consists in consciousness or R-relatedness alone. It is false, therefore, that a person's existence consists in R-relatedness alone. It is also false, then, that the conjunction of claims 2 and 3 of the Unqualified Body-swap Intuition is false. That is, it follows from the falsity of the contemporary CAT that the Unqualified Body-swap Intuition is false as well. If it is not the case that a person's existence consists in consciousness alone, it is also false that one can exist in a different body if consciousness is preserved but a substance such as a brain or a soul is not.

In chapter 6, I draw my main conclusions from the discussions of the first five chapters. I argue, first, that no version of CAT is consistent with commonsense notions about self-love and benevolence. Second, I argue that the clash arises because CAT must leave out information that is vital to these notions. I draw two more conclusions, one follows immediately, the other, is more tentative than the others. The first conclusion is that if we grant that consciousness is sufficient for a person's existence through time, other things are necessary. I argue that if consciousness is sufficient for a person's existence through time because of the access consciousness gives to the past through memory and the access it gives to the future through intentions, then the body is necessary as well, because the body is part of the content of many essentially self-regarding memories and intentions and is therefore, part of the content of our self-regarding and other-regarding aims.
CHAPTER 2
LOCKE'S CAT

Locke argues that personal identity consists in consciousness alone. In 2,27,9, he says, "In this [consciousness] alone consists personal Identity." In 2,27,10 he says, "For it being the same consciousness that makes a Man be himself to himself, personal Identity depends on that only."\(^1\)

In this chapter, I try to answer two questions and to draw some conclusions about the ramifications of the answers. The first question is, what does Locke mean when he says that personal identity consists in consciousness alone? The second question is, does Locke take his view to imply the Unqualified Body-swap Intuition? I argue that Locke does believe that consciousness alone makes our concept or idea of personal identity, but he withholds on the question of whether sameness of substance is necessary for sameness of consciousness. He, therefore, does not accept the Unqualified Body-swap Intuition.

While Locke withholds on whether it is true that sameness of substance is necessary for sameness of consciousness, he does not believe that it is absurd to suppose otherwise. Since this claim marks the distinction between the Qualified Body-swap Intuition and the Unqualified Body-swap Intuition it is worth looking at the

\(^1\) All references to Locke are from An Essay concerning Human Understanding, Peter H. Nidditch, ed. (Oxford: Oxford University Press, 1975). Punctuation, spelling, and italics are as in the original.
problems that arise for Locke on the assumption that sameness of substance is not necessary for sameness of consciousness and describing the options available to him.

2.1 Identity and Substance

I begin by considering the issues of identity and substance as they arise in Book 2, Chapter 27 of the *Essay*, the chapter that includes Locke's discussion of personal identity. I then turn specifically to Locke's discussion of personal identity.

Locke opens 2,27,2 by identifying three sorts of substances, which he calls God, finite intelligences, and bodies. Because Locke thinks of God as immutable, questions of identity with respect to God do not arise. Locke brushes the question of God's identity aside without further comment. The distinction between substances that are finite intelligences and substances that are bodies seems to be the distinction between immaterial substances—which Locke sometimes calls spirit or soul (2,23,5; 2,27,14)—and material bodies, whether inanimate bodies, such as an atom or a mass of molecules, or living bodies, such as an oak, a horse, or a man. A finite intelligence can either be disembodied, as it is with angels (4,3,28), or embodied as it is with humans.

By calling those things above "substances," Locke means to distinguish them from modes. The distinction Locke makes between substances and modes corresponds roughly to the distinction between substance and property, except that usually Locke is interested in our ideas of these things rather than the things themselves. A substance is distinguished from a mode in that our ideas of substances are ideas of things "subsisting by themselves" (2,12,6) while modes have no independent existence apart from their existence as "Dependences on, or Affections of Substances" (2,12,4). A man

is a substance, according to Locke, since it is a particular thing subsisting by itself, while gratitude is a mode since it is dependent on a substance; gratitude has no existence independent of a particular grateful thing.

In describing bodies—meaning material bodies—Locke considers an atom, a collection of atoms, and various members of natural kinds of living things. He distinguishes between those things or substances whose identity consist in sameness of substance—where by “substance” he means material parts—and those things or substances whose identity does not consist in sameness of substance. He argues, "'Tis not therefore Unity of Substance that comprehends all sorts of Identity . . ."(2,27,7).

The identity of an atom and the identity of a collection of atoms—what Locke calls "a Mass"—consists in sameness of substance, or "same particles" (2,27,3) while the identity of living creatures such as an oak, a horse, or a man "depends not on a Mass of the same Particles" (2,27,3). For Locke, this means that the identity of an atom and the identity of a mass consists in sameness of substance, while the identity of living things does not.

Living things, for Locke, are successions of distinct substances. The mass of molecules or substance that constitutes, for example, an oak at one time is not the same mass at any other time. Still, Locke believes “An Oak, growing from a Plant to a great Tree, and then lopp’d, is still the same Oak” even though he believes that the plant and the great tree are composed of different substances (2,27,3). What makes an oak the same oak if not sameness of substance? Locke’s answer is that they are the same if these distinct substances “partake of one Common Life” (2,27,4).

Locke mentions this notion of distinct substances partaking of one common life a number of times, but he never elaborates. He says, for example, that the identity of living things at different times consists in "nothing but a participation of the same
continued Life by constantly fleeting Particles of Matter, in succession vitally united to the same organized Body" (2,27,6). We can say that Locke holds that for a living thing such as an oak, X is the same oak as Y if and only if the mass of molecules that makes X and the mass of molecules that make Y partake of the same life. That leaves us with questions about what life is and under what conditions life can be said to be the same at different times.

An animal, like a plant, is "a living organized Body" (2,27,8). The mass of molecules that constitute it one time is different from the mass of molecules that constitute it at another time. The same animal, for Locke, consists in "the same continued Life communicated to different Particles of Matter, as they happen successively to be united to that organiz’d living Body" (2,27,8). A man, therefore, is nothing but "an Animal of such a certain form" (2,27,8). Therefore, the identity of the same Man consists in "nothing but a participation in the same continued life" (2,27,6).

2.2 Personal Identity

Beginning in section 9, Locke turns to personal identity. He says that a person is "a thinking intelligent Being that has reason and reflection," thus distinguishing a person from a man, which is nothing but the name of a certain kind of an animal (2,27,9; 2,27,6). This distinction follows from the fact that a man is an animal that comes into existence when a human embryo comes into existence (2,27,6) and a person is a thinking thing and therefore comes into existence when thinking begins (2,27,9).

An essential feature of a person is that it can "consider it self as it self" (2,27,9). I can think of myself as myself as I am now, as I once was by remembering, and, in a sense, as I shall be by forming intentions (2,27,10). These mental activities that distinguish me as a person are modes of thinking, with consciousness an essential mode
of thinking. Locke believes that one engages in these mental activities "only by that consciousness, which is inseparable from thinking, and as it seems to me, essential to it" (2,27,9).

Locke is making reference to a claim he has argued for at some length earlier in the Essay against the Cartesian position that holds the soul always thinks. (2,1,11-19). Locke argues there, and reaffirms here, that one cannot perceive without perceiving that one does. Since, as Locke says,

Consciousness always accompanies thinking, and 'tis that, that makes every one to be, what he calls self, and thereby distinguishes himself from all other thinking things, in this alone consists personal Identity, i.e. the sameness of rational Being (2,27,9).

Consciousness, then, is an essential property of thinking, and it is consciousness, "the perception of what passes in a Man's own mind" (2,1,19), that distinguishes one person from another at a time. I perceive what passes through my mind, not any one else's and that is what it is that makes me me, and everyone else not me.

From this he draws a conclusion about personal identity. He says, "as far as this consciousness can be extended backwards to any Action or Thought, so far reaches the identity of that Person" (2,27,9). The claim Locke sets out to defend, then, is that personal identity consists in consciousness alone.

He argues that sameness of man, does not imply sameness of person. If there were a single biological organism that had "two distinct incommunicable consciousnesses acting the same body" [sic] we should say that there are two distinct persons (2,27,23). He makes this point with the Mad Man/Sober Man example. He says,

... if it be possible for the same Man to have distinct incommunicable consciousness at different times, it is past doubt the
same Man would at different times make different Persons; which, we see, is the Sense of Mankind in the solemnest Declaration of their Opinions, Humane Laws not punishing the Mad Man for the Sober Man's Actions, nor the Sober Man for what the Mad Man did, thereby making them two Persons . . . (2,27,20).

Dr. Jeckyl and Mr. Hyde are two persons, on this view, who take turns acting in the same body, where "body" is the name of a biological organism. On this view, humans with Multiple Personality Disorder should also be thought of as distinct persons.

He argues, further, that sameness of person does not imply sameness of man. He says,

should the Soul of a Prince, carrying with it the consciousness of the Prince's past Life, enter and inform the Body of a Cobler as soon as deserted by his own Soul, every one sees, he would be the same Person with the Prince, accountable on for the Prince's Actions. (2,27,15)

Here we have sameness of soul and sameness of consciousness becoming associated with a different man.

This much would have been uncontroversial to Locke's chief opponent, the Cartesian who holds that sameness of immaterial substance is necessary and sufficient for personal identity. The Cartesian believes that the body and soul (the Cartesian word for immaterial substance) are distinct substances that are only contingently related. The Cartesian agrees with Locke that the same living human body is neither necessary nor sufficient for personal identity.

Locke distinguishes his view from what he takes to be the Cartesian view by arguing that sameness of immaterial substance is not sufficient for personal identity. He

3. Clarke, Butler, and Reid are all Cartesians in this sense. See this work, chapter 3.
asks what we should say if the same immaterial substance that was present in the ancient hero Nestor is now present in some presently living person. He says that this alone

would no more make him the same Person with *Nestor*, than if some of the Particles of Matter, that were once a part of *Nestor*, were now a part of this Man, the same immaterial Substance without the same consciousness, no more making the same Person by being united to any Body, than the same Particle of Matter without consciousness united to any Body, makes the same Person. (2,27,14).

Locke believes that X’s having the same immaterial substance Y is not sufficient for X to be the same person as Y.

Because Locke holds that sameness of man is not necessary for personal identity and that sameness of consciousness is sufficient for personal identity, Locke accepts what I am calling the Body-swap Intuition. But which version, the qualified or the unqualified version? That is, does Locke accept that X’s having the same immaterial substance as Y is necessary for X to be the same person as Y? Locke addresses this question twice and gives what seem to be conflicting answers.

In section 10, his position is that consciousness *alone* makes personal identity, even if the distinct acts of consciousness at different times are acts performed by different substances. He says if an “intelligent being can repeat the idea of any past Action with the same consciousness it has of any present Action; so far it is the same personal self . . . whatever Substances contribute to their [the idea and the repeated idea] production” (2,27,10). He says that personal identity depends on sameness of consciousness only, whether consciousness is “annexed only to one individual substance, or can be continued in a succession of several Substances” (2,27,10). So, the answer seems to be that sameness of substance is not necessary for personal identity.
In section 13, he addresses the issue of "Whether if the same thinking Substance (supposing immaterial Substances only to think) be changed, it can be the same Person" (2,27,13). In other words, is sameness of substance a necessary condition for personal identity? For Locke, that question becomes the question of whether sameness of consciousness requires sameness of substance.

Locke's answer to this question is that he does not know if sameness of consciousness requires sameness of substance, but that it is *not absurd* to think that the same consciousness may be annexed to different substances. He, therefore, holds that it is not absurd to think that personal identity may likewise be continued on distinct substances. It is important to say a bit more about consciousness in order to understand what Locke is doing here.

Locke says that "Consciousness is the perception of what passes in a Man's own mind" (2,1,19). Each conscious episode, for Locke, is an act of perception, each numerically distinct from the rest. An experience and the subsequent memory of that experience are distinct acts of perception for Locke and are not, therefore, numerically the same act at different times. Sameness of consciousness, for Locke, is not "the same individual action" (2,27,13). That is, sameness of consciousness is not the same act of perception at different times. If sameness of consciousness were the same act at different times, says Locke, it could not be transferred to a different substance. Since memories are representations of past actions, however, the possibility that an experience and the subsequent consciousness or memory of that experience might be associated with different substances cannot be ruled out on the grounds that such a case is not sameness of consciousness but only likeness of consciousness. But sameness of consciousness, even if the experience and subsequent memory are associated with the same substance, is not the numerical identity of an act at time \( t \) and an act at time \( t+1 \).
In other words, because the perception of the original experience and the subsequent perception of the representation are always numerically distinct, it is not absurd to hold that the same consciousness may obtain between different substances on the grounds that consciousness would not really be numerically the same but only qualitatively the same.

For the same reason, Locke sees no absurdity in supposing that the same consciousness "can be continued in a succession of several substances"(2,27,10) any more than he sees an absurdity in the same life being continued in the succession of substances—i.e. material particles—that make a living thing (2,27,4,10). He does not think it is absurd to suppose that consciousness can be continued in a succession of substances, but he does not know if it is true.

Locke’s answer in this section is that we don’t know whether an immaterial substance is necessary for personal identity because we do not know whether the same immaterial substance is necessary for sameness of consciousness. Locke proposes to leave unresolved the question of “whether consciousness of past Actions can be transferr’d from one thinking substance to another” (2,27,13), and thereby withholds on the question of whether an immaterial substance is necessary for personal identity.

One answer to the question of whether X’s having the same immaterial substance as Y is necessary for X to be the same person as Y is “no” and the other answer is “we don’t know.” We can reconcile these responses in that we can explain why Locke makes both.

Locke is engaged in a number of projects in chapter 27 but the one he sets out to engage in is to determine what makes our “Ideas of Identity and Diversity” (2,27,1). Much of the discussion is designed to show that the words, “person,” “man,” and “substance” are “three Names standing for three different Ideas” (2,27,7). This is a
roughly conceptual project of saying what makes our idea of a person, a man, and a substance. He also assumes that our concepts of a man, a person, and a substance are largely correct and thus having said what makes our ideas of these things, we may also say what makes things themselves. After saying that “person,” “man,” and “substance” are three names standing for three different ideas, he says, “for such as is the Idea belonging to that Name, such must be the Identity” (2,27,7). He makes a similar point when he begins his discussion of personal identity in section 9. There he says that if we want to find what makes personal identity, we must find what idea the word “person” stands for (2,27,9).

There are two projects Locke is engaged in: a conceptual project regarding our idea of personal identity and a metaphysical project regarding what makes personal identity. The way to reconcile Locke’s straightforward assertion in 2,27,10 that consciousness alone makes personal identity, whether “annexed only to one individual substance, or can be continued in a succession of several Substances,” and his agnosticism in 2,27,13, that we don’t know if sameness of substance is necessary for personal identity is to take him to assert a conceptual thesis about our idea of personal identity, and to acknowledge metaphysical facts that are beyond the reach of our ideas.

It is clear that Locke believes there are metaphysical facts that we have only an obscure idea of. In 2,23,2-3, Locke distinguishes between the "Notion of pure Substance in general" (2,23,2) and our idea of particular sorts of substances (2,23,2-3). The first is the obscure idea we have of a substratum, something that supports, or stands under the sensible qualities of an object. It is the unknown substratum in which the qualities of a thing inhere, a "something" we know not what (2,23,3). We have an idea of material substance, the thing in which its material properties inhere, such as the yellowness of gold or the hardness of loadstone. And we have the idea of an immaterial
substance, which Locke sometimes calls spirit and sometimes soul. Thinking, reasoning, and fearing: do not subsist of themselves and thus "we are apt to think these the Actions of some other Substance, which we call Spirit" (2,23,5). Our ideas of both material and immaterial substratum are obscure, a mere supposition of something, we know not what.

Locke contrasts the idea of Substratum with "our ideas of particular sorts of Substances" (2,23,3-4). We have ideas of a man, a horse, gold, water, etc., which are all substances, in this second sense of substance. Our ideas of each of these is made up of simple ideas co-existing together. The idea of the sun, for example, is just "an aggregate of those several simple Ideas, Bright, Hot, Roundish, having a constant regular motion, at a certain distance from us, and, perhaps, some other" (2,23,6). These ideas, once again, are simple ideas or qualities that we suppose to inhere is some substratum.

Our ideas of substrata—material or immaterial—are obscure, according to Locke. He says little more about substrata than that we suppose they exist. Furthermore, he thinks it is likely that sameness of consciousness requires the same immaterial substance. He says, "I agree the more probable Opinion is, that this consciousness is annexed to, and the Affection of one individual immaterial substance" (2,27,25).

Now, if our idea of an immaterial substance is an obscure idea, nothing more than a supposition of something, I know not what, our idea of the same immaterial substance is doubly obscure. It would be a supposition of a something I know not what at one time, a supposition of a something I know not what at a different time and a supposition that the one is the same as the other. This is obscurity upon obscurity.
Locke can claim that our idea or concept of personal identity consists in consciousness alone and still hold that there are facts about the metaphysically necessary conditions for personal identity to which we have no obvious access. That is what I think Locke does hold. As we have seen, Locke believes we have only an obscure idea of substrata, and so he is committed to holding that facts about substrata are beyond the reach of our ideas. Still, he thinks we make correct judgments of personal identity in life and in response to puzzle cases, and that we do so on the basis of our employing our idea of personal identity. So, we don't make our correct judgments of personal identity on the basis of our ideas of substrata.

This conceptual version of CAT is consistent with agnosticism about the relation between substance and consciousness. While it might be true that our concept of personal identity consists in consciousness alone, by Locke's own admission, our concept may not include all the facts.

2.3 "Person" as a Forensic Term

Locke says that "person" is "a Forensick Term appropriating Actions and their Merit; and so belongs only to intelligent Agents capable of a Law . . ." (2,2,26). The standard gloss on this passage is that by saying "person" is a forensic term, Locke is saying X is the same person as Y if and only if Y is responsible—morally and legally—for doing what X did and is entitled to whatever rewards or punishments justly due X. 4 I think the standard gloss is correct, but it is worth saying why.

There is good reason to think that Locke was using "forensick" in its ordinary sense, "pertaining to, connected with, or used in courts of law" (OED). The OED cites


22
the *Essay*, 2,27,26 as an early, though not the first, example of this use. Moreover, the passage itself points to actions and their merit before the law. Finally, some of Locke’s examples refer to civil laws (Cf. 2,27,20; 2,27,22). It is reasonable to conclude that for Locke, “person” is a legal term. To call something a person, for Locke, is to claim that it is responsible before the law.

Locke also holds that to say that X is the same person as Y is to make a claim about the justice of rewarding or punishing Y for what X did. We can see Locke making two kinds of claims here.

First, if X and Y are not the same person it is unjust to punish Y for what X has done. Locke says that if Socrates waking and Socrates sleeping do not partake of the same consciousness, then Socrates waking and Socrates sleeping are not the same person. And if they are not the same person, to punish the one for what the other had done, “would be no more Right, than to punish one Twin for what his Brother-Twin did . . .” (2, 27,19). Locke is saying that if X and Y are not the same person, then Y is not justly punished for what X did.

Second, Locke also seems to hold that if X and Y are the same person, then Y is justly rewarded or punished for what X has done. He says that in personal identity is “founded all the Right and Justice of Reward and Punishment . . .” (2,27,18). I take this to mean that claims of personal identity imply claims about the right of Y to some reward on the basis of what X did or claims about the justice of punishing Y for what X did.

Locke’s claim that “person” is a forensic term, then, is the claim that X is the same person as Y if and only if Y is due all rewards and punishments due X. The
operative notions of reward and punishment should be understood to mean reward and
punishment with respect to some law. More concisely, we can say that for Locke,
personal identity is necessary and sufficient for responsibility before the law.

The reason that personal identity is necessary and sufficient for responsibility
before the law is that consciousness is necessary and sufficient for personal identity and
for responsibility before the law. Locke believes it is consciousness that draws reward
and punishment \(2,27,13\). In each of the examples above that attempt to establish a
connection between personal identity and responsibility before the law, the connection
is made by means of sameness of consciousness.

For Locke, the law includes not only civil law, but divine law as well. There are
times in which Locke clearly has in mind civil laws. For example, in the madman/sober
man example, he is speaking of the justice of punishment with respect to “Humane
Laws” \(2,27,20\,\text{see also, } 2,27,22\). There are also times when Locke clearly has in
mind responsibility before the divine law. He says that even when confronted with the
divine law on judgment day, it is “reasonable to think, no one shall be made to answer
for what he knows nothing of . . . “ \(2,27,22\).

We can take claim that “person” is a forensic term to mean that persons are
responsible with respect to law, whether civil or divine and that personal identity is
necessary and sufficient for responsibility before the law. For Locke, moral goodness
and badness consists only in the conformity to law, whether divine, civil, or what he
calls “the Law of Opinion or Reputation” \(2,28,7\). Since “person” is a legal term and
morality is just conformity to the law, “person” is a moral term as well, as the standard
gloss says.
In 2.2 I argued that Locke's view is that our concept of personal identity consists in consciousness alone, even if that concept, given human limitation, does not include all the facts. When we consider the uncertainty that Locke is willing to tolerate with respect to the concept of a person, it is no surprise to find Locke uncertain about matters regarding responsibility before the law.

Recall that Locke believes it is not absurd to think that the same consciousness might be transferred to different substances. He thinks, of course, that we could not know that the same consciousness has actually been transferred. But given the assumption that consciousness may be transferred to a different substance, it would also seem possible that a person P might have a representation of having experienced an action that commonsense tells us that P did not do. Locke considers whether "one intellectual substance" might have "represented to it, as done by it self, what it never did, and was perhaps done by some other Agent" (2,27,13). On the supposition that sameness of substance is not necessary for sameness of consciousness, we can suppose that one intellectual substance has represented to it an action as having been done by itself that it did not do, and was actually done by another agent.

To illustrate: suppose that I wake up to find that I have representations of the experiences the Boston Strangler experienced on his murderous rampage. Commonsense would regard these representations as a mistake, as misrepresentations. But doesn't Locke's view imply I am the same person as the Boston Strangler?

Some things he says suggest that Locke is prepared to say that the Boston Strangler and I are the same person. For example, in the Nestor case I cited earlier,
Locke says that having the same soul as Nestor is not sufficient for sameness of person, but, he says, if a person finds himself “conscious of any of the Actions of Nestor, he then finds himself the same person with Nestor” (2,27,14). If the consciousness of any of Nestor’s actions make me the same person as Nestor, it follows that if I find myself conscious of any of the Boston Strangler’s murderous experiences, I am the same person as the Boston Strangler. Moreover, since, on Locke’s view, claims about personal identity are materially equivalent to claims about responsibility before the law, we should say that I am guilty before the law of the stranglings. Assuming that sameness of substance is not necessary for sameness of consciousness, it follows from premises that Locke seems to accept that I am guilty before the law of the Boston Strangler’s stranglings.

