3

Exceptional Persons: On the Limits of Imaginary Cases

When should we trust our judgments about far-fetched imaginary cases? In this chapter, I offer one possible answer. I propose that if the imaginary scenario is adduced to illustrate a concept structured around a set of necessary and sufficient conditions, and if these conditions play a role in how we identify candidates as falling under that concept, then our judgments about the far-fetched imaginary case may help us to separate essential features of the concept from accidental ones. But if the concept is not structured in that way, or if the features in question do not govern our application of the concept, then our judgments about imaginary non-realistic cases are likely to be misleading.

I go on to argue that the concept of personal identity falls into the second of these classes, and hence that far-fetched thought experiments may not illuminate the concept in the way that they have been purported to. The chapter includes detailed discussions of John Locke's Prince and Cobbler case, Derek Parfit's teletransportation case, and Bernard Williams's A-body/B-body case. As in the Galileo chapter (Ch. 1), these examples are illustrative, and the general claim is intended to hold more widely.

This essay and its successor—"Personal Identity and Thought Experiments" (Ch. 4)—form a natural pair. For those who prefer to read only one of them, it might be useful to know that the later paper is more intricate and subtle in its argumentation, and can fruitfully be read as a self-standing piece.

For those already familiar with the personal identity literature, the key ideas in this essay can be acquired by focusing on §§1 and 4.

***

This essay was first published in the Journal of Consciousness Studies, 5/5–6 (1998), 592–610 and is reprinted with the kind permission of Imprint Academic. The text that appears here is largely unchanged, though several now-outdated bibliographical footnotes have been omitted, and minor stylistic alterations have been made to several remaining footnotes and to brief passages in the main text. As a result of reformatting, some footnote numbers have changed. The original number of each footnote is noted in [square brackets].

53
Part I. Thought Experiments

It is of great use to the sailor to know the length of his line, though he cannot
with it fathom all the depths of the ocean. It is well he knows that it is long
enough to reach the bottom at such places as are necessary to direct his
voyage, and caution him against running upon shoals that may ruin him.

—John Locke, Essay Concerning Human Understanding, I. I. 6

1. Introduction

1.1. The Problem of Personal Identity

The question of (diachronic) personal identity (at least as it concerns many
contemporary Anglo-American materialist philosophers) is the question of
determining the necessary and sufficient conditions for a person at some later
time to be identical with a person at some earlier time. Phrased in the debate's
standard language, the question is: under what conditions are we correct in
saying that P2 (a person who exists at t2) is the same person as P1 (a person who
exists at t1)? The question is a special case of the more general question of the
identity conditions for entities over time: under what conditions are we correct
in saying that E2 (an entity of sort E which exists at t2) is the same E as E1 (an
entity of sort E which exists at t1)?

Those who think that the question of diachronic personal identity has a
determinate categorical answer tend to respond in one of two ways: either it
is suggested that some sort of physical characteristic—such as having the same
body—serves as the basis for identity over time, or it is suggested that some sort
of psychological or mental characteristic—such as having the same memories—
serves that role. It should not be surprising that these are the two sorts of
answers that have been offered. We are, after all, physical beings whose most
notable feature is our psychological characteristics; what is essential to who we
are is presumably either the distinctive set of beliefs, desires, memories, etc. that
together constitute our character, or the distinctive configuration of molecules
that together constitute our body, or, perhaps, some combination of the two.

For roughly half a century, the philosophical literature on personal
identity has centered on arguments of a certain type. These arguments defend
revisionary conclusions about the nature or importance of personal identity on
the basis of an assumed convergence of responses to purely imaginary cases.

1 [3] In addition, there are those who suggest that both factors are necessary, for instance,
that what is required for P1 and P2 to be the same person is that P2 have more than 50% of P1’s
brain, along with P1’s core psychology (Unger 1990).

2 [4, slightly modified] Whether they do, in fact, tap relevant intuitions has been a topic
discussion since their inception: cf. Quine 1972; Wilkes 1988. Or, in somewhat more
colorful colloquial form, consider the following quote, taken from a lay-directed book
entitled Persons: What Philosophers Say about You: “Philosophers frequently dream up weird
examples to test their understanding of concepts. With respect to persons, they talk about
So, for instance, one is asked to contemplate a case in which A’s brain is transplanted into B’s body, or a case in which some of C’s memories are implanted in D’s brain, or a case in which information about the arrangement of the molecules that compose E is used to create an exact replica of E at another point in spacetime.3

Thinking about these cases is supposed to help us tease apart the relative roles played by features that coincide in all (or almost all) actual cases, but which seem to be conceptually distinguishable. So, for instance, even though we can ordinarily assume that the beliefs, desires, memories, etc. which are associated with a given body will not come to be associated with another body, it does not seem to be in principle impossible that such a state of affairs should come about. Indeed, it seems that we can describe a mechanism by which such a situation might arise: for instance, A’s brain (and with it A’s beliefs, desires, and memories) might be transplanted into B’s body. And since the scenario described strikes us as something of which we can make sense, it seems we can make judgments of fact or value about which of the two factors really matters in making A who she is. We might ask, for instance, whether it would be true to say that A had survived in a body that used to belong to B, or whether it would be right to punish the B-bodied human being for A’s actions, or whether if we were A before the intended operation, we ought to worry about what would be happening to the B-bodied person afterwards. And on the basis of these judgments about what we would say in the imaginary case, we can return to the actual case having learned something about which features are essential and which accidental to our judgments concerning the nature or value of personal identity.