This creates a problem in understanding Locke’s view, and in so far as one is attracted by the idea that personal identity consists in consciousness alone, it raises a number of philosophical problems as well. The supposition on which the present case proceeds is that we have two “intellectual substances” in the beginning, one of whom seems to be remembering something that in fact he never experienced. This description of the case suggests that the one who seems to be remembering is in fact mistaken if he attributes the action to himself for, ex hypothesi, it was done by some other agent, which I assume means a different intellectual substance. On the other hand, it follows from premises that Locke seems to accept that the one who seems to be remembering is the same person as the one who had the experience. This conclusion suggests that the one who seems to be remembering is not mistaken in attributing the action to himself.

Locke seems to want to accept both of these conclusions. He regards the case in which one intellectual substance has represented to itself that which it did not do and was perhaps done by another as a case in which a “fatal error” is being made. He says,
as though to assure us that our lack of knowledge about substances will not result in our being punished unjustly, that God “will not by a fatal Error of theirs [His sensible creatures] transfer from one to another, that consciousness, which draws Reward or Punishment” (2,27,13). According to this passage, God keeps his sensible creatures from making an error that Locke describes as fatal. Of course, if my representations of the Boston Strangler’s doings are misrepresentations, then I am making error in attributing these actions to myself. If I am punished for these actions because I attribute them to myself, my error is fatal if my life is required for the lives I believe I have taken.

We can understand why Locke might think of misrepresentations as having fatal results, but it is hard to say what Locke could mean by calling them errors. If sameness of consciousness is necessary and sufficient for personal identity, and personal identity is necessary and sufficient for responsibility before the law, then I am the same person as the Boston Strangler, responsible for his actions. But Locke says that because God is good, he “will not by a fatal Error of theirs [His sensible creatures] transfer from one to another, that consciousness, which draws Reward or Punishment” (2,27,13).

But why not? No injustice would be done: if consciousness alone makes personal identity, and personal identity is necessary and sufficient for guilt before the law, I would not be suffering unjustly. Still, Locke’s initial description of the case suggests that the one with the apparent memory is making a mistake, perhaps an unavoidable mistake in attributing that action to oneself. Furthermore, Locke’s deferring to God’s goodness to assure us that even if such errors are possible, God will keep his creatures from either making the errors or for suffering unjustly for them, only highlights that Locke thinks of this as a case in which some error is being made by someone.
We are left with the following choices: We can take Locke to hold that consciousness alone, even consciousness of an act one did not experience, makes personal identity, or we can take Locke to hold that veridical consciousness alone makes personal identity. If the former, it is hard to make sense of the “fatal error” passage. If the latter, the plausibility of his view will depend crucially whether he can draw the distinction between veridical and non-veridical consciousness independently of substance, a distinction Locke does not draw for us.

We must remember, however, that these choices arise on the supposition that sameness of substance is not necessary for sameness of consciousness. Moreover, we must remember that, while Locke does not believe this is an absurd supposition, he does not know if it is true. He does not assert it; he merely considers it and withholds on whether it is true.

While this leaves much uncertainty about what Locke holds about personal identity, we can say the following with confidence. Locke promotes his version of CAT as a thesis about our idea or concept of a person, but he withholds on whether sameness of substance is metaphysically necessary for sameness of consciousness. And so, he does not hold the Unqualified Body-swap Intuition. He does accept the Body-swap Intuition, however, because he believes both that the body is not a necessary condition for personal identity and that sameness of consciousness is sufficient for personal identity.

There is one further thing that we can say with confidence. The claim that consciousness alone makes personal identity along with the assumption that sameness of substance is not necessary for sameness of consciousness raises the following dilemma. If one denies that sameness of substance is necessary for sameness of consciousness, one must either accept that consciousness is sufficient for personal
identity whether commonsense would regard that consciousness as veridical or not, or one must be prepared to distinguish between veridical consciousness, which is sufficient for personal identity, from non-veridical consciousness, which is not. That means, in order to defend the Unqualified Body-swap Intuition, one must either accept the counter-intuitive implications of allowing what seems to be non-veridical consciousness to make personal identity, or one must make a distinction between veridical and non-veridical consciousness.

In the next chapter, I consider Henry Dodwell’s view. Dodwell defends CAT and the Unqualified Body-swap Intuition, but accepts the counter-intuitive implications of allowing what commonsense would regard as a false consciousness to count as the right kind of consciousness.

Before closing this chapter, I want to say a word about the second option, that of drawing a distinction between veridical and non-veridical consciousness while at the same time holding the Unqualified Body-swap Intuition. A theorist who wants to endorse the Unqualified Body-swap Intuition has little to work with to draw the distinction. Since the Unqualified Body-swap Intuition says that the body is not a necessary condition for the existence of a person at different times, no appeal can be made to sameness of body, since that would be to deny the Body-swap Intuition in either the qualified or the unqualified form. In so far as one wants to defend the Unqualified version, no appeal can be made to sameness of some other substance such as a brain or a soul.
In this chapter, I have three aims. First, I lay out Dodwell's view in order to demonstrate his commitment to CAT and to the Unqualified Body-swap Intuition. Second, I examine the most influential criticisms that have been raised against Dodwell's view specifically and against CAT generally. Third I draw some conclusions about what is required to defend any version of CAT that implies the Unqualified Body-swap Intuition.

3.1 Dodwell

Dodwell's statement of his own views about personal identity parallels Locke's discussion in a number of ways. Dodwell, like Locke, considers the identity of a particular material substance such as an atom or a mass and a particular mode such as motion. He considers the identity of a living thing, such as an oak, an animal, or a man, and he considers the identity of a person or a self.

The identity of a particular atom or mass for Dodwell, "lies in consisting of exactly the same numerical particles, to which no Addition or Subtraction has been made."1

1. All references to the Clarke-Dodwell correspondence refer to volume 3 of *The Works of Samuel Clarke*, 4 vols. (New York: Garland, Reprinted, 1928). The correspondence consists of four letters from Clarke to Dodwell, which I call Clarke's
This is in accord with Locke's claim about the identity of atoms and masses through time.²

He follows Locke on two points with respect to the identity of a particular Mode, such as a mode of motion or thought. First, some modes exist in succession and are not identical but diverse through time; second, motion and thought—particular acts of consciousness—are among those things that exist in succession. A mode of motion, such as sliding, rolling, or tumbling, according to Dodwell, is not capable of a continued existence it perishes the moment it begins (Dodwell, Third Defense 875). Likewise, "a particular Act of Consciousness is incapable of the Continuation of its Existence; wherefore its Identity can only consist in being that very numerical Act of Consciousness that it is." (Dodwell, Third Defense 875). Both of these follow the Essay. 2,27,2, where Locke says that there are some things "whose existence is in succession, such as are the Actions of finite Beings, v.g. Motion and Thought, both which consist in a continued train of Succession, concerning their Diversity there can be no question." (Locke 2,27,2). Dodwell, like Locke, believes that each conscious episode is a particular act of consciousness. As each moment is numerically distinct from one to the next, each act of consciousness is numerically distinct from all the others. As no moment can be repeated, no act of consciousness can be repeated.

He follows Locke in his treatment of living things as well. He says,

An Oak that contains several Loads of Timber, is called the same Oak with an Oak that was an Inch long an Hundred Years ago, by partaking thence-forward of continued vegetable Life, in a like

². See this work, 2.1, pp. 11 ff.
Continued Organization of Parts: and an Animal or Man is called the same Animal or Man at Twenty Years old, that it was at a Quarter old, by partaking of a continued Animal Life, under a like continued Organization of Parts. (Dodwell, Third Defense 875).

Like Locke, Dodwell does not say what life is, nor when it is the same through time.

Dodwell then turns to the question of what we “signify by the Word Self, and sometimes call Personal Identity.” He argues that, "Personal Identity consists solely in Consciousness." (Dodwell, Third Defense 875; see also 876-877).

Dodwell understands consciousness in much the same way Locke does. He considers consciousness to be an action done by a thinking being. Each particular act of consciousness is numerically distinct from the others. He believes no one has "the same numerical consciousness" today that one had yesterday because the act of today is "a distinct numerical Act" from the one of yesterday. A particular act of consciousness in which one has represented to one an experience as having been done by oneself, Dodwell calls a memory (Dodwell, Third Defense 875-877).

He gives three examples to establish the claim that personal identity consists solely in consciousness. The first two are proposed to show that sameness of consciousness is necessary for sameness of person, while the third seems to be proposed to show that consciousness alone is sufficient for personal identity. In each of the three cases, Dodwell makes a claim about what actions one can attribute to oneself and draws a conclusion about what is the case with respect to personal identity. In other words, he moves from what one believes to be the case with respect to one’s identity to what is the case with respect to one’s identity.

First, he says, “If a Man charges me with a Murder done by some body last Night, of which I am not conscious; I deny that I did the Action and cannot possibly attribute it to my Self, because I am not conscious that I did it” (Dodwell, Third Defense
875) Dodwell’s point here is that one cannot attribute an action to oneself that one cannot remember.

The second case is similar to Locke’s Mad man/Sober man case. Dodwell uses it to defend the claim that sameness of consciousness is necessary for personal identity. He says,

Suppose me to be seized with a short Frenzy of an Hour, and during that time to kill a Man, and then to return to my Self without the least Consciousness of what I have done; I can no more attribute that Action to my Self, than I could the former, which I supposed done by another. (Dodwell, Third Defense 875).

Once again Dodwell makes the point that one cannot attribute an action to oneself that one cannot remember.

Having twice made a point about what one can attribute to oneself, he then draws a conclusion about personal identity. He says, "The mad Man and the sober Man are really two as distinct Persons as any two other Men in the World. . ." (Dodwell, Third Defense 875). In this case, the mad man and the sober man do not have the same consciousness and thus cannot attribute the actions of the other to himself, and so Dodwell concludes they are not the same person. That means, for Dodwell, sameness of consciousness is necessary for personal identity.

He then argues that consciousness alone is sufficient for personal identity. Here he makes stronger claims than Locke is willing to make. He says,

should there be so strong a Representation to my Understanding of a Murder done by me (which was really never done at all) so that I could not distinguish it in my Mind

3. Locke's mad man/sober man argument is in 2,27,20. I discuss it in 2.3, pp. 15-16.
from something really done by me I can no more help attributing this to my self, than I can any other Action which I really did. (Dodwell, Third Defense 875).

Dodwell is asking us to suppose that one is having a vivid representation—an apparent memory—of having committed a murder that the one with the apparent memory can’t distinguish from genuine memories. Dodwell says that, in this case, one cannot help attributing this action to oneself.

After saying that this apparent memory is sufficient for one attributing this action to oneself, he draws a conclusion about what personal identity consists in. He concludes that, “it is evident, that Self or Personal Identity consists solely in Consciousness . . .” (Dodwell, Third Defense 875).

Dodwell has cited three cases from which he draws conclusions about what one can attribute to oneself. He cites them in order to find “what it is that constitutes Self or Personal Identity” and he draws a conclusion from these cases about what self or personal identity consists in. The premises Dodwell thinks allow this conclusion are given immediately after his conclusion. He says that “it is evident, that Self or Personal Identity consists solely in Consciousness; since when I distinguish my Self from others, and when I attribute any past Actions to my Self, it is only by extending my Consciousness to them” (Dodwell, Third Defense 875). One distinguishes oneself from others by thinking one’s own thoughts and not another’s, and one attributes actions to oneself only by remembering having done them, according to Dodwell.

Dodwell believes that personal identity consists in consciousness alone but the only evidence he cites for this constitutive claim concerns what one can attribute to oneself. Dodwell is prepared to accept that if I have represented to me as having been done by me the Boston Strangler’s experiences, then I and the Boston Strangler are the same person. From what has been said, it is clear that Dodwell takes “a Representation
to my Understanding of a Murder done by me (which was really never done at all). . .”
to be the right kind of consciousness to constitute personal identity. Commonsense, on
the other hand, would regard such representations as misrepresentations. Recall that he
uses this example to show that personal identity consists in consciousness alone. I
conclude that Dodwell accepts the claim that personal identity consists in consciousness
alone, whether commonsense would regard that consciousness is veridical or not.

This last point distinguishes Dodwell’s view from Locke’s. Locke waivers in
his discussion of a case in which one has a representation of an action that was actually
done by another. He seems to say that such a representation should be regarded as an
error.

Dodwell waivers not at all. He asserts that such a representation makes a person
the same person as the one who did the action. Dodwell asserts something even
stronger. He holds that having a representation of an action as having been done by
oneself that was actually done by no one makes personal identity as well. For example,
suppose that I have the apparent memory of having assassinated Nixon. Unlike the
Boston Strangler’s stranglings, the assassination of Nixon is an action that was done by
no one. Dodwell, nevertheless, says that were there “a Representation to my
Understanding of a Murder done by me (which was really never done at all). . .” I
could not help attributing this action to myself. Attributing an action to oneself by
having the consciousness of having experienced it is what make personal identity, for
Dodwell.

I said earlier that it is hard to tell exactly what Locke would say about the
Boston Strangler case. But Locke never asserts, nor even suggests, that an apparent
memory of having done an action that was done by no one makes personal identity.
There is a second point that marks a distinction between Dodwell and Locke. Dodwell says, "Human Thinking or Consciousness consists of a Number of particular Acts of Thinking or Consciousness" (Dodwell, *Third Defense* 876). Each particular act of consciousness, for Dodwell, is numerically distinct from the others regardless of whether these acts "reside in a fleeting or indivisible Substance" (Dodwell, *Third Defense* 876). Since consciousness alone makes personal identity, and sameness of consciousness consists in a series of distinct conscious acts, sameness of substance is irrelevant to questions of personal identity. I take this to mean that sameness of consciousness does not require sameness of substance, for Dodwell.

Locke is more cautious about the role that substance might play in personal identity. He does not know if sameness of substance is necessary for sameness of consciousness. He therefore does not know if personal identity is preserved in the absence of the same substance. Dodwell asserts that personal identity is preserved if sameness of consciousness is preserved, regardless of whether the substances involved are the same.

One further feature of Dodwell's view of personal identity deserves mention here, though it will be developed later in this chapter. Dodwell holds that personal identity may take a one-many form. That is, he allows that one thinking thing at time \( t \) may stand in the correct relation to more than one thinking thing at some later time \( t+n \). In other words, there might be two or more thinking things with the same consciousness as one earlier thinking thing. Dodwell holds that, in this case, the two are the same person as the one. He says, "let there be ever so many thinking Beings that have a present Representation of a past Action, they can all constitute but one and the same person" (Dodwell, *Third Defense* 877). He says that if God should cause to exist "Twenty representations of the same past sinful actions," we should say "each of the
twenty distinct Beings is the same Person," and we should deny they are "distinct persons from one another" (Dodwell, *Third Defense* 878).

It is obvious that Dodwell accepts the Body-swap Intuition. He believes that the body is not a necessary condition for the identity of a person through time and he believes that sameness of mental life is sufficient for personal identity. Furthermore, he does not believe that sameness of substance is a necessary condition for sameness of consciousness, so he accepts the Unqualified Body-swap Intuition.

3.2 Clarke’s criticisms of Dodwell

Clarke raises four main objections against Dodwell’s view. First, he claims that Dodwell’s view conflicts with the theological doctrine of the resurrection from the dead. According to this doctrine, persons who have died will be raised from the dead to answer for what they have done before they died. According to Clarke, on Dodwell’s view, there can be no resurrection from the dead, only the creation of a new person who is like some person who had lived and died. The second objection follows from the first. If on Dodwell’s view there can be no resurrection from the dead, only the creation of new persons, it would be unjust to reward or punish these new persons, for they are not the ones who did the actions they seem to remember doing. Third, Clarke objects to Dodwell’s view on the grounds that it allows fission. Finally, Clarke claims that Dodwell’s view is absurd because it allows what commonsense would consider non-veridical consciousness to constitute personal identity. I consider these in order.

Clarke charges that on Dodwell’s view, “the Resurrection will be incredible, and the Justice of future Rewards and Punishments impossible to be made out” (Clarke’s *Third Defense* 851). Here, Clarke is expressing a commitment to the notion that there is a conscious afterlife in which one is rewarded or punished according to
what one did in the present life. He says that on Dodwell’s view there can be no resurrection of the same person at all, only the creation of a new person qualitatively identical to some previously existing person. He argues that if consciousness is nothing but mode on a fleeting system of matter, as Dodwell allows, then the consciousness that makes personal identity perishes entirely with the dissolution of the body. A person in the resurrection could not be identical with a person who had died, only like the previous person. Clarke gives the following argument for this claim.

For, as God’s superadding Now to a new Parcel of Matter, the like Consciousness with what I at this time find in myself, would not make that new Parcel of Matter to be the same individual Person with me, but only another Person Like me; so his superadding that Consciousness at the Resurrection, to the same Particles of Dust, of which my Body was formerly composed; will not be a Restoration of the same Person, but a Creation of a new one like me. (Clarke, Third Defense 851).

Clarke’s argument here is that a resurrected self, on Dodwell’s view, is really a new self like the old, but not identical with it. For Clarke, that would mean that on Dodwell’s view, there can be no resurrection of the dead. Clarke’s argument rests on the claim that God’s forming a new body and giving it a mental life exactly like mine while I am yet alive and well would not make that person numerically the same person as me.

His reasoning seems to be this. If God created now while I am yet alive a being with a mental life exactly like mine, that new being is obviously not identical with me, but only like me. Well, if I died while God was making this new being or before God had made it, the intrinsic relation between the new being and me is the same as it would have been had I lived: the new being would not be identical with me, but only like me.
So, on Dodwell’s view there can be no real resurrection of the dead, only a new creation.

The crucial claim for Clarke to make is the claim that this newly formed being created while I am still alive and well, is not the same person as me. This is a claim that has strong intuitive appeal. How could that newly formed being be the same person as me when I am in one place and he in another? It seems inconceivable.

Still, it is a claim that Dodwell denies. As I said in my discussion in 3.1, Dodwell accepts that personal identity can be preserved even in some cases that take a one-many form. He says, "let there be ever so many thinking Beings that have a present Representation of a past Action, they can all constitute but one and the same person" (Dodwell, Third Defense 877). So, Dodwell seems to deny Clarke’s first premise: If God created now, while I am yet alive a being with a mental life exactly like mine, that new being is obviously not identical with me, but only like me. Dodwell says that that newly created being is the same person as me.

A second argument by Clarke rests on this same claim. Clarke claims that the justice of rewards and punishments in the resurrection is impossible to make out on Dodwell’s view. This claim rests on the same claim as the charge that the resurrection is impossible to make out on Dodwell’s view. It is easy to see why Clarke believes the justice of rewards and punishments in the resurrection is impossible to make out on Dodwell’s view. It seems unjust to punish this newly created being for what I have done, for he is not the one who did the act; I am. If this new being should not be punished for what I have done while I am alive, he should not be punished just because I died before he came on the scene. Hence, it would be unjust to punish anyone in the resurrection. The same thing applies to rewards.
The argument about reward and punishment rests on the same premise as the argument about the resurrection. On Dodwell’s view, it is just to reward me and punish me for what I did yesterday if and only if I am the same person as the one who did the action. Dodwell holds that this newly created being is the same person as me if and only if he has the same consciousness as me, which ex hypothesi, he has.

Neither of these criticisms from Clarke are compelling to Dodwell, since both rest on a claim Dodwell denies; nor is Clarke’s charge that Dodwell’s view is absurd because it says that many distinct beings at a time can constitute the same person.

Clarke argues that if there were twenty fleeting systems of matter such that all had a like consciousness as some previous person,

> the Addition of the like Consciousness to all those Systems, would consequently make every one of them to be not Persons like him, but the same individual Person with him, and with each other likewise, which is the greatest Absurdity in the world . . ." (Clarke, Third Defense 851)

Clarke is saying that if consciousness is not tied to any particular substance but is a mode on a succession of substances, we can imagine the same consciousness (where “same” means something other than numerically the same) to be present on more than one system at a time. If there were twenty systems at \( t \) that have the consciousness of what a single system experienced at some earlier time \( t-n \), the twenty systems would be the same person as the one. Clarke thinks that this is absurd.

Indeed it is absurd if one assumes that personal identity is the numerical identity of a single thing at different times. If personal identity is the numerical identity of a single thing through time, then it cannot take a one-many form. But again, Dodwell denies that personal identity is the numerical identity of a single thing at different times. Dodwell believes that personal identity consists in a succession of numerically distinct
acts of consciousness, and therefore is not the numerical identity of a single thing at different times.

There is one final criticism that Clarke makes. He claims that the newly-formed thinking being with the same consciousness as I have has misrepresentations, and therefore, is not the same person as me. For this reason, this newly-formed being is not, according to Clarke, responsible for what I have done. So, he is not justly rewarded or justly punished for what I have done.

Recall that Dodwell says, "Should there be so strong a Representation to my understanding, of a Murder done by me, which was really never done at all; so that I could not distinguish it in my Mind from something really done by me; I can no more help attributing this to myself . . ." Recall that he says this in defense of the claim that consciousness alone makes personal identity and cites it as a case of personal identity. Clarke cites that quote by Dodwell and says, "It is true, I could not help it indeed; but it would be (as Mr. Locke well expresses it,) a fatal Error; and not (as you would have it) a making me to be really the Person I am not" (Clarke, Fourth Defense 902-903. Italics in original). According to Clarke, because Dodwell makes no distinction between veridical and non-veridical consciousness, he must treat cases in which one has the consciousness of having done an action that was done by no one as a case in which one is the same person as the one who did the action; and this is absurd.

Of the four criticisms I have just canvassed, this final one is the most compelling, even if Dodwell does not find it so. It is plainly absurd to hold that I am the same person as the one who assassinated Nixon just because I seem to remember doing it. The right thing to say is that I am mistaken, and that perhaps I need psychological help to find out what is causing what is surely a malfunction.
The other three criticisms are not without merit, but are conclusive only if one assumes that personal identity is the numerical identity of one thing through time. Dodwell believes that personal identity consists in a succession of distinct but related things, and is, therefore, unfazed by these criticisms. As these arguments stand, they are inconclusive. The standoff is an important one because its resolution hinges upon Clarke’s undefended assumption that personal identity is the numerical identity of a single thing through time and Dodwell’s assertion to the contrary. Hume will later attack Clarke’s assumption, which would undermine the bulk of Clarke’s case against Dodwell.