My goal in this chapter is to suggest reasons for thinking that this methodology may be less reliable than its proponents take it to be, for interesting and systematic reasons.

1.2. Questioning the Methodology

In general, two sorts of objections are offered when appeals to such scenarios are made.4 The first type involves substantive disputes about particular cases:

machines that can duplicate human behavior and appearance, brain transplants, mind interchanges, teletransportation… This tends to alienate people in other fields who wonder what philosophers have been smoking” (Bourgeous 1995: 19).

3 [5] In speaking of “A’s brain” or etc., “B’s body,” etc., I am speaking loosely, since precisely what is at issue is what sort of thing “A” or “B” might be.

4 [6] See especially Wilkes 1988. Cf. also: Johnston 1987, 1989, and 1992; Quine 1972; Johnson 1988; Snowdon 1991. For more general discussions of the methodology, see Sorensen 1992a, and the papers collected in Horowitz and Massey (2009). In the dozen years since this paper was originally published, a steadily increasing number of articles have appeared that discuss philosophical methodology in general, and thought experiments in particular. For a regularly updated list of such material, see <http://philpapers.org>.
Part I. Thought Experiments

whether, for instance, divided consciousness can be imagined "from the inside," or whether brain transplants are biologically possible. The second type concerns a more general issue: whether our concepts should (or could) support all of the implications of our beliefs concerning what is practically, physically, or conceptually possible. Those who make such objections have been partially heeded; appeals to thought experiments in discussions of personal identity are now often prefaced by a discussion of the feasibility of the scenario described, and sometimes by a discussion of the legitimacy of the methodology as such. But with a few exceptions, I think both critics and defenders have located the problem in the wrong place.

Certainly, pace extreme critics, there is nothing wrong with the methodology of thought experiment as such. After all, a thought experiment is just a process of reasoning carried out within the context of a well-articulated imaginary scenario in order to answer a specific question about a non-imaginary situation. Such hypothetical test cases play perfectly unobjectionable roles in legal reasoning, linguistic theorizing, scientific inquiry, and ordinary conversation. And there is nothing categorically wrong with thought experiments that concern technologically or biologically or even physically impossible situations. Again, such purely counterfactual test cases play unobjectionable roles in each of the domains just listed. Rather, I will argue, the legitimacy or illegitimacy of thought-experimental reasoning in a particular case depends upon the structure of the concept that the thought experiment is intended to illuminate. If the concept is structured around a set of necessary and sufficient conditions, and if these conditions play a role in how it is that we identify candidates as falling under that concept, then imaginary cases may help us to separate essential features of the concept from accidental ones. But if the concept is not structured in that way, or if the features in question do not govern our application of the concept, then imaginary cases are likely to be misleading.

---


6 [8] In particular, the argument I present below owes a tremendous amount to the work of Mark Johnston.

7 [9] In the psychological literature on concepts, this is generally referred to as the "classical view," and is widely regarded by psychologists as inadequate in accounting for all but a few of our concepts. (For surveys of this literature, see Komatsu 1992; Medin and Smith 1984; and Smith and Medin 1981.) [Added in 2009: Two valuable more recent volumes are Margolis and Laurence 1999 and Murphy 2002.] I think this dismissal is overhasty, in part for the reasons discussed in Rey 1983 and Rey 1985. But since I agree that the view cannot account for the relevant concept in the case on which I am focusing, I will not pursue these general issues further.

8 [10] For incisive discussion of these issues, see Johnston 1987. For reasons similar to my own, Johnston contends that these conditions are not met when we think about imaginary cases involving the concept "person."

9 [11] I suspect it is this distinction that explains the widespread sense that thought experiments are less problematic in science than in philosophy. It is certainly true that scientific concepts are, in general, more likely than non-scientific concepts to be structured around necessary and sufficient conditions that play a role in how we identify instances that
My main contention in this chapter is that the concept of personal identity belongs to the second of these classes. Although philosophers from Locke on are correct in recognizing that a conceptual distinction can be drawn between what Locke called the “man” (or human animal) and what he called the “person” (or set of distinct psychological characteristics), it does not follow that we are able to make informative judgments about many of the combinatoric arrangements in which these features might appear. For the fact that two features can be conceptually separated in the sense that we can imagine the one obtaining without the other in some particular case does not mean that those two features are conceptually distinct in the sense that we can make reliable judgments about them, considered in isolation. Conceptual separability guarantees conceptual distinctness only if our knowledge of the necessary and sufficient conditions of a concept is what governs our application of that concept in other cases, we have no such guarantee.

So even if we are aware that the two features need not coincide in all possible cases, the fact that they coincide in all (or even in nearly all) actual cases may mean that there is no ascertainable fact of the matter about how we would or should respond to either in isolation. For while we may be able to make sense of exceptional situations where they come apart, we do so only by relating them back to ordinary cases where they coincide. But this means that our evaluation of the exceptional case will depend upon which mapping we use in making this assimilation. And this means that our ability to make sense of such cases outruns our ability to make reliable judgments about them.

I think this is what explains both the appeal and the inconclusiveness of thought experiments about personal identity.10 As a matter of evolutionary fact, human persons are entities that are both biological organisms and self-conscious loci of psychological characteristics. And while these features are clearly conceptually separable—we can easily make sense of cases in which the purely biological kind “human animal” and that kind’s most striking characteristic, “self-conscious locus of psychological characteristics,” come apart—it is not clear that we can make informative judgments about them. For although the world forces us to think about certain exceptional cases in which a single body may be able to support more than one collection of psychological attributes (cases, for instance, of multiple personality, or of other sorts of dissociation and compartmentalization, or of memory loss and subsequent relearning), it does not present us (at least according to those whose views are the targets of my discussion) with cases in which a single set of psychological characteristics may be present in more than one body, either diachronically or

---

10 [Note added in 2009: For further exploration of the role of thought experiments in science, see Chs. 1 and 2 above.]  
11 [12] For a related but somewhat different take on this question, see Rovane 1994, as well as Rovane 1998, esp. ch. 2.
synchronically.\textsuperscript{11} And this contingent fact—that is, the fact that in almost all cases, a single mind is associated with a single body—plays a central role in how it is that we make judgments about the nature and importance of personal identity.\textsuperscript{12} For it is the lens through which we view the cases where the one–one coincidence does not hold.\textsuperscript{13}

Below, I will describe a widely discussed imaginary case to which our responses seem to be frame-dependent: how the story is told affects how the story is evaluated. This phenomenon is easily explained if we assume that the way we make sense of such cases is by assimilating them to a class of cases with which we are familiar. Whichever of the story’s features are made salient by the particular presentation will thus serve as the basis for assimilation to the ordinary cases in which the contingently associated features that together comprise ordinary cases of personhood coincide. On a biological or psychological view of personal identity, however, one or another of the responses could legitimately be criticized as mistaken. But for reasons that I will discuss after I present the case, I think both responses are at least rationally permissible.

1.3. Plan for the Rest of the Chapter

In the remainder of this chapter, I will proceed as follows. In §2, I will briefly summarize Locke’s views on the identity conditions governing objects, organisms, and persons, and describe the famous thought experiment from which modern discussions of personal identity take their inspiration. The purpose of this section is twofold: first, to provide an example of the sort of analysis that is offered in discussions of the identity conditions for entities over time; and second, to do so in the context of the Lockean framework, which has served as a jumping-off point for subsequent discussions.\textsuperscript{14} Readers familiar with Locke’s writings might skip this section without losing the thread of the

\textsuperscript{11} [13] In saying this, I do not mean to deny that there are purported cases of reincarnation, memories of past lives, soul-transfer, etc., and that commitment to the existence of such phenomena may play a central role in the conceptual schemes of millions, even billions, of people. But as far as I can tell, these purported facts play no part in the conceptual schemes of those who make use of the sorts of imaginary cases that I will discuss below. (I thank an anonymous reviewer for the Journal of Consciousness Studies for the reminding me of the provincialism of ignoring such views.)

\textsuperscript{12} [14] To repeat: this is so even if there are a small number of actual exceptions.

\textsuperscript{13} [Note added in 2009: I explore this theme in greater detail in Ch. 4 below.]

\textsuperscript{14} [15] Cf. Harold Noonan’s remark that just as “[i]t has been said that all subsequent philosophy consists merely of footnotes to Plato,” of personal identity “it can truly be said that all subsequent writing has consisted merely of footnotes to Locke” (Noonan 1989: 30). Whether this is an accurate description depends, of course, upon exactly how much new material can be included in (metaphoric) footnotes. But it certainly provides a rough indication of Locke’s centrality, and some justification for my beginning my discussion only with Locke (as opposed to with Plato or Aristotle), and for my beginning my discussion already with Locke (as opposed to with Bernard Williams or Derek Parfit).

* Which would make them, I suppose, sub-footnotes to Plato.
argument. In §3, I will identify what I take to be the crucial aspects of Locke's thought experiment, and explain how contemporary cases can be seen as continuous with it. The purpose of this section is to identify some central assumptions that underlie contemporary discussions of the nature and importance of personal identity, and to amass intuitive support for my claim that we can make sense of cases about which we cannot make judgments. Again, the details of this section are not essential to my main argument, so readers familiar with contemporary Anglo-American literature on the nature and importance of personal identity may wish to read these pages rather quickly. The principal argument is picked up again in §3.3, where I consider a particular case in some detail (that described by Bernard Williams in "The Self and the Future"), and suggest a way of understanding the case in light of a more general strategy for understanding how we make sense of exceptions.

2. The Lockean Background

2.1. Locke on the Identity of "Men" and of "Persons"

Modern discussions of the metaphysics of personal identity can be traced to Locke's chapter Of Identity and Diversity in his Essay Concerning Human Understanding. There Locke defends a view that identity is "suited to the idea" (§7); that is, that criteria for identity (over time) are criteria for identity as an X (over time), and that these criteria, in turn, can be categorized into several clusters, each involving different kinds of identity criteria. With this in mind, Locke sets out to provide a general taxonomy of types of identity over time for various sorts of bodily substances: first non-living, then living.