Dodwell’s view is unacceptable as it stands, but this is largely due to Dodwell’s account of consciousness as memory. Some of the criticisms raised by Butler and Reid reinforce the case against consciousness-qua-memory criterion of personal identity.

The version of CAT Locke defends is similar to the version Dodwell defends and they are very easy to conflate. In the next section, I consider some criticisms that Butler and Reid raise against CAT. In some cases, Locke is their intended target; in other cases, they raise a generic attack against the very idea that personal identity is constituted by consciousness alone. In one specific case, Butler cites Dodwell’s view as an example of the view he is attacking. Since I am considering these criticisms in order to determine whether any version of CAT that does imply the Unqualified Body-swap Intuition is acceptable, it is not important to decide which particular version of CAT is being attacked; not whether Locke or Dodwell could respond, but can any version of CAT?
3.3 Butler and Reid

Butler and Reid raise the issue of amnesia against what they take to be Locke's view: If memory is necessary for personal identity, if one forgets one's past, it ceases to be one's past and one ceases to be identical with the person whose past it is. Butler urges that saying that consciousness "makes personal identity, or is necessary to our being the same person is to say, that a person has not existed a single moment, nor done one action, but what he can remember."^4

Reid uses this point to raise a reductio ad absurdum against the necessity of memory for personal identity, often called "the Brave Officer" example. He says,

Suppose a brave officer to have been flogged when a boy at school for robbing an orchard, to have taken a standard from the enemy in his first campaign, and to have been made a general in advanced life; suppose, also, which must be admitted to be possible, that, when he took the standard, he was conscious of his having been flogged at school, and that when made a general, he was conscious of his taking the standard, but had absolutely lost the consciousness of his flogging.^5

The argument goes like this. Since the brave officer remembers the experiences of the schoolboy, the brave officer is identical to the schoolboy. The general remembers the brave officer's actions, so the general is identical to the brave officer. But since the general does not remember the flogging, he is not identical to the schoolboy. Identity, however, is a transitive relation: If A is identical to B, and B is identical to C, then A is identical to C. The memory criterion, therefore, generates a contradiction.


Reid says that Locke’s view has “some strange consequences, which the author
was aware of,” namely that Locke cannot rule out the possibility that consciousness can
be transferred from one intelligent being to another. Reid says that it follows from this
that “two or twenty intelligent beings may be the same person.”

There was a great deal of controversy over whether personal identity is
numerical identity or something else. Early in chapter 27, Locke distinguishes between
our idea of identity and that of similarity (2, 27, 1). Throughout the chapter, he
distinguishes between those things that are numerically identical at different times, e.g.
an atom, and those things that are not numerically identical at different times, such as
thought. Locke grants that “the same consciousness,” which, for Locke is an act of
perception, is not “the same individual Act” at different times (2, 27, 13). Clarke,
Butler, and Reid seize on this point to argue that because successive acts of
consciousness cannot be numerically the same, the same consciousness cannot be what
constitutes personal identity, where personal identity is numerical identity (Clarke’s
term) "identity in a strict philosophical sense" (Butler’s term) or "perfect identity"
(Reid’s term). Dodwell, recall, denies this: personal identity consists in a succession of
numerically distinct conscious acts, so there is nothing relevant to a person’s existence
that is numerically identical at different times.

3.4 Butler: Circularity, Interest, and Overlap

A perennial objection to a memory criterion of personal identity is the circularity
objection. Butler claimed "one should really think it self-evident, that consciousness of
personal identity presupposes, and therefore cannot constitute, personal identity, any

6. Reid, 114.
more than knowledge, in any other case, can constitute truth, which it presupposes."^7
This can be understood in two ways.

We can understand Butler to be making a point about the concept of memory.
Parfit understands Butler's point this way:^8 It is part of our concept of memory that we
can remember only our own experiences. The continuity of memory therefore
presupposes personal identity in the same way that knowledge presupposes truth. One
can no more define personal identity in terms of memory than one can define truth in
terms of knowledge. Saying that X is the same person as Y if and only if X can
remember Y's experiencing of some event e makes the same mistake as saying P is true
if S knows that P. Both of these statements may be true, but they are implicitly
circular. This is how this objection is now understood:^9 I return to this version of the
circularity objection in chapter 4.

Penelhum believes that Butler is making rather a humdrum point that one could
not be conscious of X unless X obtains:^10 Consciousness of personal identity
presupposes that there is antecedently a person who is identical to a person at a
different time. Moreover, Butler believes "consciousness of personal identity" is a
particular case of knowing that P: P must be true in order for one to know it and thus

9. See for example Parfit's Reasons and Persons 219-223; David Wiggins, "Locke,
Butler, and the Stream of Consciousness" in Amelie Rorty (Berkeley: University of
California Press, 1976); and John Perry, "Personal Identity, Memory, and the Problem
of Circularity" in Personal Identity ed. John Perry (Berkeley: University of California
one must be identical in order for one to have consciousness of personal identity. If this interpretation is correct, as I think it is, it is clear that Butler thinks that in remembering one's past, part of what is remembered is that I am identical to the one whose experience I remember. In chapter 3, we will see that Hume denies that we have any experience of a numerically identical self.

In what follows, I give special attention to an influential argument from Butler. I have three aims: First, I argue that Butler directs his argument specifically at Dodwell and to show that he has understood Dodwell correctly on the relevant matters. Second, I provide an interpretation of that argument consistent with Butler's moral philosophy. Third, to argue that Butler's argument is decisive against the view he attacks.

In order to lay out the structure of the argument, I quote the relevant passage in full and, for future reference, make my own divisions by inserting letters into the text.

[ Dodwell's view ] when traced and examined to the bottom, amounts, I think, to this: [a] "that personality is not a permanent, but a transient thing: that it lives and dies, begins and ends, continually: that no one can any more remain one and the same person two moments together, than two successive moments can be one and the same moment: that our substance is indeed continually changing: but whether this be so or not, is, it seems, nothing to the purpose; since it is not substance but consciousness which constitutes personality; which consciousness being successive, cannot be the same any two moments, nor consequently the personality constituted by it." And from hence it must follow that it is a fallacy upon ourselves [b] to charge our present selves with anything we did, or [c] to imagine our present selves interested in any thing which befell us yesterday, or [d] that our present self will be interested in what will befall us to-morrow; since [e] our present self is not in reality, the same with the self of yesterday, but another like self or person coming in its room, and mistaken for it; to which another self will succeed tomorrow. This, I say, must follow: for [f] if the self or person of today, and that of tomorrow, are not the same, but only like persons, [g] the person of

46
today is really no more interested in what will befall the person of tomorrow, than in what will befall any other person.\textsuperscript{11}

I claim that in a, e, and f, Butler is attempting to state his opponent's view, rather than, as is commonly supposed, attempting to saddle Locke's view with these results.

There are two reasons. First, Butler directs this charge against those who had carried Locke's view "to a strange length."\textsuperscript{12} In so doing, he identifies a different opponent from Locke.\textsuperscript{13} Second, Butler cites Dodwell in a footnote as holding the view he attacks (102).\textsuperscript{14}

One important key to understanding the argument is to see that he signals his conclusions at the end of (a) with the locution "from hence it must follow," indicating that the "fallacies" he charges follow from the claim that successive selves are numerically distinct and he signals a premise in the argument in (g) with the locution, "this [the fallacy charged in b, c, and d] I say must follow for . . .," which gives the reason why Butler believes the fallacy follows from the numerical distinctness of successive selves.

I claim that the best rendering of Butler's argument takes it to be a conjunction of three conditional claims, in which a, e, and f, which are purported to state Dodwell's view, imply g, and g, in turn, implies the alleged fallacies b, c, and d. When laid out in logical order, the argument is this:

On Dodwell's view,

13. Parfit, Perry, and Whiting all take this argument as directed specifically towards Locke. See this work 4.4 for a fuller discussion of the argument these writers attribute to Butler.
14. The note is to Clarke’s third defense of his letter to Dodwell. See Butler, 105.
personality is not a permanent, but a transient thing: that it lives and dies, begins and ends, continually: that no one can any more remain one and the same person two moments together, than two successive moments can be one and the same moment: that our substance is indeed continually changing: but whether this be so or not, is, it seems, nothing to the purpose; since it is not substance but consciousness which constitutes personality; which consciousness being successive, cannot be the same any two moments, nor consequently the personality constituted by it."

[48]

our present self is not in reality, the same with the self of yesterday, but another like self or person coming in its room, and mistaken for it; to which another self will succeed tomorrow.

But,

if the self or person of today, and that of tomorrow, are not the same, but only like persons,

the person of today is really no more interested in what will befall the person of tomorrow, than in what will befall any other person

And from hence it must follow that it is a fallacy upon ourselves

to charge our present selves with anything we did, or

to imagine our present selves interested in any thing which befell us yesterday, or

that our present self will be interested in what will befall us tomorrow.

Little attention has been paid to what Butler means by "interested," a notion that figures prominently in g, c, and d. The standard line in the contemporary literature on personal identity has been to assume Butler means what we mean: What a person is interested in is what a person cares about. I claim the key to understanding what Butler means by "interested" is found in his ethical writings where the idea of an interested action or an interested desire plays a technical role. An interested action or desire for Butler is what we might call a self-interested action or desire. The arguments in the *Sermons* assume
that there are interested actions and desires and conclude that there are actions and desires that are not interested, which is nothing but the conclusion that psychological egoism is false. His moral theory is built on the existence of both interested and non-interested actions and desires, and, thus, a view that undermines the distinctions so drawn undermines other normative notions as well, such as our notion of moral responsibility. I take this up shortly. That is the charge raised in b.

Butler's argument is that Dodwell's view implies that the self of today is really no more interested in what will befall the self of tomorrow than in what will befall any other person, which, in turn, implies b, c, and d. Call g the Interest Claim and b, c, and d the Fallacy Claims. In the next section, I present a way to understand what Butler means here that is consistent with Butler's ethical writings and his philosophical methodology generally.15

Butler is an empiricist and approached the subject of morals "considered as a science" (SP 6) rather than in the apriori method of Samuel Clarke, who had a great influence on Butler's early thinking. In the Preface to the Rolls Sermons, Butler compares the two methods:

There are two ways in which the subject of morals may be treated. One begins from inquiring into the abstract nature of things: the other from a matter of fact, namely what the particular nature of man is . . . The following Discourses proceed chiefly in this latter method. (SP 7).

15. Reference to the Fifteen Sermons Preached at Rolls Chapel follow the Bernard edition. (London: G.Bell and Sons LTD, 1958). I observe the standard convention in citing Butler. SP refers to the preface to the Sermons; SP1 refers to the first sermon etc.
The first three Sermons are entitled "On Human Nature" and in them Butler attempts to describe the different parts of a person's nature.  

In describing human nature, Butler considers the different kinds of motives—Butler calls them principals of action—from which people act. He argues that the different principles of action point to the human constitution as a hierarchical system, and as a system, it makes sense to think of the parts of the human constitution as having proper functions with respect to each other. Thus, Butler's metaethic says that virtue consists in acting in conformity with one's nature and vice consists in deviating from one's nature. (SP 13).

Butler's theory, then begins with purported descriptions of actual motivations. He argues that the existence of the various parts demonstrates that the human constitution is a system that functions with a purpose rather than a hodgepodge of parts at war with each other. This in turn allows him to draw a distinction between mere strength of the parts and the legitimate power or authority of the parts. The description of the different parts of human nature is descriptive not normative. The normative aspect of his theory concerns the proper function of the different parts. The descriptive part alone is relevant to the problem in the Dissertation on Personal Identity.

Butler argues that human nature consists of "a variety of internal principles" (S1.4). He lists four. We have various appetites, passions and affections; the principle of benevolence; that of self-love; and finally, the principle of reflection he calls 'conscience.' These four principles are distinct from one another; we need but to observe human activity to see that this is so. His argument in Sermon 1 shows that Butler thinks it is obvious that benevolence exists. He says,

And if there be in mankind any disposition to friendship; if there be any such thing as compassion . . . if there be any such thing as the paternal or filial affections; if there be any affection in human nature, the object and end of which is the good of another; this is itself benevolence, or the love of another (S 1.6).

Butler thinks that benevolence is not self-love, benevolence being a reflective principle that has as its goal the good of another and self-love a reflective principle that aims at the good of oneself.

What do we learn from this? Butler holds that experience furnishes us with example after example that points us towards a distinction between self-love and benevolence. Moreover, Butler takes this distinction to be fundamental for his moral theory, indeed, for any moral theory. Butler is, after all, perhaps most celebrated by his arguments against psychological egoism. He thinks it can be demonstrated empirically, from experience rather than apriori, that there is a difference between actions motivated by self-love and those motivated by benevolence, and that at least some actions are motivated by benevolence. Moreover, he assumes, as does his contemporaries that some actions are motivated by self-love and at least some of these achieve their aims.

In a number of places, Butler speaks of "interested" actions and "interested" desires. We see this in one of his arguments against the enemy he calls the "Hobbist." The Hobbist is the psychological egoist who holds that all actions are motivated by self-
interest. Actions that we mistakenly believe to be benevolent actions—Mother Theresa’s apparent sacrifice of her own good for the good of others, for example—are really examples of egoistic desires asserting themselves.

Against the Hobbist, Butler argues for the difference between self-love and benevolence. "Self-love and interestedness was stated to consist in or be an affection to ourselves, a regard to our own private good: it is therefore distinct from benevolence, which is an affection to the good of our fellow creatures" (SI 1.11). He argues that there is a difference between self-love and particular desires. He says that egoistic accounts stem from "the confusion of calling actions interested which are done in contradiction to the most manifest known interest, merely for the gratification of a present passion" (21). While arguing that even if one insists on calling all desires selfish (because, as the familiar line goes, all of one's desires are one's desires), we must still distinguish between "cool or settled selfishness" and "passionate or sensual selfishness." Having drawn this distinction Butler says that "the most natural way of speaking plainly is, to call the first only, self-love and the actions proceeding from it interested."

What was newsworthy to Butler's contemporaries, of course, is a series of purportedly empirical arguments that demonstrate the existence of benevolence and the distinction between those desires or actions that are interested and those that are not. The news is that benevolence exists and that there are aims and actions that are not interested.

So, how are we to understand the Interest Claim and the Interest Fallacies in the light of this? Consider the Interest Claim:

The person of today is really no more interested in what will befall the person of tomorrow, than in what will befall any other person
The Butler of the *Sermons* argues that some actions are not interested and that some aims are not interested; that is, there are benevolent actions and benevolent desires. The assumption that underlies Butler's refutation of psychological egoism as well as the very dispute between Butler and the egoist is that some aims and some actions are interested, for this was never questioned, not by Butler, nor by his contemporaries.

I take it that in saying that the person of today is *no more* interested in what will happen to the person of tomorrow than in what will happen to any other person, Butler has it that the person of today by definition is *not* interested in what happens to another. Thus Dodwell's view implies a falsehood, for it implies that there are no interested actions. If there are no interested actions, then it is a fallacy to believe that there are. It is a mistake to believe our present selves are interested in what befell us yesterday, for our present selves did not exist yesterday. It is a mistake, moreover, to believe that our present selves are interested in what will befall us tomorrow, for no action motivated by self-love can result in an interested action, since no action will benefit the self whose self-love motivated the action.

The charge I am attributing to Butler is that Dodwell's view implies that we are mistaken in our belief in the existence of successful egoistic actions, those that achieve their intended egoistic aim. Butler intends this as a reductio, of course, for he and his contemporaries believe that self love exists and is a rational principle of action, that we have egoistic aims, that it is sometimes rational to act to achieve egoistic aims, and that it is rational to do so because egoistic aims are sometimes successfully carried out.

On my interpretation, Butler claims that on Dodwell's view, there are no successful egoistic actions; or alternatively, egoistic intentions are never successfully carried out. The idea is that although we have both egoistic and benevolent aims, on Dodwell's view the world is such that the egoistic aims cannot be carried out because
the one they benefit is not the one they were intended to benefit. Let's say that Butler's charge against Dodwell is that Dodwell's view implies that self-love is not a rational principle of action, for it implies that actions motivated by self-love cannot be successfully carried out.

To see why Butler thinks Dodwell's view implies this, it is necessary to consider one more criticism that Clarke raises against Dodwell. It comes from the *Fourth Defense*, which is the last installment of the Clarke-Dodwell correspondence, so we do not know how Dodwell would have responded.

Clarke says that if there is no difference between raising the same person from the dead and creating a new one like the old,

\[
\text{it unavoidably follows, that I may at this present Time as possibly be created a-new, and have another Self, existing at the same time with me, by God's adding to the *Understanding* of a new-made Body *so strong a representation* of my past Actions that That Other My-self could not distinguish them in his mind from things really done by him.}^{17}
\]

This is a fission case because it involves one person having two or more successors, each of which stands in the correct relation (by Dodwell's lights) with the one, but it is a variation because it involves overlap between what seems to be an original and what seems to be a copy. It is this overlap that really helps us to see what is fundamentally at issue between Clarke and Butler, and Dodwell.

Let's imagine this case from Clarke's perspective, as he considered it in 1708, the year he wrote the *Fourth Defense*. Suppose that he, Clarke, has lived in the usual way from his birth in 1675 through the time he delivered the Boyle Lectures in 1704 and until he began his correspondence with Dodwell in 1706. Now suppose that some

---

17. Clarke, 903.
time shortly thereafter, God took a "new-made Body," as Clarke says, and gave it Clarke's memories, while Clarke, the one born in 1675 continues in the usual way through the time of the Fourth Defense in 1708. Clarke calls the being with the new body, "That Other My-self." Were there no overlap—suppose Clarke died in 1706 and "That Other My-self" emerged in 1707—it would be tempting to say that Clarke and That Other My-self are the same person. Indeed, the description given just is Dodwell's conception of the resurrection. He says that in such a case the same person is raised from the dead.

We must spell out a bit more what this means for Dodwell. For Dodwell, a person is a succession of numerically distinct beings. So, supposing that the being that died in 1706 and That Other My-self are distinct beings in a succession of beings that together constitute the person Clarke, we can say that that Other My-self is the same person as Clarke who died in 1706. On Dodwell's view, these distinct beings stand in the right relation to one another if That Other My-self has the consciousness of (has representations of) Clarke's experiences as having been experienced by him.

In a certain sense, then, Dodwell can make sense of the claim that the being that died in 1706—still supposing that there is no overlap—and That Other My-self are the same person. A person, as he seems to hold, is a succession of distinct beings, perhaps, as a train is a succession of distinct cars. What makes distinct beings part of one succession rather than another is the relations among them, just as what makes distinct cars part of one succession of cars rather than another is the relations among them.

Although he can make sense of the claim that the being who died in 1706 and That Other My-self are the same person, spelling out the essential relation between these two beings in terms of memory is not a promising option. Clarke, Butler and Reid make a compelling case that the memory criterion gives the wrong results (Amnesia
Criticism, False Memory Criticism) and can give contradictory results (The Brave Officer Case). Some other relation is needed. Setting that aside for the moment, let’s consider the case in which there is overlap.

Imagine that That Other My-self emerges early in 1707 and Samuel Clarke—the one born in 1675, the deliverer of the Boyle Lectures in 1704, and the initiator of the correspondence with Dodwell—continues in the usual way.

Dodwell is committed to saying that the relation in virtue of which the being who delivered the Boyle Lectures in 1704 and the being who began the correspondence with Dodwell in 1706 are the same person also obtains between the one who began the correspondence with Dodwell and That Other My-self. In other words, the one who began the correspondence with Dodwell and That Other My-self are every bit as much the same person as the one who delivered the Boyle Lectures in 1704 and the one who began the correspondence with Dodwell are the same person.

That gives us two pairs of beings to consider initially:

Pair 1: The one who delivered the Boyle Lectures in 1704 and the one who began the correspondence with Dodwell in 1706.

Pair 2: The one who began the correspondence with Dodwell in 1706 and That Other My-self.

Each pair consists of two distinct beings, the first of which precedes the other in time. Dodwell regards the beings in each pair to constitute the same person and is therefore committed to seeing the relation between the first and second beings as equally interested. That is, he is committed to seeing the egoistic intentions of the first that are carried out by the second as successful in the case of Pair 2 as in the case of Pair 1; the one case is no more a case of interested actions than the other.

Now consider the overlap. Consider the following pair:
Pair 3: The one who began the correspondence with Dodwell in 1706 and the one who wrote the Fourth Defense in 1708.

Pair 2 and Pair 3 together describe a fission case. The first being in view in each pair is numerically the same, while the second in view in each pair is numerically distinct from the first and numerically distinct from one another.

Pair 1 and Pair 3 together describe what seems to be a portion of the life of a single person, Samuel Clarke. Clarke, after all, is the one who delivered the Boyle Lectures, began the correspondence with Dodwell, and continued that correspondence by writing the four Defenses.

The first being in pair 3—the one who began the correspondence with Dodwell in 1706—is accounted for already: it is included in the life of the single person, Samuel Clarke. That leaves That Other My-self.

Butler would say the one who delivered the Boyle lectures is interested in what will befall the one who began the correspondence with Dodwell, who is in turn interested in what will befall the one who wrote the Fourth Defense in 1708. Butler would also say that the one who wrote the Fourth Defense is not interested in what will befall That Other My-self. By this is meant that the egoistic intentions of the one who began the correspondence with Dodwell cannot be successfully carried out by That Other My-self.

This last claim is an important one. It relies on the assumption that the relation between the one who began the correspondence with Dodwell in 1706 and the one who wrote the Fourth Defense in 1708 has something that the relation between the one who began the correspondence with Dodwell in 1706 and That Other My-self lacks. This unspecified “something,” for Butler, is a necessary condition for successful egoistic actions. Notice that the argument does not rely on the claim that the unspecified something is identity or that identity is a necessary condition for “it.”
Dodwell, as I have said, denies that there is anything that the one pair has anything the other lacks. That means that Dodwell must treat both cases the same with respect to interestedness: either both are interested or neither are. As I said, we don’t know what Dodwell would say here but I want to say more about the two options available to him.