The primitive bearers of bodily identity on the Lockean picture are atoms, which are the basic units of matter. So long as an atom exists, there can be no question about its identity; being fundamentally simple, the atom exists unchanged as long as it exists at all (§2). At the next level of complexity come

---

15 [16] My discussion in that section is not intended to do justice either to the details of Locke's view, or to the many issues of identity over time that thorough discussion of the view would require. For discussion of the former, the reader might fruitfully consult Alston and Bennett 1988, Mackie 1976 (especially Chs. 3, 5, and 6), or (for a general overview) Noonan 1989, Ch. 2; a bibliographic exploration of the extensive literature on the latter question might reasonably begin with the sources listed in Noonan 1989: 255–9. [Added in 2009: Again, the reader who wishes to explore more recent work on the topic might consult the extensive listings at <http://philpapers.org/>.]

16 Locke 1710/1975, Bk. II, Ch. xxvii. Section numbers in the main text, indicated by the convention "(§x)", refer to sections of this chapter in Locke.

17 [18] I here neglect Locke's discussion of the other sorts of substances—God and finite intelligences—about which identity judgments might be made. For reasons that Locke explains clearly in §2, these do not raise the sorts of puzzles that are raised when we make identity judgments about complex bodies.
Part I. Thought Experiments

simple clumps of matter, where “two or more Atoms [are] joined together into the same Mass”; here continued existence requires the continued contiguity of the body’s subparts, whatever the arrangement (§3).

A very different sort of criterion governs Locke’s remaining categories—plants, animals, and human beings; for entities of these sorts, identity is a function not of sameness of matter, but of sameness of life. A plant, such as an oak tree, is the same plant so long as it follows a natural course of events determined by its organic unity as a living entity of a particular sort, regardless of radical changes in its form (say, from acorn to sapling) or the particular matter that makes it up (§4). Similar criteria govern animals, whose identity comes from the internally driven unified participation of the various parts in a continuing life (§5). Finally, Locke turns to the criterion of identity for man, which is nothing more than a special case of animal identity (§6).18

Thus far, Locke has introduced two sorts of identity criteria: non-living (non-artifactual) bodily substances retain their identity through identity of matter, and living bodily substances retain their identity through identity of life. This latter criterion is intended to satisfy two constraints simultaneously: first, to permit continued attribution of identity in the face of change (§6), and second, to do so without eliminating the ground for drawing distinctions between separate individuals (§7). That is, in providing a theory of identity for human beings over time, a distinction must be drawn between the sorts of changes which are entity-preserving (such as growth, or getting drunk, or the loss of a limb), and the sorts of changes which are entity-destroying (such as transformation into a beast, or death). The criterion of “participation in a single continued Life,” which is nothing more or less than the criterion for animal identity, is meant to capture precisely this distinction.19

The main negative conclusion of the first part of Locke’s discussion is that none of the criteria of identity for bodily substances over time can do justice to what Locke will call personal identity. So the discussion thus far can be seen as prefatory to Locke’s famous distinction between “man” and “person,” and the corresponding distinction between the identity suited to the idea of the one, and the identity suited to the idea of the other (§7). Whereas man is an animal, that is, a “living organized Body” of “a certain Form” (§8), a person is “a thinking intelligent Being, that has reason and reflection, and can consider itself as a self, the same thinking thing in different times and places” (§9). That is, a person is a self-conscious reflective being whose awareness of itself as a thinking thing over

18 [17] Throughout this section, for convenience of exposition, I follow Locke’s terminology in using “man” as a general (gender-neutral) term for human animals, and “person” as a general term for human beings considered as psychological entities. In later sections, I make use of the terms “Lockean man” and “Lockean persons” respectively.

19 [19] “The Identity of the same Man consists . . . in nothing but a participation in the same continued Life, by constantly fleeting Particles of Matter, in succession vitally united to the same organized Body” (§6).
time is what serves to make it the same self. The criterion of identity for personhood is “sameness of... rational Being,” and the personal analogue to “participation in a single continued Life” is participation in a single consciousness; life unifies men, consciousness unifies persons.

2.2. The Prince and the Cobbler

With this distinction in place, Locke goes on to discuss a number of imaginary cases designed to buttress his analysis of personal identity, the most famous of which is the story of the prince and the cobbler. He writes: “Should the soul of a prince, carrying with it the consciousness of the prince’s past life, enter and inform the body of a cobbler, as soon deserted by his own soul, everyone sees that he would be the same person with the prince, accountable only for the prince’s actions: but who would say it was the same man?” (§15).

Locke’s story is motivated by a desire to avoid putting excessive weight on a certain uniformity in the world—that persons and men coincide with sufficient regularity that the distinction between them seems to have gone unnoticed. In order to establish that the two are conceptually separable, he employs a technique well known from scientific methodology: that two features can be shown to be discrete if it is possible for each to obtain without the other. If the story Locke has described makes sense to us, then the conceptual distinction he wishes to draw must also make sense. But this alone does not show that we can make reliable judgments about Lockean persons as self-standing entities. So let us look a bit more closely about why Locke’s story seems to make sense, and why that does not show what Locke and his followers have taken it to show.