Since he must treat both cases the same with respect to interestedness, he can either accept that the one who began the correspondence with Dodwell in 1706 is not interested in what will befall the one who wrote the *Fourth Defense* in 1708, or he can argue that the one who began the correspondence with Dodwell in 1706 is interested in what will befall That Other My-self. On this second option, a Dodwellian may accept the claim that egoistic intentions are never successfully carried out, but rather than this counting against his view, it becomes an indictment against commonsense notions of what it is to act in one’s own best interest. Adopting this option is to adopt an error stance towards commonsense views of a person’s existence through time, commonsense views about what persons have a reason to do, or both.

In chapter 3, I consider take up both of these options. I argue that the first option is unacceptable in its own right and useless as part of a defense of the Unqualified Body-swap Intuition. I begin to consider the second option there as well, the discussion of which will continue through chapter 4. Interestingly, both options have been found in Hume.

3.5 Summary
I have described the central criticisms Clarke, Butler, and Reid raise against Dodwell’s CAT. Three of these criticisms are directed specifically at the consciousness-qua-memory criterion: The False Memory criticism charges that memory, understood as
merely having the representation of an action's having been done by oneself, is not sufficient for personal identity. The Amnesia criticism charges that memory is not necessary personal identity. The Brave Officer case shows that the supposition that memory is both necessary and sufficient leads to a contradiction. These three criticisms show that Dodwell’s consciousness-qua-memory version of CAT is unacceptable as it stands.

Some of the criticisms, however, are not nearly so decisive. Consider the problem of fission. It is certainly strange to say that twenty distinct beings at one time can be one person. But is it false? After all, twenty distinct railroad cars at one time can be one train. Are persons like that? It is certainly strange to say that they are but it is not logically absurd.

The problem of destruction and replication is also hard to assess on its own. Imagine once again that I awoke with a different body than the one I seem to remember going to bed with. Suppose that the brain and body of the one whose past I seem to remember was scanned and destroyed and the brain of the body I have was reconfigured to make me a mental duplicate of the one whose body was scanned and destroyed. Am I the same person as the one whose past I seem to remember? Dodwell would say yes, Clarke, Butler, and Reid would say no. But with what right do either Dodwell or his opponents make these claims?

As I said, Dodwell has two possible answers to Butler’s Interest Argument. He must treat cases that are the same with respect to consciousness the same with respect to interestedness: he can either accept that the one who began the correspondence with Dodwell in 1706 is not interested in what will befall the one who wrote the Fourth Defense in 1708, or he can argue that the one who began the correspondence with Dodwell in 1706 is interested in what will befall That Other My-self. For reasons I
make evident in chapter 4, only the second option is consistent with the Unqualified Body-swap Intuition.

Let me summarize these criticisms against the CAT so far in order to see more clearly what is needed to defend the second option. What is to count as success in arguing that the one who began the correspondence with Dodwell is as interested in what will befall That Other My-self as in what will befall the one who wrote the *Fourth Defense*?

Since these two beings are numerically distinct, a Dodwellian must grant that personal identity may obtain between distinct beings. He must accept, therefore, the criticism that Clarke, Butler, and Reid make that on Dodwell’s view personal identity is not the strict philosophical identity of a single enduring being through time.

He must also specify the relation that holds between distinct beings in virtue of which distinct beings are the same person. The criticisms directed specifically towards the memory criterion of the Dodwell’s CAT show that memory alone cannot be the correct relation. As we shall see, there are better candidates to play this role than memory alone. The crucial question here is whether a view that holds that consciousness alone (where the concept of consciousness is broader than the concept of memory) is the essential relation that must hold between beings is a plausible candidate.

Moreover, the preferred relation must be spelled out non-circularly. If one claims that personal identity consists in consciousness alone, consciousness must be spelled out in such a way that it does not presuppose personal identity between the related entities.

Finally, can a theory that meets the above demands be used as an argument for the second option that the one who began the correspondence with Dodwell is as
interested in what will befall That Other My-self as in what will befall the one who wrote the *Fourth Defense*? That is, is there a defensible version of CAT that implies the Unqualified Body-swap Intuition?
In this chapter, I consider the relation between CAT and interestedness. Does CAT imply that egoistic actions are never successful or does it imply that one's egoistic intentions may be successfully carried out by a being with which one does not share a body, a soul, or any other substance, and is related to a previous being only in that it has the same consciousness, however we wish to spell out "consciousness"?

I take up this issue by considering Hume's important contribution to the discussion of personal identity. The traditional interpretation of Hume's discussion of personal identity says that Hume believes that persons are nothing but bundles of experiences, impressions and ideas, related by causality, resemblance and contiguity, and that therefore, it is always false to say that the same person exists at different times, since no two bundles are identical at different times. I say more about this below. For now I want to note that the on the Traditional Interpretation, Hume takes an error stance towards commonsense notions of personal identity. I will eventually argue that taking an error stance towards commonsense notions of personal identity commits one to an error stance towards commonsense notions about what persons have a reason to do. I argue that not only would an error stance towards practical rationality not be acceptable to Hume, it implies that the Unqualified Body-swap Intuition is false.
The traditional interpretation of Hume's view says that attributions of personal identity at different times are always mistaken because the self is not the same through successive moments. It is a short step from that claim to the claim that the self we intend to benefit is never the self who receives the benefit; which is just the claim that there are no successful egoist actions.

4.1 The Traditional Interpretation

According to Hume as Penelhum interprets him, the problem of personal identity is the problem of trying to justify a practice that seems strange and even paradoxical. That is the practice of talking about people as single beings in spite of the fact that they are constantly changing, and over time have changed completely.

On this understanding, the problem of personal identity is an instance of the more general problem of the identity of changing things. The practice Penelhum has in view is the practice of attributing identity to changing things at different times, a practice in which all competent users of the language presumably engage, whether philosophers or ordinary folk. We are inclined, for example, to say of trees, brutes, and loved ones that they are the same through time, in spite of their gaining and losing parts.

The philosophical problem concerns the justification of that practice. How could such identity attributions be correct? How can a thing remain the same if its parts have changed?

According to the traditional interpretation, Hume gives an explanation of the mistaken practice but he claims there is no justification for this practice. On this

1. Penelhum, 214.
interpretation, philosophers and common folk alike, systematically err by attributing personal identity at different times. The traditional interpretation says that Hume charges that identity attributions are mistaken in two ways.

First, we mistake a succession of changing objects for an uninterrupted and invariable object because the "action of the imagination" by which we consider the two are "almost the same to the feeling" (T 254). It is the resemblance of the ideas of a succession of related objects with the idea of an uninterrupted and invariable object that is "the cause of the confusion and mistake" (T 254). The succession of related objects, for Hume, is the succession of perceptions which "successively make their appearance; pass, re-pass, glide away, and mingle in an infinite variety of postures and situations" (T 253). He says of this succession, "there is properly no simplicity in it at one time, nor identity in different" (T 253). So, Hume believes that we mistake a succession of related objects for an uninterrupted and invariable object and he explains the mistake by saying the two are easily confused.

He argues further that we try to justify our mistake by inventing an uninterrupted and invariable entity to unite the successive objects. He says, "our propensity to this mistake" is so great that at last "we yield to it, and boldly assert that these different related objects are in effect the same [when they are not]" (T 254). That is the first mistake.

In order to justify this first mistake, we make another; namely, "we feign the continu'd existence of the perceptions of our senses, to remove the interruption; and run into the notion of a soul, and self, and substance to disguise the variation" (T 254). The second mistake concerns the philosophical story we tell to justify the mistake we make in calling non-identical entities the same. We justify (or attempt to justify) the first mistake by inventing or positing the existence of a soul, a self, or a substance.
He states the position he attacks:

There are some philosophers, who imagine we are every moment intimately conscious of what we call our SELF; that we feel its existence and its continuance in existence; and are certain, beyond the evidence of a demonstration, both of its perfect identity and simplicity (T 251).

Hume does not say which philosophers he has in mind, but the position he attacks here is the roughly Cartesian position that personal identity consists in the persistence of a simple substance that is numerically identical through time. Descartes may have held this, and Samuel Clarke and Joseph Butler, noteworthy contemporaries of Hume held, surely did.\(^2\)

Against these philosophers, he argues that he has no experience of such an entity. Rather, he claims, a careful introspective examination reveals nothing but a stream of closely related perceptions. He says,

> For my part, when I enter most intimately into what I call myself, I always stumble on some particular perception or other. . . . I never can catch myself at any time without a perception, and can never observe anything but the perception (T 252).

Hume concludes, even more famously, that the whole of mankind are "nothing but a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in a perpetual flux and movement" (T 252). Moreover, since Hume denies that bundles are identical at different times, he adopts an Error Theory of personal identity attributions. He says that the mind is not simple, as was held by the Cartesians, but that inspection reveals "There is properly no *simplicity* in it

---

at one time, nor *identity* in different; whatever natural propensions we may have to imagine that simplicity and identity" (T 253). Hume holds there is a difference between what is the case with respect to what persons are and what we are inclined to believe is the case.  

4.2 A problem for the traditional interpretation

On the traditional interpretation, Hume's statements commit him to an error theory of attributions of personal identity. The traditional interpretation takes Hume to hold that attributions of identity to changing things, including persons, are false. The error lies in believing that such attributions are true and the error is compounded by pretending to experience an unchanging thing such as a self, a soul, or a substance. Penelhum summarizes what he takes to be Hume's position this way:

> What he is actually claiming is that we are constantly making a mistake in referring to a person from day to day as the same person (in using the same proper name, for example), or in referring in this way to *anything* that has changed in the slightest. For, strictly speaking, a changed person would be literally another person. . . . If we make a mistake in doing this, [using the same proper name, for example] it is a mistake we all make all the time, and a mistake of which the correction would require a complete overhaul of the concepts and syntax of our language.  

3. Hume makes a subtle move from talk of the identity of persons to talk of the identity of minds. He says, "when I enter most intimately into what I call *myself*" and "the rest of mankind," indicating that he is speaking of persons or selves (T 252). The bundle he has in mind, however, is comprised of perceptions alone. In other words, while Hume sets out to talk about the identity of a person or self, he ends up talking about the identity of a mind. I will not pursue this issue here since its solution is quite independent of the conclusions I wish to draw from Hume's discussion.

4. Penelhum, 224.
I now turn to a problem that arises on this interpretation. It is easily shown how the Error Theory conflicts, irreconcilably so, with Hume's refutation of psychological egoism, his discussion of sympathy, and with commitments he has to the rationality of the indirect passions pride and shame, love and hatred.

Hume believes psychological egoism, or what he calls the "selfish system of morals" of Hobbes, is a mistake, even if it is an understandable mistake. He thinks that the "deduction of morals from self-love, or a regard to private interest, is an obvious thought..." But, he says, "the voice of nature and experience seems plainly to oppose the selfish theory." (EM 40). He argues in the second appendix that some actions demonstrate "disinterested benevolence, distinct from self-love" (EM 92). He is therefore committed to the existence of both self-love and benevolence.

He equates benevolence with sympathy, a prominent notion in both the Enquiry and the Treatise, which is fundamental to his account of morality. He explains sympathy by using the familiar analogy to "strings equally wound up" that vibrate sympathetically: when one is plucked, the others vibrate as well (T 576).

The argument against the selfish system of morals demonstrates the existence of benevolence. The discussion of sympathy attempts to explain the mechanism that gives rise to it. He says, "no passion of another discovers itself immediately to the mind" but that we "infer the passion" (T 576). We have no direct experience of the passions of others. Our own passions, on the other hand, arise immediately in the mind or soul. (T 275).

According to Hume, indirect passions such as pride and shame\(^6\) take the self as their object, while indirect passions such as love and hatred take other selves as their object. He says, "Tis evident, that pride and humility, tho' directly contrary, have yet the same OBJECT. This object is self . . ." (T 277). Consider one of Hume's examples. Suppose a man is proud of a beautiful house that belongs to him or which he himself has built. Hume says, "the object of the passion is himself, and the cause is the beautiful house" (T 279). Hume believes that pride always takes the self as its object, while the object of love and hatred is always "some other person, of whose thoughts, actions, and sensations we are not conscious" (T 329).

For Hume, then, self-love, pride, shame and other self-regarding passions exist as do benevolence, love, hatred, and other other-regarding passions exist. So, Hume believes in their existence. Does he believe that they are contrary to reason?

It is crucial for Hume's moral enterprise in Book III that self-love, benevolence and the indirect passions don't turn out to be contrary to reason, since, according to Hume, moral distinctions are derived from sentiment, which includes self-regarding passions, such as self-love, pride and shame, and other-regarding passions such as love and hatred.\(^7\) If the sentiments from which morality has been derived are contrary to reason, the rationality of morality is called into question as well.

Now, the allowable grounds for a claim that a passion is contrary to reason are, for Hume, quite minimal. That is because passions, unlike beliefs—what Hume calls

\(^6\) The passions, for Hume, include both emotions (T 276) and desires (T 417).

\(^7\) Hume's *An Enquiry Concerning the Principles of Morals* is an extended argument for the claim that moral distinctions are derived from sentiment, what is felt, rather than reason. See especially Appendix I, Concerning Moral Sentiment. (Indianapolis: Hackett Publishing, 1983).
judgments— are not capable of truth or falsity. A passion, says Hume, is an "original existence" and does not represent objects (T 415). The role reason plays in the moral enterprise, according to Hume, is to discover the means to bring about a desired end, where the ends are given by sentiment not from reason, from what we desire or feel rather than from what we believe (T 455 ff.). He argues that "Reason is and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (T 415). Reason serves to discover the means to satisfy our passions. Our ultimate aims are not discovered by reason; nor are our ultimate aims subject to criticism by reason. He says,

'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter. (T 416).

In order for a passion to be contrary to reason, it must, according to Hume, be accompanied by some judgment which is itself unreasonable. Hume gives two grounds on which "any affection can be call'd unreasonable" (T 416). The first ground is if "a passion, such as hope or fear, grief or joy, despair or security, is founded on the supposition of objects, which do not really exist" (T 416). Here a passion is considered contrary to reason on the basis of a judgment of existence that is itself contrary to reason. For example, if one believes that there are ghosts in the sugar bowl and therefore one fears having sugar in one's coffee, we may say one's fear of sugar is contrary to reason. Strictly speaking, however, it is the judgment that there are ghosts
in the sugar bowl that is contrary to reason. If there were ghosts in the sugar bowl or if believing that there are is a reasonable belief, it would not be contrary to reason to fear having sugar in one’s coffee.

The second ground is if "in exerting any passion in action, we chuse means insufficient for the design'd end, and deceive ourselves in our judgments of causes and effects" (T 416). Here a passion is contrary to reason on the basis of an unreasonable judgment of cause and effect with respect to an act undertaken to satisfy a passion. If a passion cannot be satisfied by the means chosen, Hume is willing to say that the passion may be said to be contrary to reason, although again, he emphasizes that strictly speaking, it is not the passion itself that is contrary to reason, but the judgment of cause and effect that accompanies it.

Let us say that, for Hume, a passion is innocent of the charge of irrationality until proven guilty, and that they may be proven guilty on only two grounds. A passion is unreasonable if it is founded upon an unreasonable existence claim or an unreasonable claim about the means to satisfy the passion. On the traditional interpretation, both of these conditions hold with respect to pride and shame, love and hatred.

If the self is a mere philosopher's fiction, as the traditional interpretation affirms, then pride and shame, which take the self as their object, are contrary to reason to the extent that they are "founded upon the supposition of objects, which do not really exist" (T 416). If selves do not exist and are mere fictions we invent to justify our mistake, then pride and shame are irrational because they are directed towards my self, which does not exist. Love and hatred are likewise irrational because they are directed towards another self, which on the traditional view does not exist.
Moreover, if I am making a mistake in referring to myself at different times, then I am irrational to act in ways that presuppose the mistake. If I want to act out of self-love, for example, the very desire itself includes that desire that the self who is the beneficiary of the action is the same self as the one who intended the action. If there are no selves, there are no means available to satisfy that desire. Thus, any means I choose will be insufficient to satisfy it. On the traditional interpretation, therefore, any self-regarding passion will be contrary to reason. For these reasons, the traditional interpretation, which says that attributions are always mistaken, is incompatible with the rationality of self-love and benevolence, as Hume understands these notions.

It is also incompatible with either version of the Body-swap Intuition, which says that consciousness is sufficient for a person's existence through time. On the traditional interpretation of Hume, nothing is sufficient for the existence of a person at different times.

The view the traditional interpretation attributes to Hume is plausibly thought of as a version of CAT. Hume claims that a person's existence consists in a succession of perceptions, conscious episodes. But, on the traditional interpretation, the fact that what we call personal identity consists in consciousness alone means that it is a mistake to refer to a person at different times. This version of CAT implies that the Body-swap Intuition is false.

4.3 An Alternative Interpretation

The second interpretation agrees with Hume's indictment against attributions of perfect numerical identity to persons at different times but holds that Hume holds a view whereby attributions of personal identity may be properly made. The motivation for such an interpretation is easily found.
Hume was aware that there is a connection between the notion of the self as it appears in Books II and III of the Treatise and the notion of the self as it is defended in Book I. He distinguishes the metaphysics of personal identity—what he calls "personal identity as it regards our thought or imagination"—from personal identity "as it regards the passions or the concern we take in ourselves" (T 253). He believes, however, that the two are not at variance with each other. He says "our identity with regard to the passions serves to corroborate that with regard to the imagination" (T 261). It does so, according to Hume, "by making our distinct perceptions influence each other, and by giving us a present concern for our past or future pains or pleasures" (T 261). This last quote seems to indicate that Hume takes himself to have provided a corrected view of the self in Book I. He seems to affirm that one's present concern for one's own good both past and future, that is, personal identity as it regards "the concern we take in ourselves" comports with rather than conflicts with the result of the inquiry of personal identity as it regards the thought or imagination. He gestures at this in a crucial point he make about pride and shame. He says that a "connected succession of perceptions, which we call self, be always the object of these two passions [pride and shame]" (T 277).

It seems, then, that Hume takes himself to have a corrected view of the self, a positive view of identity attributions to changing things. Hume says, "the true idea of the human mind is to consider it as a system of different perceptions or different existences, which are link'd together by the relation of cause and effect. . . ."(T 261).

He continues by comparing the soul to a commonwealth, "in which the several members are united by the reciprocal ties of government and subordination." In comparing the soul to a commonwealth Hume says that "as the same individual republic may not only change its members, but also its laws and constitutions; in like manner the same person may vary his character and disposition, as well as his impressions and ideas without losing his identity" (T 261, emphasis mine). Finally, he allows that correct identity attributions may be made about trees and persons. He says, "An oak that grows from a small plant to a large tree is still the same oak, though there be not one particle of matter or figure of its pasts the same. An infant becomes a man, and is sometimes fat, sometimes lean, without any change in his identity" (T 257).

Commenting on this last passage, Ashley and Stack say "Hume thus seems to say that 'a succession of related objects'—different perceptions and dispositions—can be a single identical object, a person in this case." They argue that Hume is making "ascriptions of personal identity on the basis of some new standard of identity." They point to the fact that Hume says that it is the uninterrupted progress of the succession of experiences that constitutes the “imperfect identity” of changing things (T 256).

Hume does not say much about imperfect identity, beyond distinguishing it from perfect identity (T 254). Moreover, it is not clear that Hume is making identity ascriptions on the basis of a new standard of identity. It is reasonable to assume Hume was drawing rather a standard distinction between numerical identity and other relations we call identity, whether out of convenience or confusion. After claiming that we often mistakenly ascribe perfect identity to a succession of distinct but closely

9. Ashley and Stack, 91

10. Ashley and Stack, 91.
related objects and claiming that a succession of closely related experiences constitutes the imperfect identity of changing things, he gives a similar example, this time drawing the distinction in terms of numerical identity and specific identity, where specific identity seems to be just qualitative identity (T 257). He says that "a noise that is frequently interrupted and renew'd" has only "a specific identity or resemblance, and there is nothing numerically the same, but the cause, which produc'd them" (T 258).

It would seem, then, that Hume's notorious claim that the self is nothing but a bundle of perceptions is not the claim that there are no selves or persons, nor that all attributions of personal identity are false. The claim is, rather, that there is no enduring self, no soul or immaterial substance that we can find that provides numerical identity through time. What we do find are perceptions variously related that do not retain their numerical identity through time, but which none the less constitute a self or person. On this interpretation, selves exist, though they are not perfectly identical souls, selves, or substances as his opponents thought. There is no more to a self than a bundle of perceptions, contrary to what some philosophers say.

The alternative interpretation is consistent with what Hume says about the self-regarding and other-regarding indirect passions in as much as it holds that persons exist at different times. Recall that on the traditional interpretation, attributions of personal identity are always mistaken. But Hume holds that self-regarding passions such as pride and shame, and other regarding passions such as love and hatred take selves as their objects. So, it seems the traditional interpretation implies that the above passions are contrary to reason, since they take objects—identical selves at different times—that do not exist. Hume does not seem to accept this, however.

On the alternative interpretation, personal identity is said to consist in a connected succession of perceptions and, therefore, attributions of personal identity are
sometimes true. Since attributions of personal identity are sometimes true, persons can exist at different times and, therefore, pride and shame, which take selves as their objects, cannot be criticized on the grounds that they are directed towards objects that do not exist.

The alternative interpretation is superior to the traditional interpretation for two reasons. First, it does not give a result that surely would have been so obviously an embarrassment to Hume. Hume’s view as described by the alternative interpretation does not obviously undermine the central project of the Treatise. Second, the alternative interpretation is no mere apologetic designed to avoid an embarrassment of Hume’s own making, but is warranted by what Hume says. He makes plain declarations that are hard to square with the traditional interpretation. He says that “the true idea of the human mind is to consider it as a system of different perceptions or different existences. . . .” (T 261), which is contrasted with the false view that the human mind is a single substance at different times. He says that the same person may vary his character and disposition, as well as his impressions and ideas “without losing his identity” (T 261, emphasis mine), which seems a straightforward denial of what the traditional interpretation attributes. Furthermore he says that “an infant becomes a man, and is sometimes fat, sometimes lean, without any change in his identity” (T 257), which, once again, seems to contradict what the traditional interpretation attributes to Hume. These passages make sense on the alternative interpretation.

I will, therefore, take Hume to hold the view the alternative attributes to him. Specifically, I take Hume to hold that a person or self is a connected succession of impressions about which we may make correct attributions of personal identity. I also accept that Hume’s view cannot be assailed on the grounds that it implies that pride,
shame, love, and hatred are contrary to reason on Hume’s first ground. Since a person can exist at different times, passions that take the self as their object, these passions are not necessarily founded upon the supposition of objects that do not exist.

4.4 Rationality and the pride of the philosophers
Recall that Hume allows a second ground on which to criticize the rationality of a passion, though in the end he grants that the passion itself is not the source of the irrationality but rather the judgment that accompanies it. A passion can be said to be contrary to reason "when exerting any passion in action, we chuse means insufficient for the design'd end and deceive ourselves in our judgments of causes and effects" (T 416). The false judgment concerns the connection between the designed end of an action, and the means chosen to achieve that end.

For Hume, the ends that the passions give us are, strictly speaking, neither rational nor irrational. We could say that for Hume the passions are non-rational. Reason, Hume contends, can neither “justify nor condemn” a passion (T 416). Passions accompanied by judgments, however, become an item for rational evaluation. The indirect passion of pride, for example, is neither in accord with reason nor contrary to it unless it is accompanied by an unreasonable judgment.

Consider once again Hume’s example of pride. Suppose a man is proud of a beautiful house that belongs to him or which he himself has built. Hume says the object of the passion of pride is himself and the cause of the passion is the beautiful house. The beautiful house by itself does not produce pride unless one takes the house to be related to oneself. Hume says, “Beauty, consider’d merely as such, unless plac’d upon something related to us, never produces any pride . . .” (T 279). In this case, the house
is related to the man in that the one who built the house or bought the house is the same person as the man who is now proud. Is pride, therefore, beyond the reach of rational criticism on the alternative interpretation? It depends.

For Hume, claims of the form "X is proud" imply a claim about personal identity. Since these claims are sometimes true, we cannot say that pride per se is contrary to reason. But when pride becomes a motivation for action it can be criticized if it motivates actions that cannot achieve their intended aim. It is not hard to find an example of what Hume would be committed to regarding a unreasonable pride.

Consider the philosophers against whom Hume directs his discussion of personal identity. As I said before, Hume says that there are some philosophers, who imagine we are every moment intimately conscious of what we call our SELF; that we feel its existence and its continuance in existence; and are certain, beyond the evidence of a demonstration, both of its perfect identity and simplicity (T 251).

In so far as these philosophers have self-regarding passions such as pride and in so far as these passions are conjoined with the false belief Hume charges above, these passions can be said to be contrary to reason on Hume’s second ground.

If the pride of the philosophers becomes a motivation for actions, it can be said to be contrary to reason on Hume’s second ground. The reason is, the pride of the philosophers is conjoined with a false belief about the object of the passion, namely that the object of the passion and the subject of the passion are numerically the same. When the pride of the philosophers becomes a motivation for action, their pride is contrary to reason because no action is sufficient for the satisfaction of a passion that includes that the person with the passion and the self that is the object of the passion are numerically the same.
For example, suppose one of these philosophers is proud of his beautiful house. Hume would say the cause of the pride is the house and object is himself. Now suppose the pride of this philosopher becomes a motivation to paint the house a more beautiful color. The passion that motivates the painting of the house can be criticized if no action can satisfy it. If we suppose the philosopher to believe that by painting the house, he himself, the numerically identical self who now has the passion, will receive some benefit, then no action can satisfy that passion. The passion that motivates the painting of the house can be satisfied only if the one with the passion and the object of the passion are numerically the same person, which Hume believes is never the case. Therefore, the pride of this philosopher is contrary to reason.

Actions motivated by the pride of these philosophers are contrary to reason because they rest on an unreasonable belief about the metaphysics of persons. Because they have a false and unreasonable belief that the subject of the passion and the object of the passion are numerically the same, no action can achieve the end motivated by the passion of pride. The same can be said for shame, love, and hatred, each of which takes the self as its object.

This is a result that Hume might approve of. Not only do his opponents have faulty metaphysical views about personal identity, but their faulty metaphysics results in unreasonable passions, which further result in unreasonable actions.

4.5. Summary and conclusions

I have considered two interpretations of Hume’s theory of personal identity and I have argued that the better interpretation takes Hume to hold that persons exist and that personal identity attributions are sometimes true. Furthermore, I have shown how the
alternative interpretation can be used to raise an argument against his opponents' view of personal identity.

The problem that arises on the traditional interpretation occurs because of two commitments. First, on the traditional interpretation, personal identity attributions are always false. Second, claims of the form ‘X is proud’ imply a claim of personal identity. The conjunction of these two commitments implies that pride is contrary to reason.

This line of argument poses not only a serious problem for Hume, but would have provided a serious challenge to commonsense notions of praiseworthiness and blameworthiness. Let me explain.

The following kind of reasoning seems to be sound. If Hauptmann is guilty of murdering the Lindbergh baby, then Hauptmann and the murderer are the same person. This seems to hold whether we are talking about moral guilt or legal guilt. If Hauptmann is not the same person as the murderer, it would surely be a miscarriage of justice to punish Hauptmann, either with the force of law or with the force of public opinion. Not only do claims about moral guilt and legal guilt have this result, claims about desert quite generally have this result. If Lindbergh is praiseworthy for having crossed the Atlantic, it seems to follow that Lindbergh and the one who crossed the Atlantic are the same person.

Suppose it is right to say that by claiming that Hauptmann is guilty of the murder, we are thereby making an ascription of personal identity. If all such ascriptions are false, however, so is the claim that Hauptmann is guilty of the murders. The trouble is, not only is Hauptmann off the hook, so is everyone else. No statement of the form “X is guilty for having done some act A” is true if such claims imply a claim about personal identity. Claims of the form “X is praiseworthy for having done act A,” which
also seem to imply a claim about personal identity are likewise false. X is praiseworthy for having done A only if X and the one who did A are the same person. If personal identity attributions are always false, so are claims about praise worthiness. Well then, if claims of personal identity are always false, it seems that we are neither guilty of anything nor praiseworthy for anything.

This means that the view of personal identity expressed by the traditional interpretation of Hume poses a threat to commonsense notions of praiseworthiness and blameworthiness. If it is true that personal identity attributions at different times are always false, we are neither guilty nor praiseworthy for anything.

The line of argument that runs from the alternative interpretation to the irrationality of the philosophers might pose a threat to commonsense as well. According to the argument I raised in 4.4, the faulty metaphysics of the philosophers results in unreasonable passions, which further result in unreasonable actions. The alleged irrationality here can be traced to the belief of the philosophers that there is something numerically the same in a person’s existence at different times. A passion, such as pride or an action motivated by pride is contrary to reason. It would also seem to follow that when one of these philosophers wants to praise or blame a person, this desire too is irrational, since it is accompanied by the belief that the person who performed the praiseworthy or blameworthy action is the same person as the one now deemed praiseworthy or blameworthy.

I said that ordinary commonsense claims about praiseworthiness and blameworthiness imply claims about personal identity. If that is true, whether these commonsense notions are rationally criticizable depends crucially on what ordinary speakers intend when they claim that some X is praiseworthy or blameworthy. Do they mean “X the succession of impressions”? or do they mean “X the same numerical
self'? If the former, these commonsense notions are vindicated from this charge of irrationality. If the latter, we have an argument that these notions are contrary to reason. If they mean neither, the next task is to figure out what ordinary speakers do intend.

But is it true that ordinary commonsense claims about praiseworthiness and blameworthiness imply claims of personal identity? The main writers I have discussed all believe they do.

This is most obvious in Locke's case. Locke says that "person" is a forensic term which, for him, means that claims about responsibility before the law are materially equivalent to claims about personal identity. The law, for Locke, includes not only civil law but the moral law as well. Dodwell and his opponents contend over whether CAT is consistent with well-accepted beliefs about reward and punishment in an afterlife.

Hume accepts that pride and shame presuppose personal identity. Not only this, but he believes that self-love generally presupposes personal identity. If I act out of self-love, my aim is to benefit the same person who forms the intention.

This leads back to the problem that I considered at the end of chapter 3. Dodwell believes personal identity consists in relations between more basic entities. He calls the basic entities "beings" and advances "sameness of consciousness" as necessary and sufficient for two or more distinct beings to constitute one person. Sameness of consciousness, for Dodwell, may obtain not only between distinct beings at different times but also between distinct beings at the same time.

The problem is, Dodwell must say the first being in each pair below is equally interested in what will befall the second being in each pair.
Pair 1: The one who delivered the Boyle Lectures in 1704 and the one who began the correspondence with Dodwell in 1706.

Pair 2: The one who began the correspondence with Dodwell in 1706 and That Other My-self.

Pair 3: The one who began the correspondence with Dodwell in 1706 and the one who wrote the Fourth Defense in 1708.

Since Dodwell must treat each of the above cases as equally interested, either the egoistic intentions of the first being in each pair can be carried out successfully by the second being in each pair or neither can be.

The traditional interpretation of Hume allows us to see some of the consequences of claiming that there are no successful egoistic actions, the most important of which for my purposes is that this line implies that the Body-swap Intuition is false.

The alternative interpretation takes a first step towards defending the claim that the egoistic intentions of first being in each pair can be carried out successfully by the second being in each pair. It is an advance over Dodwell’s view in that, while Hume’s view is a version of CAT because he holds that a person’s existence consists in related impressions and ideas—mental events—he does not limit himself to memory but includes causation as well. This move allows him to maintain CAT and avoid the criticisms that arise specifically against Dodwell’s consciousness-qua-memory theory.

This is only a step, however. Hume does not, after all, set out to defend CAT as part of an argument for the Body-swap Intuition. He restricts his interest to ordinary, ubiquitous cases such as are described in Pair 1. He shows no interest in defending the rationality of actions in cases such as are described in Pair 3. A defense of the Unqualified Body-swap Intuition requires attention to both. Nor does he consider replication or fission cases. We know that he has no interest in such cases because their

82
primary function was to illustrate responsibility in an afterlife. Hume limits his discussion to personal identity in this life, which is to say, he does not concern himself with the Body-swap Intuition. This leaves it open that Hume's version of CAT picks out only a contingent feature of persons.

Finally, there is the question of circularity. We saw an obvious difficulty in this regard with respect to memory, which seems to presuppose identity. Since Hume believes that memories are among the succession of experiences that constitute a person, there is a prima facie demand that he spell out memory in non-circular terms. As we shall see in chapter 4, meeting this requirement brings a similar burden of spelling out causation non-circularly.

Hume's view provides a transition to the contemporary problem of personal identity. The most influential work has been reductionist in that the persistence of a person is said to consist in the existence of more basic entities. Parfit, Perry, and Shoemaker promote psychological reductionist views, which is to say they endorse CAT. Unlike Hume, they use their version of CAT as an argument the Unqualified Body-swap Intuition.
CHAPTER 5
PARFIT AND THE CONTEMPORARY CAT

In this chapter, I consider the contemporary version of what I have been calling CAT, that of Derek Parfit, John Perry, and Sydney Shoemaker. I have three aims.

First, I describe the views of noteworthy defenders of the contemporary CAT in order to demonstrate their commitment to the Unqualified Body-swap Intuition. Second, I turn my attention to the most influential and most carefully worked out version of the contemporary CAT, that of Derek Parfit, to set out his position, particularly, on his responses to the early modern criticisms of CAT. Third, I evaluate his position.

This evaluation proceeds in two stages. In the first stage, I argue that Parfit's position is undefended. What Parfit must show but fails to show is that the body (with or without the brain) is not an important part of a person's diachronic existence. The conclusion at the end of this stage is that there is no reason to accept Parfit's CAT. The second stage focuses on a family of arguments that trace their pedigree to Butler. These arguments concern the justification of self-love. I argue that the considerations that arise here show that we ought to reject Parfit's CAT. Finally, I say why this, in turn, implies that we ought to reject the Unqualified Body-swap Intuition as well.
5.1 The Contemporary CAT and the Unqualified Body-swap Intuition

The contemporary CAT is a descendant of the versions of CAT found in Locke, Dodwell, and Hume. In the twentieth century, Grice, Quinton, Perry, Lewis, Shoemaker, Parfit, and others,\(^1\) have offered accounts of a person's existence through time in terms of relations among mental events.

Quinton imagines a case in which, he argues, we should say of some person B, "a dark, tall, thin, puritanical Scotsman of sardonic temperament," that he had switched bodies with another person C, "a fair, short, plump, apolaustic Pole of indestructible enterprise and optimism.\(^2\) His argument that we should so regard B and C as having switched bodies is that, confronted with the case Quinton describes, we would so regard them. He says, "we would say that the two had exchanged bodies. . . . our concern and affection would follow the character and memory complex and not the original bodily associate is clear.\(^3\) He concludes that "the soul, defined as a series of mental states connected by continuity of character, is the essential constituent of personality. The

---


2. Quinton, 63.

3. Quinton, 64-65.
soul, therefore, is not only logically distinct from any particular human body with which it is associated; it is also what a person fundamentally is." Quinton's view is a version of CAT since it says that a person's existence consists in relations among mental states or events.

Shoemaker imagines a society in which human bodies cannot remain healthy for more than a few years. The society has coped with this by developing technology that can clone duplicate bodies and store them for a future body change. The body-change, as Shoemaker imagines it, consists in a person's "total brain-state being transferred to the brain of one of his duplicate bodies." He calls this the brain-state transfer (BST) procedure. Shoemaker's story is an advance over Quinton's in that it describes a mechanism that helps us better to understand the view he advocates. He argues that such a society can, with no mistake, regard body-changes as person-preserving. He argues further that if they are right in regarding the BST-procedure as person-preserving, and if they mean by 'person' what we do, then it seems that we ought to regard the BST-procedure as person-preserving.

Shoemaker makes it clear that what is transferred is information only, and that "if we have personal identity here, it is not carried by the identity of any body." But can such a procedure preserve personal identity? The fission criticism raised by Clarke, Butler, and Reid applies here as well.

4. Quinton, 65.
7. Shoemaker, 110.
As described, it would seem that there is nothing preventing us from taking two identical bodies from the stock of clones and investing them both with the same psychological make-up as the original. The original cannot be identical with both copies since the two copies are not identical with each other.

Quinton does not consider the possibility of psychological continuity branching, but Shoemaker and Parfit do. They say that, while we cannot say that both offshoots are identical to the one, we can say that a person has survived as both offshoots, though the one is identical to neither. They claim further that a person's existence through time should be thought of in terms of survival and not identity.

Parfit's version of Shoemaker's BST-procedure, which calls the *Simple Teletransportation* case, asks us to imagine oneself about to step into a machine that will destroy my brain and body, while recording the exact state of all my cells. It will then transmit this information by radio. Traveling at the speed of light, the message will take three minutes to reach the Replicator on Mars. This will then create, out of new matter, a brain and body exactly like mine.8

Parfit believes that this is not a case of personal identity, but it is a case in which what is rationally significant in a person's existence through time is preserved: he believes that one's prospects here are as good as if one had traveled to Mars by spacecraft.9 All that matters is preserved in Simple Teletransportation just as well as it would be if one took a spaceship.

Quinton, Shoemaker, and Parfit all believe that psychological continuity is sufficient for the existence of a person through time, which Shoemaker and Parfit think

8. Parfit, 199.
is best described as “survival.” Furthermore, they all believe that sameness of substance is not necessary for psychological continuity. And, finally, because they all believe that a person can exist in a different body if only psychological continuity is preserved, whether or not that continuity obtains between the same substance or not, they all accept the Unqualified Body-swap Intuition. In the next section, I turn specifically to Parfit's view.

5.2 Parfit's view
In what follows, I set out Parfit's position and the key arguments for that position and turn to the question of whether that position can support the Unqualified Body-swap Intuition. I first describe how Parfit understands the views he opposes and then turn to his account of consciousness.

5.2.1 Parfit's opponents
Parfit describes three positions one could take about personal identity and then presents what he takes to be the best versions of each. The first two views he calls "reductionist views" because they claim "that the fact of a person's identity over time just consists in the holding of certain more particular facts." He rejects the physical view in this form because people can and

sometimes do lose body parts throughout their lives. He says that we cannot reasonably require the existence of the whole body through time, but only "enough of the brain to be the brain of a living person." He calls the following "a better version" of the physical view:

The Physical View: X today is one and the same person as Y at some later time if and only if (1) enough of Y's brain continued to exist and is now X's brain, and (2) this physical continuity has not taken a branching form.\(^1\)

The second kind of view that Parfit calls "reductionist" takes relations among mental events to be the essential feature of a person's existence through time. These are what Parfit calls psychological views of personal identity.

A psychological view of personal identity says that X is the same person as Y if and only if three conditions hold. First, X is the same person as Y if there is "psychological continuity and connectedness" between them.\(^2\) Connectedness is a two-place relation that holds between, for example, the experience of an event and the subsequent memory of the experience, or the forming of an intention and the subsequent action undertaken in virtue of having formed the intention. As Parfit notes, connectedness is not a transitive relation and cannot therefore be a correct criterion for identity.\(^3\) Recall Reid's Brave Officer example, where he says,

Suppose a brave officer to have been flogged when a boy at school for robbing an orchard, to have taken a standard from the enemy in his first campaign, and to have been made a general in advanced life;

\(^{11}\) Parfit, 204.

\(^{12}\) Parfit, 204.

\(^{13}\) Parfit, 206-207.

\(^{14}\) Parfit, 206.
suppose, also, which must be admitted to be possible, that, when he took the standard, he was conscious of his having been flogged at school, and that when made a general, he was conscious of his taking the standard, but had absolutely lost the consciousness of his flogging.\textsuperscript{15}

The problem once again is this. Connectedness is not a transitive relation and identity is. Since the schoolboy remembers the experiences of the brave officer, the schoolboy is connected to the brave officer. The brave officer is connected to the general since the general remembers taking the standard. But since the general does not remember the flogging, he is not connected to the schoolboy. Identity, however, is a transitive relation: If A is identical to B, and B is identical to C, then A is identical to C. Connectedness alone is not, therefore, an adequate criterion of identity. Continuity consists in overlapping chains of connectedness and is therefore transitive. The general is psychologically continuous with the schoolboy because the schoolboy is connected with brave officer and the brave officer is connected with the general. The holding of psychological continuity and connectedness Locke calls “Relation R.”

Second, psychological continuity and connectedness must have “the right kind of cause. According to Parfit, there are three views one can take about what counts as the right kind of cause.\textsuperscript{16} On what he calls the "Narrow version" the right kind of cause must be the ordinary cause, the same brain functioning in the ordinary way.\textsuperscript{17} On what

\begin{itemize}
  \item Parfit, 207.
  \item Parfit does not name a theorist who holds this view, but if one requires the existence of the same brain, the view is of no use in defending the Unqualified Body-swap Intuition, which denies that sameness of substance, brain or soul, is necessary for sameness of consciousness.\
\end{itemize}
he calls a "Wide version", the cause need not be the ordinary cause, but only a reliable cause. If, for example, my strange experience this morning were the result of a scanning/reconfiguration device and this device reliably reproduces the mental life of the original, the causal condition of the Wide version would be satisfied. On what he calls the "Widest version" any cause, reliable or not, will do. Even if the scanning/reconfiguration device I described earlier works only now and then, when it does work, the wide psychological view would say that psychological continuity and connectedness has the right kind of cause.

The third condition is a non-branching clause, which says that R may not take a one-many form. If Relation R takes a branching form so that there are two persons who are R-related to a former person, the non-branching clause is violated.

He states the psychological view this way:

The Psychological View: X today is one and the same person as Y if and only if (1) psychological connectedness and continuity hold between X and Y, (2) this continuity has the right kind of cause, and (3) it has not taken a branching form.  

Parfit describes the physical view and the psychological view as reductionist because they hold that there is nothing more to a person's existence than a brain, a body, and an interrelated series of mental and physical events. What Parfit terms a Further Fact view is non-reductionist because it holds that there is a further fact to a person's existence beyond the existence of a brain, a body, and a series of mental and physical events. He cites the view that one is an immaterial "Cartesian ego" that persists from one moment to the next as the best exemplar of a Further Fact view.

18. Parfit, 207.

Parfit's strategy against these opponents is to argue first, that we must become reductionists, and second, to argue that the non-branching clause is ad hoc. He concludes from both of these that it is not personal identity that has rational significance in a person's continued existence, but R-relatedness, psychological connectedness and continuity among mental events, is what has significance, whether R-relatedness obtains on different substances or not.

5.2.2 On what matters
The locution "what matters" plays a central role in much of what is interesting in Parfit's discussion. What he sets out, ostensibly, to do, is to use "what matters" in some kind of factual sense. When he says that personal identity is not what matters in a person's existence, he sets out to say what is essential to a person's existence. But he is constantly making claims about what has value in a person's existence, using "what matters" as "what has rational significance."20

The relation between facts and values generally is a matter of no small controversy. In Parfit, we are forced to confront it directly because he claims to challenge our beliefs about the nature of persons and our existence through time, and says that when we see the truth we ought to revise our beliefs about what we have a reason to do, both our moral theories and our beliefs about rationality. So, ostensibly, Parfit sets out to defend a claim about the facts and then to derive an ought from those facts—a Herculean, task to be sure.

The locution "what matters" may be used to make factual claims about what is or is not the case. For example, what matters in whether some pile of metal is gold or not is only the number of protons in the nucleus of the metal atoms. "What matters’

may also be used to make claims about what people do care about, or what they would care about if they thought about it carefully. It may also be used to make claims about what people ought to care about. Consider the three cases below.

First, something may be said to matter if there is someone who cares about it. In this first sense, it is correct to say that it matters that Godfrey is failing calculus if it matters to Godfrey, his professor, administration, family, or friends.

Second, something may be said to matter if it would after careful reflection in a cool hour. Suppose that Godfrey is a literature major, does not need the course to graduate, and learned everything about calculus he needed to learn, in spite of not being able to work quickly enough to pass the exams. Suppose that were the various parties for whom Godfrey’s failure matters to reflect carefully and in a cool hour. It might matter to them at first, but upon further reflection they might cease to care about Godfrey’s failing calculus.

The third sense is not as straightforward. In this sense, what matters is thought not to depend on there being anyone for whom it matters, either in fact or after careful reflection. Moral standards are often thought of this way. It might, for example, matter morally that Godfrey passes calculus because he deserves to. Some things might be said to matter morally even if they don’t in fact matter to anyone at all, even after careful reflection.

The first interpretation of the third way to understand “what matters” can itself be understood in two ways, depending on how one understands the nature of the oughtness of normative claims. For some, the oughtness of normative claims just consists in the aims of agents, what matters to me or us or both. This third sense of “what matters” is really not a further sense at all.
For others, there are objective normative facts that rational creatures can
discover, facts about what one ought to do that transcend facts about what I, in fact,
want to do or what I would want to do after careful reflection. This gives a distinctive
third sense of "what matters." Something matters if it would matter to an ideally
rational being successfully confronting the facts.