2.3. Crucial Elements of the Lockean Story

Locke’s scenario has the following crucial elements:

(1) The set of psychological characteristics (“personality”) previously associated with one body comes to be associated with another body, in such a way that it seems to the rest of the world that the Y-body manifests the personality previously associated with (the) X-body; and it feels to the Y-body person (that is: the X-soul person) as if she has gotten a new body.

---

20 [20] “For since consciousness always accompanies thinking... in this alone consists personal identity, i.e. the sameness of a rational being; and as far as this consciousness can be extended backwards to any past action or thought, so far reaches the identity of that person” (69).

21 [21] Locke’s less famous cases—including those of the Christian Platonist, the Mayor of Queensborough, the Day-man and the Night-man, and so on—are described in the surrounding sections.
Part I. Thought Experiments

(2) The story describes a mechanism—the movement of the consciousness-carrying soul from one body to the other—by which the changes described in (1) come about.

(3) The mechanism is such that the X-personality is manifest in the Y-body because some substance that was present in the X-body is now present in the Y-body.

That is, the Lockean story depends upon the following three things: (1) that we can make sense of a story in which two personalities “switch bodies”; (2) that we can describe a mechanism by which such a switch might take place; (3) that that mechanism involves some transfer of some sort of (material or immaterial) substance. What (2) and (3) help us to see is that we do think the scenario described in (1) is coherent; by providing us with a narrative about how the surprising state of affairs described in (1) might come about, they help make the situation seem less mysterious. But from the fact that we can tell a story about how the world might come to be configured in some way other than the way it actually is, it does not follow that we will be able to make judgments about the various combinations of the features Locke has isolated. As the combinations grow more complex, we see that the initial illusion of certainty about the simple case was only that: an illusion.

So in the next section, I will trace the ways in which a number of widely discussed contemporary thought experiments can be seen as arising out of Locke’s original case. My purpose in doing so is twofold: to demonstrate the continuity of contemporary cases with their early modern predecessors, and to show how the increasing complexity which these cases introduce in no way disrupts our ability to make sense of them, but wreaks havoc with our ability to make informative judgments about them.

3. Variations on the Lockean Story

3.1. Transferring Matter

In direct descendants of Locke’s story, only (2) is altered. Instead of the “soul ... carrying with it the consciousness of the prince’s past life,” it is the brain of the first character (carrying with it the consciousness of the character’s past life) that is imagined to have been transplanted into the body of the second character. As with Locke’s story, the intuition this is standardly taken to evoke is that the person has moved from one body to another. A typical presentation of this sort of story is the following:

Imagine ... that in the twenty-first century it is possible to transplant brains, as it is now possible to transplant hearts, and let us suppose that the brain of a Mr. Brown is transplanted into the skull of a Mr. Robinson ... The result of the operation, call him Brownson,
will then be a completely healthy person... with Robinson’s body, but in character, memories and personality quite indistinguishable from Brown... Most modern philosophers who have reflected on this case... have found that they could not honestly deny that Brownson, in the case imagined, was Brown. (Noonan 1989: 4-5)²²

Let us call this the “brain transplant case.” A minor variation on this story is one in which the brains of the two characters are switched, so that, for instance, just as Brown’s brain is transplanted into Robinson’s skull, Robinson’s is transplanted into Brown’s. (See e.g. Perry 1975a: 3-6.) On the standard interpretation, just as we would be inclined to say that Brownson is Brown, so too would we be inclined to say that Robin is Robinson. Let us call this the brain switch case.

The brain has a certain complexity that has led philosophers to consider a variation on the brain transplant case referred to as fission. Following the appearance in the philosophical literature of an article entitled “Brain Bisection and the Unity of Consciousness” (Nagel 1971), in which the results of Sperry’s split-brain research were presented, philosophers began to consider the possibility that a single brain might in principle be able to support two loci of consciousness.²³ And if this is so, they reasoned, then it seems to be a shallow rather than a deep truth that various processes are localized in one or another brain hemisphere, rather than being spread throughout the brain as a whole.

But from here it seems only a minor idealization to imagine a case in which all of the features of Brownson’s personality are realized in his brain twice over, such that the transplant of either half to Robinson’s body would be sufficient to give us a human being with all of Brownson’s memories, beliefs, desires, etc. But if either half would be sufficient, then we might also coherently imagine a third case, in which after being removed from Brownson’s body and divided into two, one half of Brownson’s brain is transplanted into Robinson’s body, and the other half is transplanted into a third body, that of Robinson II. But then, is Brownson, or Brownson II, or both, or neither? It seems clear that he cannot be both, for surely Brownson and Brownson II are distinct individuals, and if Brown were identical with both, then they would have to be identical with each other. But it seems equally problematic to say that he is one as opposed to the other; for what could it make the case that he is Brownson instead of Brownson II? And it seems no more plausible to say that he is neither one; after all, were it not for

²² [23] To the best of my knowledge, this version of the story derives from Shoemaker 1963.  
²³ [24] Discussions of similar cases, in which persons were simply hypothetically “split like amoebas,” were present in the literature before mid-century. But it was only when philosophers were able to identify a mechanism by which such a process might “actually” occur that discussion began to take on the proportions it currently has. (To be fair, there are additional factors, both micro- and macro-sociological, which may also explain the swelling literature.) [Added in 2009: Apparently, a similar debate occurred in the eighteenth century; for details, see Martin and Barrett 1999.]
Part I. Thought Experiments

the existence of Brownson II, he would surely be identical to Brownson, and it seems odd to say that twice-over survival is tantamount to death.