The first sense of "what matters" is factual: something matters if someone cares
about it. The second sense is counterfactual: something matters if it would matter to
those who reflected carefully. But there is also a normative air about it. If I see by
Godfrey's face that his failing calculus matters to him a great deal, I might say to him
that it does not matter that you are failing calculus and I intend to convey the second
sense of "what matters." By saying it does not matter that you are failing calculus, I am
saying that it would not matter to you if you considered it carefully. This is very close
to saying that it ought not to matter to him.

The third sense of what matters is clearly normative: It matters that people get
what they deserve, even if it does not matter to them. But where this sense of what
matters is intended, if it is not to collapse into the second sense, it also makes a factual
claim. There is some fact that transcends what we in fact want or that we would want if
we if we reflected carefully and in virtue of this fact something might be said to matter
and therefore ought to matter to us.

These are the three important senses of "what matters." As I have said, they
may not even be distinct, in the end. One might argue that what we ought to care about
is just what we would care about given rational deliberation. As we shall see, Parfit
means different things by "what matters." Keeping track of what he means and what his
arguments entitle him to provide an important key to seeing what is wrong with his view of diachronic personhood. I argue that, ultimately, Parfit needs a claim that what matters, matters in the third sense.

5.2.3 Parfitian consciousness, circularity, and the R Relation

In one sense, Parfit opposes both the physical view and the psychological view, since he opposes the non-branching clause both have. Nevertheless, he ends up defending a view that endorses the first two conditions of the psychological view. Since Parfit believes that a person’s existence consists in R-relatedness alone, it is a version of CAT. Since it is a version of CAT, Parfit must spell out R-relatedness non-circularly if his view is to be acceptable at all. In so far as his aim is to argue for the Body-swap Intuition, he must spell out R-relatedness without reference to sameness of body. Finally, in so far as his aim is to argue for the Unqualified Body-swap Intuition, he must spell out R-relatedness independently of sameness of substance, whether brain, soul, or some other.

Parfit addresses the circularity issue, which he attributes to Butler. Parfit states the objection this way:

> It is part of our concept of memory that we can remember only our own experiences. The continuity of memory therefore presupposes personal identity. The same is true of your Relation R. You claim that personal identity just consists in the holding of Relation R. This must be false if Relation R presupposes personal identity.

---


22. Parfit, 220.
His answer to the circularity problem defines Relation R as relations among quasi-memories, quasi-intentions, quasi-desires, etc.) which do not presuppose personal identity.\textsuperscript{23} The definition of these q-states begins with the definition of their ordinary counterparts. Parfit characterizes ordinary memory as follows:

I remember having an experience only if

1. I seem to remember having an experience,
2. I did have this experience,
and
3. my apparent memory is causally dependent, in the normal way, on this past experience.\textsuperscript{24}

The source of the circularity is condition (2) which really means, "I am identical with the one who had the experience," and perhaps (3) which would likely be circular if "the normal way" were spelled out. The new notion replaces condition (2) with

2'. Someone did have this experience,

and condition (3) with,

3'. my apparent memory is causally dependent, in the right kind of way, on that past experience.\textsuperscript{25}

He illustrates this concept with the following case:

\textit{Venetian Memories.} Jane has agreed to have copied in her brain some of Paul's memory-traces. After she recovers consciousness in the post-surgery room, she has a new set of vivid apparent memories. She seems to remember walking on the marble paving of a square, hearing the flapping of flying pigeons and the cries of gulls, and seeing light sparkling on green water. One apparent memory is very clear. She seems to remember looking

\textsuperscript{23} The notion of a quasi-memory was introduced by Sydney Shoemaker, "Persons and their Pasts" \textit{American Philosophical Quarterly} 7, 1970.

\textsuperscript{24} Parfit, 207.

\textsuperscript{25} Parfit, 220.
across the water to an island, where a white Palladian church stood out brilliantly against a dark thundercloud.\textsuperscript{26}

Jane seems to remember these Venetian episodes. But since she has never been to Venice, she did not experience them. Since someone did experience them, condition 2' is satisfied. If one deems the causal chain between Paul's experience and Jane's apparent memory of these Venetian episodes to be an appropriate causal chain, Jane has quasi-memories of Paul's Venetian experiences. Parfit believes that in one sense Jane's quasi-memories give her mistaken information, for she seems to remember herself walking on the marble paving, looking across the water and seeing the thundercloud. He recognizes that in order for Jane's quasi-memories to "give her knowledge about Paul's experiences, she must know [by asking Paul] roughly how they have been caused."

Parfit's R-relation, then, is to be understood as a relation among quasi-memories, quasi-intentions, quasi-desires, etc. rather than their ordinary counterparts. Call these "q-states." Q-states—q-memories, q-desires, etc.—are described so that they do not presuppose the existence of the same person (in order to avoid circularity), the same body (in order to be a genuine alternative to a bodily view), or the same immaterial substance (in order to be a genuine alternative to a Cartesian view).

5.2.4 Personal Identity is not what Matters

Parfit argues that personal identity is not what matters. Until further distinctions are required, "what matters" may be taken to mean "what has rational significance." His claim is that personal identity is not what has rational significance in a person's

\textsuperscript{26} Parfit, 220.

\textsuperscript{27} Parfit, 220.
existence through time. The argument proceeds against the reductionist views glossed above, though it is readily applicable to the Cartesian view as well. He presents a case that satisfies all but the non-branching conditions of the physical view and all three versions of the psychological view. In particular, he presents a fission case in which one person bears the same intrinsic relations to two distinct persons. He argues that the non-branching clause is ad hoc, that only the intrinsic relations matter.

The argument relies on a thought experiment Parfit calls *My Division*.

My body is fatally injured, as are the brains of my two brothers. My brain is divided, and each half is successfully transplanted into the body of one of my brothers. Each of the resulting people believes that he is me, seems to remember living my life, has my character, and is in every other way psychologically continuous with me. And he has a body that is very like mine.28

The argument proceeds by noting that both reductionist views allow that a brain-transplant preserves personal identity. He then cites some controversial research that suggests that each hemisphere of the cerebral cortex has its own stream of consciousness, so that consciousness can be said to divide.29 The argument goes like this. We know that people have survived the loss of one hemisphere or the other either by disease or accident. Thus, if a brain-transplant preserves personal identity as both reductionist views grant, so would a half-brain transplant, since one hemisphere is supposed to be enough to support consciousness. Both reductionist views must treat a

28. Parfit, 253-255.

29. Parfit cites the famous Sperry experiments in which severe epilepsy patients had their corpus collosum surgically severed to alleviate the symptoms of epilepsy. Some of the patients exhibited what some have described as two separate spheres of consciousness. Cf. Thomas Nagel's "Brain Bisection and the Unity of Consciousness" in *Personal Identity*, ed. by John Perry. For a more cautious treatment of this phenomenon, see, Dennett's *Consciousness Explained*, chapter 13 and Kathleen Wilkes's *Real People* chapter 1.
half-brain transplant as preserving the identity of the brain donor. But, of course, then we must allow that if either hemisphere of the brain were transplanted rather than the other, the intrinsic relations specified by both criteria are satisfied. If both are transplanted into different bodies, the non-branching clause is violated. But why, apart from the fact that it is not a case of personal identity, should it matter that the outcome we want (X can remember Y's past) is duplicated? Parfit asks rhetorically, "how could a double success be a failure?"  

The conclusion Parfit wants to draw from this is that personal identity is not what matters in the sense that personal identity is not what is essential to a person's existence through time. On the reductionist views, the fact that we have unique successors is a contingent fact about us.

5.2.5 What matters is R with any cause

Parfit argues that what does matter is Relation R, regardless the causal mechanisms that contribute to its production. As we shall see, when Parfit says that Relation R is what matters, he means that Relation R has rational significance and therefore is what ought to matter to us.

Parfit's answer to the circularity problem gives him the resources to describe a person's existence though time without reference to a persisting body or immaterial substance. His Relation R is a relation among q-states. Since it is possible that there are q-memories that are not memories, he must show that these q-states are, as Parfit says, "as good as" their ordinary counterparts.  

30. Parfit, 256.

31. Like the ambiguous "what matters" the evaluative relation "as good as" invites the question "to or for whom?" or "with respect to what?" I take this up in 5.3.1.
as something else in one way and not in another way. A rock may be as good as a hammer for pounding a tent stake and not for removing a nail. In other words, he must show that q-remembering that is not remembering is just as good in the relevant way as remembering.

Suppose that I awoke in a different body as the result of the scanning-reconfiguration device I mentioned earlier. If so, I was q-remembering this morning and not remembering the past experiences and past intentions of some person. According to Parfit, this is good enough. The fact that Relation R is not brought about in the usual way is not important. In his words, Relation R with any cause is what matters.32

It should be emphasized that this is a crucial argument for Parfit. Unless he establishes the claim that q-remembering, q-intending, etc. are as good as their ordinary counterparts, he has not established his reductionist view of persons.

The argument that the ordinary cause does not matter uses an analogy between the ordinary cause of Relation R and the ordinary cause of seeing. Suppose one were to lose one's sight and then have mechanical eyes installed that give one experiences qualitatively indistinguishable from those had with natural eyes. Parfit asks whether we should say that such a person is seeing. He answers,

If we insist that seeing must involve the ordinary cause, we would answer No. But even if this person cannot see, what he has is just as good as seeing, both as a way of knowing what is within sight, and as a source of visual pleasure. If we accept the Psychological Criterion, we could make a similar

-------------------
32. See, for example, Parfit, 215 ff.
claim. If psychological continuity does not have its ordinary case, it may not provide personal identity. But we can claim that, even if this is so, what it provides is as good as personal identity.\textsuperscript{33}

His point is that if one's visual experiences with artificial eyes are qualitatively indistinguishable from one's visual experiences with natural eyes and artificial eyes provide the same experiences of one's environment as natural eyes, then, with respect to what is important about seeing, artificial eyes are as good as natural eyes. The reason is, the experiences given by one are indistinguishable from the other and the relation between the experiences and the world are the same. Parfit claims this case is analogous to worries about the cause of Relation R. An artificial cause of R is as good as the ordinary cause even if it does not provide personal identity. If the ordinary cause is not what has rational significance, then the continuity of the brain does not matter. Thus regardless of the cause of my strange experience this morning, all that matters is preserved between me and the person whose life I q-remember, even if the process were duplicated.

It seems that most of the trouble that Parfit raises, he raises in the context of fission. Of course, if fission takes place, we cannot say that we have personal identity. Does Parfit believe that in the absence of fission destruction and replication preserves personal identity? In other words, does Parfit believe that personal identity consists in non-branching R-relatedness. Parfit's Simple Teletransportation case is just such a case. Of this case he says, "we should probably decide not to call my Replica me"\textsuperscript{34} even though it would be correct to say that one has survived as one's replica. Parfit

\textsuperscript{33} Partit, 209.

\textsuperscript{34} Partit, 215.
does not endorse any view of personal identity because he thinks a non-branching clause is required for any theory of personal identity, and he thinks the non-branching clause is ad hoc.

5.3 Analysis
Let me summarize the discussion so far. Parfit promotes his version of CAT as supporting the Unqualified Body-swap Intuition. I have granted for the sake of argument that his Relation R meets the amnesia criticism, the false memories criticism, and the circularity criticism. On this last issue, I emphasize the point once again, that in granting Parfit success against the Circularity Criticism, I am granting that success on this front precludes any appeal to the ordinary cause, the same brain, body, or soul. I am not convinced that R-related mental events can be spelled out non-circularly but if they can, they must be spelled out without reference to their relation to the same person, the same soul, or the same brain.

As we shall see, it all comes down to a question of value, a question of what matters in more than a purely factual sense of “what matters,” a question of whether R-relatedness is as good as ordinary survival. Each of the evaluative terms is ambiguous, having both normative and non-normative senses. Something is valuable if some cares about or if it is worth caring about. Something matters if there is someone who cares about it or if one ought to care about it. Something may be as good as something in one respect and not another. The significance of the ambiguity will be clear shortly.

5.3.1 Is personal identity what matters?
The reductionist views Parfit challenges are vulnerable to the claim that personal identity is not what matters because, according to them, the locus of personhood—the
brain for the physical theorist and R-relatedness for the psychological theorist—can divide. A non-branching clause is required because, as Parfit holds, a criterion of identity must preserve the logical properties of identity. The continuity condition of the psychological criterion is there because continuity, unlike connectedness, is transitive. The non-branching clause is needed because identity is a one-one relation, but it is ad hoc because the intrinsic relations that reductionists take to be constitutive of personal identity can branch. While it avoids the problem of branching, it leaves us with the following counter-intuitive result. If A's brain is divided and successfully transplanted into separate bodies resulting in persons B and C, the non-branching clause is violated, so neither view would say that B or C is identical with A. Both views, however, are committed to saying that A would have been identical with B or C had the other died on the operating table. If B had died on the operating table, A would have been identical with C, but only contingently so. Because, as Parfit says, a criterion of personal identity must preserve the logical properties of identity, it must be transitive and one-one, as Parfit notes, but it must also preserve the necessity of identity, which he does not note. And both reductionist views fail in this latter regard.

That means, I claim, that Parfit's conclusion that personal identity is not what matters comes too fast. One may object that the reductionist views considered are not correct accounts of personal identity because they do not preserve the logical properties

35. Parfit, 206.


37. The idea is if a = b, then necessarily a = b. See Saul Kripke, Naming and Necessity (Cambridge, Massachusetts: Harvard, 1980).
of identity. All Parfit is entitled to is the claim that personal identity, as it is
misdescribed by his reductionist opponents, is not what has rational significance.

5.3.2 Does R alone matter to us?
Parfit defines Relation R as continuity among q-events, q-memories, q-intentions, etc.
He does not often make it explicit that the R-related mental events must be described in
their q forms, but it is clear he is committed to it. Q-states have been described so that
they don't presuppose the identity of a person, a body, or an immaterial substance. He
has earned the right to use this relation in his account of the diachronic existence of a
person but he must show further that R-resources alone are up to the task of describing
what matters in a person's existence, in one or more senses of "what matters." Since R-
relatedness can, by definition, obtain between different bodies, it must be shown that R-
relatedness that does obtain between different bodies is just as good in the relevant
sense as R-relatedness that obtains between the same body at different times.

As we have seen, he uses the analogy with ordinary seeing to argue that R with
an artificial cause is as good as R with the ordinary cause. We could grant that what
one does with artificial eyes is as good as seeing, but not merely because the
experiences themselves are indistinguishable from the ones given by natural eyes, but
because, ex hypothesi, they put us in the same relation to the world as natural eyes do.
If Jack's experiencing the apple tree with his artificial eyes is as good as with his natural
eyes, it is because Jack has this experience in the presence of an apple tree with either
set of eyes. It is not the having of experience that matters, but that such experiences
allow us to form true beliefs, such as, "there is an apple tree before me."

We are invited to compare the ordinary cause of seeing with the ordinary cause
of Relation R. It is reasonably clear how we should evaluate the goodness of visual
experiences. But how shall we evaluate the goodness of Q-events? We should ask not only whether the relevant mental states stand in the right relation to each other, but also whether they stand in the right relation to the world.

Beliefs, it would seem, are easy enough: a belief is good if it is true or warranted. If the person in my original story whose experiences are R-related to mine believed that Napoleon was defeated at Waterloo, my belief is as good as his if I have the same belief as he and the belief is true. These two R-related events are as good with the unusual cause as they are with the ordinary cause. What if my belief is that I saw the presidential motorcade move through the city last year? Suppose that the person with whom I am R-related believed this and that his belief is true. Is my belief as good? Well, is it true or warranted? It is not obvious that it is. Ordinarily, to say that I remember seeing the presidential motorcade, is to say that I remember my seeing it. In the present case all we are allowed is the claim that I q-remember my seeing it. We want to know if q-remembering is as good as remembering. And so the question has raised itself again: In what ways are q-states as good as their ordinary counterparts?

The answer is clearer in the case of the aims one tries to achieve. Suppose the person with whom I am R-related went to bed wanting to spend the morning working in the garden. Ordinarily, if I have a desire and believe that doing certain things will satisfy that desire, I get what I desire if I carry out the task successfully. Is that the case here? The person with whom I am R-related wanted to plant the garden. Does he get what he wants if I plant the garden? It depends. If his desire to plant the garden was merely the desire that the garden be planted, then he gets what he want no matter who plants it. If he wants that the garden be planted in exactly the way in which he intends to plant it, an R-descendent can carry out his intentions in a way that gives him what he wants. But if what he wants is to plant it—to be out in the fresh air with the sun shining
on his back—then his desire cannot be satisfied by this merely R-related descendant. His desire cannot be satisfied unless he participates in its satisfaction. Call such desires \textit{essentially self-regarding desires}.

This suggests that some q-related mental events are as good as their ordinary counterparts and some are not. This point can be made more vividly with the following. Suppose that the whole story of my strange experiences this morning is as follows. Last night, a sadistic psycho-surgeon visited my bedroom, took the person with whom I am R-related back to his lab, scanned his brain and reconfigured the embodied brain of a different person to be R-related to the original. Suppose further that the original is alive and well, still in the lab and watching my morning activities by closed-circuit television. In what ways, then, is mere R-relatedness—R-relatedness that obtains between different substances—as good as ordinary survival?

First, consider this issue in terms of aims, whether desires or aims that reason gives: when does the person in the lab get what he aims at? Second, I want to consider this issue in terms of beliefs: what are the appropriate beliefs to have in the light of the first issue?

I think it is clear from what I have said before that the person in the lab can be said to get what he wants in the following respects. Suppose he watches me plant the garden. He gets what he wants in that it is getting planted, and it is being done in exactly the way he would have planted it. But he also wanted to enjoy the fresh air and feel the sun on his back. In short, he wanted to enjoy planting the garden, and in that respect he does not get what he wants. He does not get what he wants because his desire to plant the garden is essentially self-regarding. He does not get what he wants unless he participates in the satisfaction of the desire. If Relation R preserves all that matters, all that matters is preserved between me and the person whose life I q-
remember. It is clear, however, that the relation that typically obtains between an essentially self-regarding desire and the satisfaction of that desire is better than the one that obtains between an essentially self-regarding desire and the satisfaction of the corresponding q-desire. Therefore, all that matters with respect to at least one desire is not preserved between me and the one whose q-memories I have.

The person in the lab with whom I am R-related does not get what he wants unless he enjoys the sun on his back. The satisfaction of this desire requires more than R-relatedness for its satisfaction. One's beliefs should, therefore, be brought in to line with this. If I believe that by carrying out my aims—those that I find myself to have—I will achieve the ends for which they were intended, I believe falsely in all those cases in which I try to carry out the essentially self-regarding aims of the one whose q-memories I have.

Now we can see why R with an unusual cause is not always as good as R with the ordinary cause. It is not merely that we have mental states that stand in special relations with one another that makes a mental life such a special thing to have, but that these mental states stand in certain relations with the world. The having of beliefs is indeed a fine thing, but finer still is the having of beliefs that are true. Moreover, we value things—objects, states of affairs—and we can form intentions and act to maximize value. It is a fine thing that we can act with the aim to maximize value. But finer still that we can actually maximize value. A merely R-related descendent—one who is R-related but shares no substance with one's predecessor—has false beliefs about himself: about himself and about his existence through time. Moreover, acting on those beliefs does not achieve essentially self-regarding aims.

I believe that this is enough to show that we ought to reject Parfit's view as it pertains to our beliefs and aims. It is simply not true that what actually matters to us is
Relation R with any cause. Ironically, Parfit agrees. He says that the natural view is that personal identity is what matters. He grants that personal identity is what matters to us and that the ordinary cause of R does matter to us but he also holds that these things ought not to matter. Parfit promotes his view as revisionary as standing against what we are inclined to believe and as urging revision of our moral and rational aims.

5.3.3 Is R alone what really matters?

In order to defend his view in its normative form—what ought to matter to us—Parfit has two options. First, he can argue that our aims are grounded in false beliefs. Second, he can argue that the aims themselves are intrinsically irrational.

Parfit most often promotes his reductionist view as proceeding against our beliefs. He states the main project of *Reasons and Persons* as though it proceeds from the metaphysics of persons to normative claims about what we have a reason to do. He says,

I believe that most of us have false beliefs about our own nature, and our identity over time, and that, when we see the truth, we ought to change some of our beliefs about what we have a reason to do. We ought to revise our moral theories, and our beliefs about rationality.^[39]

This suggests that the revisionism Parfit urges is first metaphysical and then normative, that the envisaged normative revision follows from the metaphysical. The following consideration, however, show that the mistake Parfit charges is deeply normative.

In the case I have been considering, what mistake can he charge against me and the person with whom I am R-related? My R-related predecessor believes that he is not

---


I, and I believe that I am he. Surely, he is right and I am wrong. He believes that there is a difference between the satisfaction of his self-regarding aims from the beginning of his life until today and their satisfaction on this day. I believe that by achieving the aims I find that I have, I achieve the aims for which they were intended. Once again, he is right and I am wrong. The conclusion, then, is that Relation R does not preserve all that matters.

It is plain that were Parfit's view correct, some revision of our aims is required. If R is all that matters and R does not accommodate essentially self-regarding aims, then essentially self-regarding aims do not matter and they should be disregarded. Even if I knew the whole story of my strange experiences this morning, I should plant the garden and believe that in so doing I can carry out all the aims that matter.

Parfit often promotes his view as opposing not our beliefs about ourselves simpliciter, but our beliefs about the rational or moral significance of certain facts about our existence. When Parfit says that what matters is Relation R with any cause, he means that R alone is what it is "rational to care about in our future." He takes the question of what it is rational to care about in one's future to be equivalent to the following: "Suppose that I am an egoist, and that I could be related in one of several ways to some resulting person. What is the relation that would justify egoistic concern about this resulting person?" (283). He believes that all it is rational to care about is preserved between me and my R-related predecessor. Thus, he should accept that what I have called essentially self-regarding aims are not rationally justified.

If non-branching R-relatedness were all there is to a person's existence, Parfit might be entitled to this normative claim, especially if it could be shown that we believe otherwise. The argument would go like this. There is nothing more to personal

---

40. Parfit 282.
identity than non-branching R-relatedness. We believe that there is more to a person's existence than non-branching R-relatedness and our concern for ourselves is rooted in this mistaken belief. We are inclined to believe, therefore, that there is something ordinarily preserved in our existence that is not preserved in a case in the scanning/reconfiguration procedure. But everything present in one is present in the other and thus treating the cases differently is rationally arbitrary.