3.2. Reconfiguring form

The case of fission is indeed deeply perplexing, and to say more about it here would distract us from my main line of argumentation even more than we have been already. Instead, I turn to cases in which (3) is denied, that is, cases in which it is imagined that it is not the transfer of some substance, but rather the transfer of some information by which the circumstance described in (1) comes about. Consider three such cases which we can call the brain state transfer case, the brain state exchange case, and the brain state duplication case. For the first (brain state transfer), we might imagine a machine that scans all of the information about the configuration of the pre-switch X's brain, and then uses this information to reconfigure the brain of pre-switch Y so that the brain in Y's body comes to support all of the memories, beliefs, and desires of the pre-switch X.25 The outcome of such a process is functionally identical to the outcome of the brain transfer described above; the Y body manifests the memories, beliefs, and desires of the pre-switch X, and does so because pre-switch X did. And, of course, as with the brain transfer case, we might imagine slight variations: a more complicated version of this story wherein the brain structures of the two characters are switched (brain state exchange), or a case in which the duplication (or multiplication) occurs without disrupting the original (brain state duplication).

But if the correct set of instructions is enough to allow us to reconfigure one brain in such a way that it supports the psychological characteristics previously associated with another brain, then with a slightly more complicated set of instructions, we ought to be able, at least in principle, to assemble a brain, or indeed a whole human being, out of simple bits of matter of the same sort of which the original was composed. So we might imagine a case of "teletransportation," where one or more exact duplicates of a human being would be brought into existence at a spatially remote location as a result of information garnered about the structure of the original being.

And if this is in principle possible, then it also seems possible, in principle at least, that matter might come to be configured in that way without this configuration being the result of information-transfer as described above. That is, it seems that what might be called "independent replication" is at least in principle possible, wherein one or more exact duplicates of a human being would be brought into existence at a spatially remote location not as a result of


information garnered about the structure of the original being, but as a result of
some causally independent process.

Let me summarize the variations I have described on the basis of the assump-
tions that underlie them:

(4) Assuming that the source of psychological content is localized, this could
be removed and transplanted to another body (brain transfer, brain
exchange).

(5) Assuming that the source of psychological content is localized, re-
dundantly realized, and divisible such that either half might ade-
quately serve as the basis for full psychology, it could be divided and
doubly transplanted (fission).

(6) Assuming that the source of psychological content might be realized in
another physical entity of similar basic structure, the matter of that entity
might be reconfigured in a way that would render it structurally identical
to the original (brain state transfer, brain state exchange, brain state
duplication).26

(7) Assuming that the same arrangement of the same kind of matter
would produce the same macrostructural properties, one or more exact
duplicates could, in principle, be generated at a spatially remote location
(telereportation).

(8) Assuming that the same arrangement of the same kind of matter would
produce the same macrostructural properties, one or more exact duplicates
could, in principle, be spontaneously generated at a spatially remote
location (independent replication).

It seems to me undeniable that we can make sense of such scenarios. Each rep-
resents a state of affairs that seems metaphysically, perhaps even physically, possible.
But this does not mean that we can make reliable judgments about how we would
or should respond to such scenarios, were we to encounter them. In the next
section, I will say more about this contention in the context of one particular case.

3.3. The Self and the Future

In "The Self and the Future," (Williams 1970/1973b) Bernard Williams de-
scribes an imaginary case in which one is asked to contemplate a machine of
a sort rendered practicable by (6) above: when two individuals, A and B, are
hooked up to the machine, it reconfigures the A-brain in such a way that it
comes to be associated with all of the psychological states previously associated

26 [Note added in 2009: For novel-length exploration of this idea, see Robert J. Sawyer's
Mindscan. (Thanks to Susan Schneider for directing me to this book, which, in keeping with
its themes, I read partly in paper form and partly on my Kindle.)]
Part I. Thought Experiments

with the B-brain/B-body person, and reconfigures the B-brain in such a way that it comes to be associated with all of the psychological states previously associated with the A-brain/A-body person. With this in place, Williams asks the reader to consider the following two stories.

In the first, one imagines A faced with the prospect of being connected up to the machine in question. From the machine will emerge two persons: the first, the A-body person, will have the body previously associated with A, but all of the psychological states previously associated with B; the second, the B-body person, will have the body previously associated with B, but all of the psychological states previously associated with A. Before the operation, A is told that one of the two resultant persons will be given a large financial reward whereas the other will be tortured. A is asked to decide, on purely self-interested grounds, whether the reward should go to the A-body person or the B-body person.

As the case is presented, it seems sensible for A to direct the reward towards the B-body person. Among the other evidence that seems to support this decision as being correct is the fact that, when the operation is over and the goods are distributed, the B-body person—whose memories and desires correspond to those of the pre-operational A—will say rightly: “This is just the outcome I selected! And how glad I am that I so chose.” Whereas, presumably, if the reward went to the A-body person, the B-body person would remark with outrage: “Why am I sitting here in great physical discomfort, when what I requested was the reward?” (Cf. Williams 1970/1973b: 48-50.) The intuition evoked by this first version is that we seem to be able to make sense of there being some sort of procedure whereby two persons could, so to speak, “swap bodies.” To the extent that I bear to my future self a relation of rational prudential concern, I might properly bear that relation to someone with whom I shared no physical matter at all.

The second scenario is the following. One imagines A to be in the hands of a particularly dastardly surgeon, who tells A: “Tomorrow, you will be subjected to great physical discomfort. But before this happens, you will be operated upon with the following effect. ‘[Y]ou will not remember being told that this is going

27 [27] As one of the central cases in the personal identity literature, Williams's story has received wide discussion. In light of this extensive literature, my rather flat-footed presentation of the story may strike readers familiar with these discussions as rather naive. But my purpose in presenting this case is to talk about what role imaginary cases in themselves can play in deciding the sorts of questions they are credited with deciding. I have no doubt that imaginary cases considered in conjunction with well-worked-out philosophical theories can do many, many things.

28 [28] If, however, as Derek Parfit has argued, “identity is not what matters” for rational prudential concern, then the case does not allow us to conclude anything about the identity of the person towards whom the prudential concern is directed. (See Parfit 1971, 1984/1987.) I think Parfit is wrong about this, but presenting my reasons here would take us too far afield. [Added in 2009: For further discussion of this issue, see Ch. 4 below.]
to happen to you, since shortly before the torture something else will be done to [you] which will make [you] forget the announcement. Indeed, you will 'not remember any of the things [you are] now in a position to remember.' In fact, at the moment of torment you will 'not only not remember the things [you are] now in a position to remember, but will have a different set of impressions of [your] past, quite different from the memories [you] now have'—a set of memories and impressions that exactly fit the past of some other person." Of this situation, Williams writes: "Fear, surely, would still be the proper reaction: and not because one did not know what was going to happen, but because in some vital respect at least one did know what was going to happen—torture, which one can indeed expect to happen to oneself, and to be preceded by certain mental derangements as well."^{29}

The intuition evoked by this second version—which, of course, is just a one-sided presentation of the original scenario—is that we seem to be able to make sense of there being some sort of procedure whereby two persons could, so to speak, "swap minds." To the extent that I bear to my future self a relation of rational prudential concern, I might properly bear that relation to someone with whom I shared no psychological connections at all. My biological animal seems also to be me.^{30}

Having presented these two versions of the story, Williams goes on to introduce a series of cases leading up to the story with which we were just presented.^{31}

(9) First case: "A is subjected to an operation which produces total amnesia."

(10) Second case: "amnesia is produced in A, and other interference leads to certain changes in his character."

(11) Third case: "changes in his character are produced, and at the same time certain illusory 'memory' beliefs are induced in him" that do not correspond to the memories of any actual person.

(12) Fourth case: "the same as [(11)], except that both the character traits and 'memory' impressions are designed to be appropriate to another actual person, B."
Part I. Thought Experiments

(13) Fifth case: "the same as [(12)], except that the result is produced by putting the information into A from the brain of B, by a method which leaves B the same as he was before."

(14) Sixth case: "the same happens to A as in [(13)], but... a similar operation is [also] conducted in the reverse direction," resulting in corresponding changes in B.

In general, consideration of this series of stories has evoked three different sorts of responses. Williams himself contends that our inclination to say that A would be right to be prudentially concerned about what would happen to the A-body person in the first (amnesia) case carries through all the way to the fifth (brain-state transfer) case; and since the difference between the fifth and the sixth cases does not involve any difference in what happens to A, then it carries through to the sixth case as well. So the correct criterion of identity, Williams thinks, is a bodily criterion. By contrast, advocates of the psychological criterion of personal identity conclude just the opposite. Already in the first case, they contend, what matters for prudential concern has been lost; the Lockean person associated with the A-body has already been eliminated by the process that produces amnesia. So from the very beginning, the A-body person does not have warranted prudential concern for any of the characters described in the six cases. Finally, some have taken the scenario to show that both lay legitimate claim on our intuitions, and that what the case shows is that our concept of person is not definitively committed to the primacy of one or the other.

In short, the Williams story perfectly illustrates the claim I have been making: that our ability to make sense of imaginary scenarios in which features that coincide in nearly all actual cases are recombined in novel ways far outruns our ability to make judgments about them. We may well feel that a scenario is perfectly coherent, without knowing what we would do or say were we to encounter it. In such circumstances, our evaluation of the case is likely to depend upon how the case is presented.  