This route to the normative claim is not available to him, however, because he believes that more than R-relatedness is ordinarily preserved in a person's existence. Parfit is a reductionist, and therefore accepts that there is nothing more to a person's existence than a brain and a body and an interrelated series of mental and physical events. If a brain, body and an interrelated series of mental and physical events are present in an ordinary case of diachronic personhood, there doesn't seem to be anything rationally arbitrary about treating a case in which an interrelated series of mental events alone is preserved differently from an ordinary case in which other things are preserved as well; unless Parfit can show otherwise.

The claim that Parfit must defend, then, is that we ought to pay attention only to R-relatedness because, with respect to the presently undefined relevant standard, R is the only thing that has value. He must argue, then, that a particular brain and body are not important.

41. Parfit does make this claim, 262. In light of the fact that Parfit does not endorse a view about personal identity, it is best to take him as claiming that what the physical view and the psychological view call personal identity is nothing more than non-branching R-relatedness, not that personal identity is described so correctly.

42. Parfit, 210-211.
What, then, is the relevant standard? Parfit suggests that the relevant standard is what really matters to us, what we really care about upon careful reflection in a cool hour. The argument is that, as a matter of fact, we don't really care about a particular body, or would not if we thought about it enough. He says,

I believe that physical continuity is the least important element in a person's continued existence. What we value in ourselves and others, is not the continued existence of the same particular brains and bodies. What we value are the various relations between ourselves and others, whom and what we love, our ambitions, achievements, commitments, emotions, memories, and several other psychological features.  

Parfit seems to be suggesting that what we really value, upon cool and sober reflection, is not a particular brain or body, but, rather psychological relations between ourselves and others. If Parfit it right that in fact we do not value a particular body and we do value these other relations, and if R-relatedness preserves these various relations, then R preserves what really matters to us. The normative version of his thesis, therefore, requires the claim that we do not value a particular brain or body in our concern for persons.

Parfit seems to think that if we value these psychological features, we do not value a particular brain and body, that either one values persons for their bodies or for their minds, but not both. He is quick to disparage love for a particular body by equating it with lust and to praise love for an R-related descendant as "the best kind of love." He does not argue for this either/or claim, nor does he present an argument that we do not love a particular body. So, the claim that what really matters to us is R alone is undefended.

43. Parfit, 284.

44. Parfit, 295.
The remaining option for Parfit is to make a straightforward normative claim: what ought to matter to us is R alone. I consider this strategy in the next section.

I summarize the argument of this chapter so far. Parfit believes that what matters in a person's existence is R-relatedness alone. He also believes that, since R might take a branching form, it is wrong to think that it is identity that matters in a person's existence through time. I have argued that Parfit is not entitled to draw the conclusion that personal identity is not what matters. He is entitled only to draw the conclusion that personal identity, as incorrectly described by the two reductionist views, is not what matters.

I have also argued that the claim that what matters is R alone is ambiguous. If we understand “what matters” to mean what matters given our beliefs and aims, it is simply false: It is not the case that what matters to us is R alone. I also argued that the normative version, which says that what really matters or what ought to matter to rational creatures is R alone, requires the claim that either we do not or we ought not to value a particular body. Since this is a claim he has not defended, his CAT is undefended on this point as well.

Parfit promotes his view as a metaphysical view about what a person's existence through time "necessarily involves or consists in" but the metaphysical discussion is limited to the claim that personal identity is not what matters, a claim I have argued is undefended. The remaining discussion centers on values, either those that we have or those that we ought to have.

5.4 The Special Concern we have for ourselves

In chapter 3, I argued that Butler’s Interest Argument concludes that on Dodwell’s version of CAT, there are no successful egoistic actions. A different but related argument has been attributed to him by contemporary reductionists such as Parfit, Perry, and Whiting. The argument attributed to Butler has to do with what Butler calls self-love, what modern writers call the "special concern we have for ourselves." John Perry describes the specialness of self-concern this way. He says,

Most of us have a special and intense interest in what will happen to us. You learn that someone will be run over by a truck tomorrow; you are saddened, feel pity, and think reflectively about the frailty of life; one bit of information is added, that the someone is you, and a whole new set of emotions rise in your breast.

Whiting describes it this way:

Most of us believe that we have reasons to care about our own future selves and that these reasons differ, either in degree or in kind, from the reasons we have to act on behalf of one another.

These writers believe Butler is making the following kind of charge. On a psychological view, there is no reason for the special concern we have for ourselves.

---


47. Perry, 67.

48. Whiting, 547.

The demand for a reason may be a demand for an explanation of why we care about ourselves or it may also be a demand for a justification.

Perry’s influential treatment of Butler’s Interest Argument takes Butler to be demanding an explanation of the special concern we have for ourselves. He takes Butler to charge that a psychological view cannot explain our interest in an afterlife or why we care about ourselves tomorrow.\(^{50}\) Perry advances a view of personal identity he claims is, “a descendant of the memory theories of Locke, Grice, and Quinton,”\(^{51}\) and he argues that Butler’s alleged demand for explanation can be met with the resources his psychological view provides. This, he believes, answers the argument attributed to Butler.

Other psychological theorists, such as Whiting and Parfit, take Butler’s argument as a demand for a justification for the special reasons we have. In its crude form, the objection is that on a psychological view of personal identity, there are no special reasons why one should be concerned for oneself. Alternatively and equivalently (and just as crudely), on a psychological view the special concern we have for ourselves is not justified.

Perry argues that Butler’s demand for an explanation can be met but there are features of our special concern for ourselves that are not justified. Whiting, on the
other hand, argues that "psychological accounts of personal identity can explain, and in some sense justify, the sort of concern that each of us has for her own future selves."\(^{52}\) Parfit, as we shall see, straddles the fence between these positions.

These writers draw a distinction between concern for strangers and the special concern we have for ourselves. Perry's explanation of the special concern we have for ourselves is in terms of the aims or projects we have. The new set of emotions that rise in my breast when I learn that some calamity is to happen to me is explained, according to Perry, by the fact that a calamity that happens to me will interfere with my projects. He says,

I expect to have tomorrow much the same desires, goals, loves, hates—in a word, projects—as I have today. There is, in the normal case, no one as likely to work on my article, love my children, vote for my candidates, pay my bills, and honor my promises, as me.\(^{53}\)

This, he believes, explains why we value personal identity: The best way to have one's projects advanced is to do them oneself. In many cases, such as finishing an article, a favorite example among philosophers, it is ordinarily the only way. In my imaginary story, however, I am just as well able to complete the article begun by the one whose past I seem to remember. If I can carry out all the projects of the one whose past I seem to remember, all that matters, according to Perry, is preserved between us.

As for what I have called essentially self-regarding aims, those that cannot be achieved by another, even by a duplicate, Perry says there is no justification for these. The example he gives is the aim that this article be completed. He says, "I want not

\(^{52}\) Whiting, 547.

\(^{53}\) Perry, 75.
merely that this article be completed, but that it be completed by me."^54 He argues that
the only justification for this project (that the article be completed by me) is the project
that the article be completed. The special concern we have for ourselves is justified
only in so far it is typically efficacious in bringing about other ends.

Perry's strategy is to explain and justify self-concern (the special concern we
have for ourselves) in terms of aims that do not include ourselves. He regards any aim
that cannot be justified in terms of non-private projects as unjustified. Whiting and
Parfit try a different strategy, adding a third category, the special concern we have for
loved ones. Whiting says,

We ordinarily take ourselves to be justified in having special
concern for other persons to whom we stand in certain special
relation—for example, our parents and children, friends and lovers.
This suggests that the psychological continuity theorist might take
friendship as a model for how psychological continuity can justify
concern.\(^55\)

This is what Parfit says in response to a claim he attributes to Chisholm that, "in the
absence of personal identity, psychological continuity provides no reason for special
concern."\(^56\) Consider Parfit's My Division in which one person becomes two. Call the
resulting persons *Lefty* and *Righty*. Parfit says,

Suppose I shall be Righty. Lefty will not be a mere stranger. My
relation with Lefty is and will be very close. We might claim that,
compared with the future of a mere stranger, I have reasons to be
more concerned with Lefty's future. . . .[it is] plausible to claim that
psychological continuity gives me a reason for special concern.\(^57\)

\(^54\) Perry, 80.

\(^55\) Whiting, 557.

\(^56\) Parfit, 309.
The contrast in those cases is between the concern we have for strangers and concern for an R-related descendent with whom one is not identical. Parfit admits that the special concern that is appropriate for a mere R-related descendent is "not like the special concern we have about our own future." The kind of concern R justifies is concern for "someone else who can, in various ways act on my behalf."^58

Parfit distinguishes generalized benevolence—the concern for an arbitrary member of one's species—from the special concern we have for ourselves and argues that one's relation with one's replica is more like one's relation with oneself than the relation between oneself and a stranger. He concludes that one has a reason for special concern for one's replica. The argument for this draws on the "special concern" we have for our children and loved ones:

> Given the nature of our relation to our children, or to someone we love, we can plausibly claim that we have reasons to be specially concerned about what will happen to these people. And the relations that justify this special concern are not the deep separate fact of personal identity. If these relations give us reason for special concern, we can claim the same about Relation R. We can claim that this relation gives each of us a reason to be specially concerned about his own future.^59

Parfit, like Perry, draws a distinction between concern we have for strangers and the special and intense concern we have for ourselves. Parfit and Whiting draw the further distinction between concern for strangers and the special concern we have for loved ones. Parfit and Whiting argue that since the special concern we have for others is

---


58. Parfit, 310.

59. Parfit, 311.
justified and does not rely on the identity relation between the one with the concern and the loved one, the special concern we have for ourselves may likewise be justified towards an R-related duplicate. I want to consider these three kinds of concerns for persons in order to show a fatal problem for this strategy.

Suppose that my wife and I visit the dentist for a routine examination and that after we have both been x-rayed and examined I overhear some key words between my dentist and an assistant. I hear the words "problem" and "great pain" and I assume that the conversation is not about me or my wife, for we are there for a routine examination. As I think about the patient whose grave plight is just ahead, I am moved, as I sometimes am when I reflect on the human condition. When I hear the word "Lent" in their conversation, I assume that they are referring to my wife, for I am sure that I am there for a routine visit. At this news, my concern changes from a general sympathy for the woe and weal of some arbitrary member of my species, to a specific concern for one of those members, one with whom I have a special relationship. When at last I hear the word "Mister" my concern changes yet again. Call it fear; call it terror; call it what you would like. The feeling that accompanies this latest news is different from the other two cases, for the pain will not come to a mere stranger; nor will it come to a loved one whose pain I can experience only vicariously or sympathetically.

The suffering I will undergo is not sympathetic or empathetic, but the thing itself. This feature of self-concern is not merely different in degree from benevolent concern, but different in kind. I may be more concerned for my wife than I am for myself, preferring that I suffer rather than she. But my fearful anticipation is, like my desire to plant the garden, essentially self-regarding. Even my R-related ancestor watching this episode via closed-circuit television cannot appropriately feel anything more than sympathy or empathy for my upcoming ordeal.
My fearful anticipation and my desire to plant the garden are forward-looking and essentially self-regarding. The nasty set of emotions that arises in me at the prospect of my suffering is appropriate only if I have reason to believe that the suffering will come to me. My desire to plant the garden is not satisfied by another's planting the garden and, hence, my eager anticipation of doing it is not appropriate if someone else will do it.

Perry, Parfit, and Whiting must treat the reasons essentially self-regarding aims generate as irrational. The contemporary CAT is committed to holding that all that really matters in a person's survival—what ought to matter to us—is preserved between me and the one whose past I seem to remember, and is thereby committed to holding that whatever distinctions are not describable in R-terms have no rational significance. If we continue to make such distinctions, according to the contemporary CAT, we are drawing rationally arbitrary distinctions.

The position I am attributing to Parfit and Perry holds that the special concern we have for ourselves, which I have argued includes essentially self-regarding aims, is irrational (making essentially self-regarding aims irrational as well) but, it is claimed, there is a rationally justifiable special concern that one can have towards an R-related duplicate.

Recall that Whiting's proposal is to use the special concern we have for others as "a model for how a psychological view can justify concern."60 Perry's proposal is to justify the special concern we have for ourselves in terms of the projects we have. Parfit's proposal is similar. He believes the special concern we have for loved one's is justified, and that the relations that justify this concern are not "the deep further fact of personal identity"(311).

60. Whiting, 557.
These writers make these proposals as a response to what I have been calling Butler’s Interest Argument. I want to remind the reader of some of the details of that argument in order to evaluate position taken by Parfit and Perry on the special concern we have for ourselves.

I take Butler to argue that Dodwell’s view implies there are no successful egoistic actions. The reason is that Dodwell would say that if the one whose past I seem to remember were destroyed after he went to sleep and his mental life were reproduced in the brain that I have, I would be the same person as the one whose past I seem to remember. Butler charges that if Dodwell’s view is true, the self of today is no more interested in what will befall the self of tomorrow than in what will befall anyone else. This charge relies on the claim that the action undertaken by a merely R-related descendent is not an interested action. Since Dodwell thinks all so-called egoistic actions are like those carried out by a merely R-related descendent, there are no successful egoistic actions.

If Butler’s argument is allowed to stand, it is easy to see an argument that the special concern we have for ourselves is not justified. If there are no successful egoistic actions, the special concern we have for ourselves that motivates egoistic actions is plausibly thought to be contrary to reason. The reason is, the special concern we have for ourselves motivates actions that cannot achieve their desired end.

Notice that there is nothing normative about Butler’s argument, even if one can draw normative conclusions from it. It is a claim that Dodwell’s view implies a straightforwardly factual mistake: we know what interested or egoistic actions are and we know what benevolent actions are, and the actions carried out by a merely R-related descendent are not successful egoistic actions.
The contemporary CAT, however, does not imply the factual mistake Dodwell’s CAT does. Parfit and Perry do not believe that, as a matter of fact, interested actions can be carried out by a merely R-related descendent; nor do they believe that a merely R-related descendent is properly called the same person as the one of whom it is a duplicate. So, they don’t believe their view implies there are no successful egoistic actions.

They do believe, however, that what is rational about egoistic actions can be carried out by a merely R-related descendent and that what cannot be carried out by a merely R-related descendent is not worth caring about. If the one whose past I seem to remember were told that the plans he made today would be carried out by me, it should make no rational difference to his person-related plans, and it should make no difference with respect to his genuine person-related reasons. It should make no difference, for example, that the rewards of his exemplary life of prudence should come to me; nor should it make any difference to me. I should enjoy reaping the fruit of the prudential aims and actions of the one whose past I seem to remember.

That is what I believe the contemporary CAT is committed to. But they are also committed to showing that essentially self-regarding aims do not provide genuine reasons for acting, since essentially self-regarding aims cannot be successfully carried out by a being that is merely R-related to the being that intended them. What this means is that the contemporary CAT must say that the special concern we have for ourselves that includes essentially self-regarding aims is not rational, but there is a justifiable special concern that one may have toward a merely R-related descendent.

So, that describes the task set before the contemporary CAT. The proposal, once again is to argue that because the special concern we have for others is justified and does not rely on identity but other relations, a kind of special concern is justified
towards a merely R-related descendent as well. Parfit, Perry, and Whiting attempt to argue that since the concern we have for loved ones is special and justifiable, and does not rely on there being identity between the one with the concern and the one for whom the concern is directed, there is a justifiable special concern one may have for an R-related descendent with whom one is not identical.

I agree that the concern we have for loved ones is special and justifiable, and does not rely on identity, but we must look more carefully at this other kind of special concern. The reason is that the contemporary CAT requires more than the claim that the special concern we have for loved ones is justifiable and not dependent upon there being identity between me and object of my concern. The contemporary CAT must also say that the special concern we have for loved ones is dependent only upon there being R-relatedness between the object of my special concern at one time and the object of that special concern at a different time. The reason is, the contemporary CAT claims that R-relatedness is all that it is rational to care about in a person’s existence, whether one’s own or another’s. In other words, if R alone has rational significance in a person’s existence with respect to the justifiable special concern has for oneself, then R alone has rational significance with respect to the justifiable special concern one has for a loved one.

Recall that Parfit says this when he claims that physical continuity is “the least important element in a person’s continued existence”\(^6\) I quote this claim again:

What we value in ourselves and others, is not the continued existence of the same particular brains and bodies. What we value are the various relations between ourselves and others, whom and what we

---

61. Parfit, 283.
love, our ambitions, achievements, commitments, emotions, memories, and several other psychological features.

I said before that this claim is undefended. The debt left unpaid in that context—it left his version of CAT undefended—is payable once again. Moreover, it is an unpaid debt due in normative coin. He must show that the various relations that really matter or ought to matter are describable in R terms alone. It won’t do for Parfit simply to claim that the things we value in a loved one’s continued existence are ambitions, achievements, commitments, emotions, memories, and several other psychological features. He must show that what we really value or ought to value are q-ambitions, q-achievements, q-commitments, q-emotions, q-memories, and other q-events.

Now, it is false that what actually matters to us is R and not identity. Furthermore, nothing that Parfit has said shows that R and not identity would matter to us, even upon careful and sober reflection. That means we have no reason to accept Parfit’s reductionism. I believe the following two considerations show that we ought to reject Parfit’s reductionism.

First, recall that Parfit’s reductionism implies that the special concern we have for ourselves is not justified and the special concern that we can have for a merely R-related descendent is justified. An argument either that the special concern we have for ourselves is justified or an argument that the special concern we one have for a merely R-related descendent is not justified would suffice to demonstrate the falsity of Parfit’s reductionist view of persons. The falsity of either of these implies the falsity of Parfit’s view.

I believe that if there is a justifiable special concern we may have for a merely R-related descendent, then the special concern we have for ourselves is justified as

well. And that is what I shall argue. If I am right, that, in turn, implies that at least one of the two claims above implied by Parfit's reductionism is false.

Recall that I said that contemporary reductionists recognize the distinction between a generalized benevolence, the special concern we have for loved ones, and the special concern we have for ourselves. I drew these distinctions earlier in the story of my visit to the dentist's office. There is a distinction between the concern I have for a person I believe is a stranger and the special concern I have for a loved one, and there is a distinction between the special concern I have for a loved one and the special concern I have for myself. I marked this distinction earlier by distinguishing between the "general sympathy for the woe and weal of some arbitrary member of my species" and a "specific concern for one of those members, one with whom I have a special relationship." I further distinguished both of these from the special concern I have for myself. If I expect that I will undergo some suffering, the suffering I expect to undergo is not sympathetic or empathetic, but the thing itself.

If Parfit's reductionism is true, then sympathetic concern I have for the suffering of others is justified but not my special concern for my own suffering. We are justified in the kind of sympathetic concern one has for an R-related descendent of oneself or an R-related descendent of a loved one, but not in the essentially self-regarding concerns we have for ourselves. This seems absurd.

Recall Hume's statement about sympathy: He explains sympathy by using the analogy of strings equally wound up that vibrate sympathetically: when one is plucked, the others vibrate as well. When we have sympathy with someone, we resonate with what they feel. We sympathize with someone, for example, if we feel embarrassed for someone.
For example, suppose a cherished colleague were to preside over some grand event and had left some important article of clothing embarrassingly undone. Imagine the colleague seems not to have noticed but the audience has. It is natural to feel embarrassed for the good professor. It feels just like being embarrassed oneself, but is not confused with it. I know that this feeling that I feel is not in virtue of my having left some article of clothing undone.

In order to accept Parfit’s reductionism, we would need to accept that the sympathetic concern I have for the cherished colleague at the grand event and for my wife at the dentist’s office are justified, but the concern I would have for myself were I in the position of the cherished colleague, the concern I have when the dentist tells me about the pain I must endure, or even the concern I have for myself when the pain finally arrives—all of these—are justified if reductionism is true.

I think that it is just absurd to claim that sympathy is justified but none of the other concerns are. If sympathy is justified at all, it must be, at least in part, because the thing itself is bad to experience. If those I care about experience these things, it is possible for me to enter into their sufferings and indirectly to experience the bad things as well, but the badness of humiliation and the badness of pain have to be among the things that make it right to be concerned. Notice, if I learned that leaving one’s clothing undone in just that way was a symbol of solidarity with some righteous cause—that is, if I learned that this was not a case of humiliation—it would be wrong for me to be embarrassed in the way I was when I thought it an inadvertent faux pas of an increasingly absentminded friend.

If the badness of humiliation and the badness of pain justify sympathetic concern—that is, make it rational or, at least, not contrary to reason—surely the
badness of humiliation and the badness of pain justify the direct concern we have for ourselves. It is a bad thing to experience these things oneself, and it is bad in a very different way from the indirect concern we have for others. It is surely one thing to feel as though one were being humiliated, and quite another to feel humiliated. It is one thing to feel as though one’s own teeth were being drilled, and quite another to have one’s teeth drilled. If sympathetic concern is justified and provides reasons for acting—if it is right to feel vicarious shame or vicarious pain—and if feeling so gives one reasons for acting, surely it is right to feel shame and to fear pain. It is surely right to say that these passions sometimes give us reasons for acting.

That means we should reject and not accept the claim that sympathy is justified and self-love is not. If sympathy is justified, self-love is as well. Since reductionism implies that sympathy is justified and self-love is not, we can also conclude that reductionism is false.

There is a second consideration, very closely related to the consideration of self-love and sympathy. That concerns self-love and benevolence. I claim that Parfit’s Reductionist view of persons implies that neither self-love nor the special benevolence we have for loved ones are justified. I also claim that if the reductionist view does imply this, we ought to reject reductionism. The argument for both of these claims relies on Butler’s position on the “selfish system of morals,” a position I described in chapter 3.

Butler’s position is that benevolence as well as self-love exists, and that they are recognizably distinct from one another. He believes it is false to say that all apparent benevolent actions are really egoistic actions. But more than this, he believes that the ability to draw the distinction is a necessary condition for moral action. If by our nature we are capable only of egoistic actions, then we cannot reasonably be expected to act
benevolently, since we cannot do so. We do, however, find ourselves motivated by benevolent concerns as well as by egoistic concerns. Butler’s position is that we know what benevolent concern is just as we know what egoistic concern or self-love is. I think it is obvious that Butler’s position is right.

I remarked earlier on a different distinction, that between generalized benevolence that one feels toward an arbitrary individual and the special concern one has for a loved one. This is a distinction, as I said before, that is recognized by Whiting and Parfit. The benevolent concern we have for loved ones is special, according to Parfit and Whiting. But, if I am right, the Reductionist must grant that the special concern we have for loved ones is not justified either. If I am right about that, then reductionism implies there is no reason for either self-love or the special benevolence we have towards loved ones.

The reason is, it is as obvious that the special benevolent concern I have for loved ones is not the same as the concern that would be appropriate for a merely R-related descendent of a loved one as it is that the special benevolent concern I have for loved ones is distinct from generalized benevolence. If I were confronted with a merely R-related descendent of a loved one, my concern for that person would not be the special benevolence I have for the loved one. In order to see why, I want to consider some further points of agreement between Parfit and me.

Parfit says a number of things about what we care about in our concern for loved one that I agree with. I mentioned earlier that when Parfit denies that physical continuity is important in a person’s existence he says that what are important are, among other things, memories and commitments.\(^{63}\) I believe that Parfit is right. It matters, for example, that a loved one and I have memories of past experiences

\(^{63}\) Parfit, 284.
together. If one makes a commitment to a person, it matters that one keep one’s commitment to that person. Moreover, I think that Parfit is quite right when he says that mutual love requires a shared history.

In a case in which R-relatedness is preserved but no substance is, one has only shared q-memories of experiences with those one believes to be loved ones, and therefore, one has only the illusion of a shared history with them. And what of commitments with one who is merely R-related? Such commitments are q-commitments, of course, whatever in the end that could mean. For example, if one has a commitment to lifelong monogamy, one might be q-committed to lifelong monogamy with many. It is hard to say how one could keep one’s commitments and one’s q-commitments.

I think that Parfit is right to say that the concern we have for loved ones includes memories, commitments, and a shared history. I also think that these are among the things that distinguish generalized benevolence from the special benevolence we have for loved ones. Parfit’s reductionism, on the other hand, can deliver only q-memories, q-commitments, and the illusion of a shared history. This adds yet another kind of concern. It is not the same as generalized benevolence because, after all, a merely R-related descendent won’t really be a stranger; she will seem to know a lot about me and I about her. And it is not the same as the special benevolent concern we have for loved ones, because a merely R-related descendent is one with whom one shares only q-memories, q-commitments, and the illusion of a shared history. Call the concern one may justifiably have for merely R-related descendent, Mere R Concern.

We can conclude, then, that the special benevolent concern we have for loved ones, which requires memories, commitments and a shared history, is not the same as mere R concern, which requires only q-memories, q-commitments, and the illusion of a
shared history. Parfit's reductionism implies only the latter is justified; so Parfit's reductionism implies the special benevolence we have for loved ones is not justified.

I believe that this provides another argument against Parfit's reductionism. I argued earlier that Parfit's reductionism implies that self-love, or the special concern we have for ourselves, is not justified. Parfit and Perry are accept that. The present argument shows that Parfit's reductionism implies that the special benevolence we have for loved ones is not justified. Parfit's view, then, implies that neither self-love nor the special benevolence we have for loved ones is justified.

If reductionism implies these two concerns, self-concern and special benevolence, are not justified, then, reductionism should be rejected. I already argued that if sympathy is justified, so is the special concern we have for ourselves. But a similar argument can be raised concerning special benevolence and Mere R Concern. If Mere R Concern is justified at all (and I'm not saying it is), it is absurd to hold that the special benevolent concern we have for loved ones is not. The reason is simple: If keeping one's q-commitments, sharing q-memories, and having the illusion of a shared past have value, surely keeping one's commitments, sharing actual memories, and actually sharing a past must have value as well.

We ought, then, to reject Parfit's reductionism. Other conclusions follow from this section.

First, it follows immediately from this that we ought to reject the contemporary CAT, since Parfit's reductionism is the most carefully worked out expression of the contemporary CAT.

Second, the problems that are fatal to Parfit's view seem to be inherent in any version of CAT. Recall that in order to argue that a person's existence consists in consciousness alone, one must do two things initially. First, one must spell out "the
same consciousness” or R-relatedness non-circularly, so that it does presuppose sameness of person, or one has not given an account of the a person’s existence in terms of consciousness. Second, if one wants to claim that a person’s existence consists in consciousness alone one must spell out sameness of consciousness/R-relatedness without appealing to sameness of body, brain, soul, or any other kind of substance.

Parfit proposes to meet these demands with the various q-notions but these q-events are not as good as their ordinary counterparts in what seem to be the most important ways. Q-memories and q-beliefs are sometimes false in cases in which their ordinary counterparts are true; q-desires are sometimes unsatisfiable in cases in which their ordinary counterparts are satisfiable; q-intentions sometimes cannot possibly be carried out when their ordinary counterparts can. Moreover, the problems arise in precisely those cases in which the q-attitudes diverge from their ordinary counterparts.

The conclusion we should draw from this is that a view that a person’s existence consists in consciousness alone simply gets it wrong. Where it gets it right, it does so only accidentally, and in the crucial cases, it simply gets it wrong. Once consciousness is described non-circularly and without reference to the same brain, body, or other substance, it is no longer the only thing we value when we value consciousness, and it is surely not the main thing we value.

Finally, it follows from the falsity of the contemporary CAT that the Unqualified Body-swap Intuition is false as well. If it is not the case that a person’s existence consists in consciousness alone, it is also false that one can exist in a different body if nothing other than consciousness alone is preserved.
5.5 Summary

Let me summarize the argument of this chapter. The contemporary CAT purports to be an account of what diachronic personhood consists in. The claim is that a person’s existence through time consists in consciousness or R-relatedness. I have argued that key premises in Parfit’s argument are undefended, and thus we have nor reason to accept the contemporary CAT. One such important premise is that the body is not an essential part of a person’s existence through time. He makes use of it twice, first, implicitly by choosing opponents who accept it, and second, explicitly when he claims that physical continuity is the least important element in a person’s existence. A second is that R and not identity matters, in any important sense of “what matters.”

I have raised two arguments urging that we reject the contemporary CAT. The first is that the reductionist view of persons implies that sympathy is justified and self-love is not. I argued that this is absurd, which implies that that the reductionist view of persons is false. The second is that the reductionist view implies that neither self-love nor benevolence are justified. I argued that because self-love and benevolence constitute our paradigm for what rational action is, self-love and benevolence are justified if anything is. Therefore, the reductionist view of persons ought to be rejected.

The reductionist view of persons says that a person’s existence consists in R-relatedness alone. It is false, therefore, that a person’s existence consists in R-relatedness alone. It is also false, then, that a person can exist in a different body if nothing but consciousness alone is preserved. That is, it follows from the falsity of the contemporary CAT that the Unqualified Body-swap Intuition is false as well. If it is not the case that a person’s existence consists in consciousness alone, it is also false that one can exist in a different body if nothing other than consciousness alone is preserved.
CHAPTER 6
BRAIN TRANSPLANTS AND THE IMPORTANCE OF THE BODY

I have examined four versions of CAT. Locke’s version says that consciousness alone makes our concept of personal identity. Dodwell’s version says that consciousness alone makes personal identity. The traditional interpretation of Hume says that a person’s existence consists in nothing but consciousness, and therefore there is no personal identity. The contemporary CAT says diachronic personhood consists in R-relatedness alone.

I have argued that Dodwell’s version and the contemporary CAT ought to be rejected. The arguments against these versions give us a more general argument against any version of CAT.

6.1 CAT and the clash with our norms
All the versions of CAT that I have examined clash in some important way with our moral norms. Locke himself calls out attention to a clash or potential clash between moral responsibility and his version of CAT in the “fatal error” passage. Dodwell’s view clashes with our moral norms because he does not seem to recognize the need to distinguish between veridical and non-veridical consciousness. Furthermore, Dodwell allows that twenty distinct beings at a time can constitute a single person, which clashes with various norms. On the traditional interpretation of Hume, our moral norms are contrary to reason because they rely on false beliefs and give us unachievable aims.
Parfit recognizes a number of cases in which his reductionist view of persons is at odds with important intuitions we have about what we have a reason to do.

Each of the versions of CAT I have examined clashes in some important way with our moral norms. It can be shown that the clash between CAT and our norms is not limited to these particular versions of CAT, but inevitable with any version. The clash, I argue below, is traceable to the inability of any version of CAT to meet the twin demands of purging circularity while preserving veridicality. Let me explain.

In chapter 1, I considered Locke's response to the question of what to say about cases that our intuitions tell us are cases of false representations and therefore preserve neither personal identity nor moral responsibility. I said there that Locke's options were either to accept the counter-intuitive implications that consciousness alone, even a false consciousness, makes personal identity and is therefore the bearer of moral responsibility, or to draw a distinction between veridical and non-veridical representations, between true and false memories, and to urge that veridical consciousness and it alone makes personal identity. If this second option is taken, there is an additional requirement that veridical consciousness be spelled out non-circularly.

Dodwell chooses the first option of accepting many counter-intuitive implications and his view is all the more implausible because of it. The criticisms raised against his view help us to see the need for more than simply having the representation of having experienced something, but that the representation importantly gets it right. In short, we see the need for veridical consciousness, not just consciousness.

The contemporary CAT is an unhappy union of the two options that I said were available to Locke. On the one hand, the contemporary CAT introduces the notions of
quasi-memory, quasi-intentions, etc., which, as I argue below, are only accidentally veridical. On the other, Parfit and others are not only willing to recognize the clash between their view and our norms, they argue on that basis that our norms should be revised. I believe the clash arises because once the notion of R-relatedness is spelled out non-circularly, information is left out that is vital to our norms.

I said earlier that a view that holds that a person's existence consists in consciousness alone, independently of sameness of body, brain, or soul—that is, any version of CAT—must be able to spell out sameness of consciousness non-circularly. Quasi-memory and the like are designed to meet this objective. Recall that Parfit characterizes ordinary memory as follows:

I remember having an experience only if

1. I seem to remember having an experience,
2. I did have this experience,
and
3. my apparent memory is causally dependent, in the normal way, on this past experience. (207)

We saw in chapter 4, that conditions (2) and probably (3) are circular with respect to the project of giving an account of diachronic personhood in terms that include memory. Condition (2) really means, "I am identical with the one who had the experience." Condition (3) would likely be circular if "the normal way" were spelled out.

The notion of quasi-memory replaces condition (2) with

2'. Someone did have this experience,

and condition (3) with,

3'. my apparent memory is causally dependent, in the right kind of way, on that past experience. (220).
By purging circularity, which any version of CAT must do, the result is a notion in which veridicality is compromised as well. Seeming to have had an experience, which is what condition (1) tells us, is not sufficient for the experience to be veridical. Conditions (2) and (3) give us the conditions under which seeming to remember having an experience is actually to remember correctly that one had the experience.

In the case Parfit calls "Venetian Memories," Jane has copied in her brain some of Paul's memory traces. She seems to remember various experiences of Venice that she did not experience but Paul did. As I said, even Parfit agrees that Jane's quasi-memories give her mistaken information.

I think it is clear why: In the interest of purging conditions (2) and (3) of circularity, conditions (2') and (3') leave out information that is vital to our norms. If I am remembering having done some shameful act, it is appropriate for me to be ashamed. If I am quasi-remembering, whether it is appropriate for me to be ashamed or not depends on whether I am the one that did the act, the very information that must be left out to avoid circularity and to remain a version of CAT. If I remember having committed some crime, I am responsible for having committed it. If I quasi-remember having committed some crime, whether I am responsible depends on whether I am the one who committed it, the very information that must be left out in order to avoid circularity and remain a version of CAT.

We should conclude no version of CAT is consistent with our norms, since any version must meet these same demands. Moreover, I have argued that the clash arises because CAT must, by its very nature, leave out information that is vital to our norms. What CAT leaves out is whatever makes the difference between someone's experiencing some experience and a particular someone's experiencing some experience, and this information is vital to our norms.
6.2 If R matters, other things matter as well

The arguments that led to the conclusion that CAT and the Unqualified Body-swap Intuition should be rejected allow another general conclusion. Many of the arguments I raised against Parfit proceeded on the assumption that consciousness/R-relatedness matters, has value, or is an important if not essential feature of a person's existence at different times. They concluded that other things matter, have value, or are an important if not an essential feature of a person's existence at different times.

In my argument that R with an unusual cause is not always as good as R with the ordinary cause, I argued that it is not merely that we have mental states that stand in special relations with one another that makes a mental life such a special thing to have, but that these mental states stand in certain relations with the world. I repeat what I said there:

The having of beliefs is indeed a fine thing, but finer still is the having of beliefs that are true. Moreover, we value things—objects, states of affairs—and we can form intentions and act to maximize value. It is a fine thing that we can act with the aim to maximize value. But finer still that we can actually maximize value. A merely R-related descendent has false beliefs about himself: about himself and about his existence through time. Moreover, acting on those beliefs does not achieve essentially self-regarding aims.¹

I used a similar argument to argue that if sympathy is justified, the special concern we have for ourselves is justified as well.

These arguments show that if R matters in a person's existence, if R has rational significance, other things matters as well; other things have rational significance. If R

¹. This work, 107
justifies concern for persons, oneself or loved ones, other features of our existence do as well. If a person’s existence consists in R-relatedness, it consists in other things as well.

I have drawn three further conclusions in this chapter. I argued, first, that no version of CAT is consistent with the norms associated with pride, shame, moral responsibility, rational prudence, benevolence. Second, I have argued that the clash arises because CAT must leave out information that is vital to our norms. I have also drawn a third conclusion from the arguments against CAT in chapter 4. I have argued that if R matters in a person’s existence, other things matter as well.

Parfit would not contest the first conclusion. He believes that reductionism does clash with our norms. If it is true, as I have argued, that the clash is the result of information required by our norms but unavailable to a defender of CAT, then any version of CAT will be too impoverished to accommodate our norms. As the third conclusion says, even assuming that consciousness matters in a person’s existence, it is not all that matters in a person’s existence.

6.3 The importance of the body

In what follows I present a brain-transplant case, rather a standard one, in order to argue that the body, not just the brain, matters in a person’s existence. Barbara Harris’s novel, *Who is Julia*? provides an insightful thought experiment, spelling out details that short philosophical puzzles fail to provide.

The story centers around a tall, beautiful, well-educated blonde named Julia North and a plain-looking simple-minded brunette named Mary Frances Beaudine. Julia is severed in half by a street car as she saves Bobby, Mary Frances’s son. Mary

Frances suffers a stroke at the scene and shortly thereafter is declared brain dead. A team of physicians is in place to do a brain-transplant, which is described as an experimental procedure.

To procure permission from Jack Beaudine, Mary Frances's husband, the hospital administrator tells him,

We have a team of surgeons here who are ready to perform a brain transplant. That is, they are ready to put a healthy brain in your wife's body if you will give us permission. You must understand, however, that this means you will be giving up your wife's body. She's the donor. With the other woman's brain she will no longer be your wife. The recovery period is slow, but when it's over, and if it's successful, the woman whose brain lies in your wife's body will only look like your wife. Her memory, her education, her emotions, and personality will be as they were in her former body. (9).

At first, Jack is indignant and refuses to grant permission. He capitulates, reluctantly, when he learns that the brain-donor and body-recipient is the one who saved Bobby, his son.

Julia's husband, Don North, is told of the operation after the fact. The physician tells Don about the accident saying,

We couldn't restore her body, the body she sacrificed to save the life of a child. That child's mother died of a cerebral hemorrhage the same moment your wife was hurt. We couldn't restore the brain of that young woman. There was only one avenue open to us, and we took it. We removed your wife's brain, and we have reconnected it in the body of that other woman. (41).

He is at first indignant, then incredulous, asking how it is possible for Julia to be alive when her body is dead. He is reassured by the doctor who gives the standard defense of the Body-swap Intuition. He says, "if we are successful as we believe we can be, your wife's mental faculties—her personality—will be intact. She'll be alive for you to talk
with—to go on living with . . . She is alive, Don. Her body is different, but everything else will be the same." (42).

Throughout the story, the husbands are confronted by what is the same and what is different about the Julia-brain/Mary Frances-body. Don looks at the still-unconscious person lying in the hospital bed and thinks, "This homely, ordinary woman was not, could not be Julia" (60). He recalls her lovely cultivated voice and wonders if it will sound the same. When he is able to be near her, "He smelled her skin and her hair, fresh and clean, but strange to his nostrils" (178). Don is struck by the differences between Julia as she was and the Julia/Mary Frances composite.

Jack is struck by the sameness between Mary Frances as he knew her and the Julia/ Mary Frances composite. When Jack sees Mary Frances lying in the bed breathing on her own, he is convinced that she is alive and no argument can persuade him otherwise. When she wakes, Jack becomes even more convinced. After all, she looks, sounds, and smells like Mary Frances. Furthermore, that very body is the one that bore their son.

Both husbands are responding correctly to the facts. For Don, much of what mattered to him about Julia is not preserved between Julia and the Julia/Mary Francis composite; for Jack, much of what mattered to him about Mary Francis is preserved. If the Qualified Body-swap Intuition is true, we should ignore these samenesses and differences as irrelevant to what matters in a person’s existence.

When Jack sees the Julia/Mary Frances composite, much of what he loved about Mary Frances is present. The sound of the voice, even though uttering strange words, makes the same sound as Mary Frances’s; the smile is her smile; this, in an important sense, is the mother of their child. If they had planned to have two children together,
the Julia/Mary Frances composite could bear their second child. The similarities become more striking if we imagine more intimate similarities, more personal characteristics.

From one perspective, much of what matters in a person's existence is missing between Julia and the Julia/Mary Frances composite. From another perspective, much of what matters in a person's existence is preserved between Mary Francis and the Julia/Mary Frances composite.

The conclusion we should draw from this is that both the brain and the body matter in a person's existence. If an R-supporting brain were transplanted in a different body, some of what matters in a person's existence is preserved. If the same living body persists, some of what matters in a person's existence is preserved as well. The reason is, our interest in persons is complex: it is not merely interest in a simple thing such as a soul nor even in a single organ such as the brain. It includes, but is not limited to, interest in a particular body with a particular origin and a particular history.

Our interest in persons is as complex as the moral and prudential norms that govern them. Some of these norms are backward-looking, historical. Robert Nozick, for example, spends a great deal of time arguing that justice is historical, depending of what was actually done. Whether it is just to punish Hauptmann depends on whether he did it. Other norms are historical as well. Pride and shame look to the past as does reward and compensation. These norms presuppose the existence of persisting persons, of entities with a history. Even if my strange experiences are the result of a brain-transplant, I have mistaken beliefs about my past and unachievable aims. Likewise for the Julia/Mary Frances composite.

Many of our norms are historical in this way and so, according to Nozick, is love. Love is historical, according to Nozick, in that it depends upon what actually occurred. An adult may come to love another because of the other's characteristics; but it is the other person, and not the characteristics, that is loved. The love is not transferable to someone else with the same characteristics, even to one who "scores" higher for these characteristics. And the love endures through changes in the characteristics that gave rise to it.\(^4\)

Parfit argues that love is historical. He says, "loving someone is a process, not a fixed state. Mutual love involves a shared history" (295). If it is true as I have argued that love is historical, the body cannot be treated as an unimportant contingent feature of our existence.

Some of these norms are forward-looking: Taking care of myself—the prime directive of prudence—requires that I take care of my body, that I recognize its potential for growth and decay, and that I treat it appropriately for things of this kind. Beneficence, which most moral theories recognize as a duty, brings the same demand with respect to other persons as prudence brings with respect to oneself. Promoting the greater good requires, in part, knowing what constitutes a good for things of this kind.

What this means is, if R matters because of the access it gives to the past through memory and the access it gives to the future through intentions, then the body matters as well. The reason is that the body is part of the content of many essentially self-regarding memories and intentions, and is therefore part of the content of our norms.

We don't even need science fiction to argue for the importance of the body. The familiar tragedies that befall persons every day drive the same point home. The

---

ordinary process of decay that human beings undergo is familiar and tragic. We suffer decaying cartilage, thinning or graying hair, loss of stamina, loss of strength. Initially we suffer benign instances of forgetfulness and later, typically, more severe instances follow. Not every human being suffers these losses toward the end though the end is always the same. Accident or disease may cause such losses more abruptly. It remains true that persons can suffer the loss of limbs and of organs and still survive as the same person. They may also suffer the loss of memories, desires, virtue, or other psychological characteristics.

Consider the significant psychological change that Alzheimer's disease brings. There is no question that Ronald survives as the same person when at first his memory begins to slip. The occasionally faltering Ronald is surely the same person as Ronald in his prime. A decline may continue over a period of years before there is even a question of whether Ronald still exists. When there is finally a question whether he is the same person, much of what we value about personhood is gone. And even when Ronald's last lucid moment has passed, one's obligations to him do not cease.

So, while it is true that a person can be said to exist in a different body, it is also true that one can exist in as a semi-coherent being in the same body. Either existence preserves some of what matters; neither preserves all of what matters.  

Worthwhile disputes continue over which existence is better. But these are disputes about values rather than disputes about facts. If this is true, one more conclusion follows. This is a conclusion about problems associated with what is often called the metaphysics of personal identity.

The various brain-transplant cases allow us to consider features of our existence in isolation that normally go together. By imagining the Julia case, we are agreeing about the facts relating to sameness of brain, sameness of body, and R-relatedness. The dispute concerns which of those matters in a person's existence. Because each of them matters in its own way, the really interesting questions that arise concern what matters, what we do value, what we would value upon careful and sober reflection in a cool hour, or what we ought to value.

My final conclusion, then, is that the metaphysical problem of personal identity is at bottom a problem of value. If we are inclined to think of questions of value as Hume did, as felt rather than discovered, the solution to the various problems of personal identity will rely on claims about what we care about. If we are inclined to think of questions of value as objective facts discovered and not felt, the solution to the various problems will rely on claims about what we ought to care about, what has objective value.

I have claimed that the same body matters in a person's existence through time. By this I mean that after some years of careful and often sober reflection, sometimes in cool hours, sometimes in a frenzy, the body seems to me to have significance in a person's existence through time. I have granted the facts of the case, facts about brains, bodies, and R-relatedness. And I believe that the body is a valuable feature of a person's existence because it is presupposed by many person-regarding norms.
Having granted the facts, the remaining puzzles about our nature and our existence through time are questions of value, what we do value and what we ought to value. I might be in error in valuing some of the things I value. Even if this is the case, I need not fear a challenge to my values on the basis of false beliefs I have about the nature of a person.
BIBLIOGRAPHY