4. The More Rational Decision

4.1. Decision Rationality

Consider the story: you have $10 in your wallet. Someone approaches you and asks if you would like to buy a ticket? A: No. But in the following case: A is at the station with a ticket arriving at the train. Someone asks A if he would like to buy another one. A: Yes.

Clearly, however, the case is in the form of the following: A was offered a ticket which is the same as the $10 in his wallet, the $10 lost: $10 and the $10 expected gained: $0. If this is the case, then this particular ticket is imagined as equally trivial. Of course, there is no reason that one cannot imagine the second case to the same extent, and irrational to that extent. We have two stories that are intended to justify our intuitions.  

[32] More complicated versions of this view draw the line somewhat later along the spectrum. See e.g. Noonan 1989.

[33] Modified for an interesting discussion of this case and its implications, see Revan 1998, ch. 2; for remarks on Revan’s account, see Gendler 2002a.

[34] One might say, for instance, that the story provides evidence in favor of a “closest continuator” theory: that “to be something later is to be its closest continuator” (Nozick 1981: 33). According to such a theory, to be X later is to be whatever entity it is that most closely matches the profile of characteristics associated with being X (provided that the entity matches this profile sufficiently closely to be a candidate). The closest continuator view raises certain perplexing metaphysical puzzles about the ex nunc determination of identity. (See e.g. Johnston 1987 and Noonan 1989 for objections.) But I am not convinced that the question of personal identity is straightforwardly metaphysical in the way critics of the closest continuator theory assume it must be. [Added in 2009: For further brief remarks on this suggestion, see Gendler 1999.]
4. The Moral of the Stories

4.1. Decision-making and Assimilation to a Class of Familiar Cases

Consider the following situation, which is standard fare in discussions of rationality.\footnote{[35] This case, along with many others, is first discussed in Tversky and Kahneman 1981. Again, to those familiar with the enormous literature on these subjects [added in 2009: a wider group of philosophers than a decade ago when this essay was first published] my discussions here may seem superficial. As before, the point I am trying to establish is a simple one, so the many nuances of recent discussions are not relevant to my purposes. The reader interested in following up on these issues might (thankfully) begin with the papers collected in Kahneman, Slovic, and Tversky 1982 and [reference added in 2009] Kahneman and Tversky 2000.} Suppose that on your way to see a play you lose a $10 bill from your wallet. Upon arriving at the theater, you discover this loss. However, you still have money in your wallet, and tickets to the play are $10 each. Do you still buy a ticket? A vast majority of people answer “yes.”\footnote{[36] The question was phrased as follows: “Imagine that you have decided to see a play where admission is $10 per ticket. As you enter the theater, you discover that you have lost a $10 bill. Would you still pay $10 for a ticket to the play?” Out of nearly 200 subjects, 88% answered “yes”; 12% answered “no” (Tversky and Kahneman 1981: 457).} By contrast, consider the following case. You have already purchased a $10 ticket for the play but upon arriving at the theater, you discover that you have lost the ticket. Would you buy another ticket? Most people answer “no.”\footnote{[37] “Imagine that you have decided to see a play and paid the admission price of $10 per ticket. As you enter the theater you discover that you have lost your ticket. The seat was not marked and the ticket cannot be recovered. Would you pay $10 for another ticket?” Out of 200 subjects asked, 46% said “yes”; 54% said “no” (Tversky and Kahneman 1981: 457).}

Clearly, however, at some level these cases are the same. They differ only in the form of the $10 item lost: in the first case, the $10 lost is in the form of cash, which is the paradigmatically interchangeable commodity; in the second case, the $10 lost is in the form of a movie ticket which is directly related to the expected gain (seeing the play).\footnote{[38] Tversky and Kahneman diagnose the outcome as follows: “The marked difference between [the two cases] is an effect of psychological accounting. We propose that the purchase of a new ticket in [the second case] is entered in the account that was set up by the purchase of the original ticket. In terms of this account, the expense required to see the show is $20, a cost which many of our respondents apparently found excessive. In [the first case], on the other hand, the loss of $10 is not linked specifically to the ticket purchase and its effect on the decision is accordingly slight.” (Tversky and Kahneman 1981: 457.)} But as far as their implications for action in this particular case are concerned, this difference is trivial: the purchase of the ticket is imagined to have been basically effortless; to acquire a second would be equally trifling. Indeed, the similarity between them can be seen from the fact that one can easily talk oneself out of either reaction by assimilating the first case to the second, or the second case to the first. So it seems \textit{prima facie} irrational to treat the cases differently. If we can see that, deep down, all we have are two alternate descriptions of the same state of affairs, how can we justify our inclination to deal with them asymmetrically? To do so seems to be
Part I. Thought Experiments

to give judicially import to factors that ought not to matter: how could the mere fact that we describe something in one way rather than another justify such a striking difference in the attitudes we bear towards it?

Although I agree that it is not rationally mandatory that the attitudes we take towards the two situations should differ, I also do not think that it is rationally prohibited. That is, I think that it is rationally permitted that we should take these different perspectives, and I want to say a few words about why. When we make decisions, we make sense of the particular scenario with which we are confronted by assimilating it to a class of familiar cases.\textsuperscript{39} Since in general it is economically unwise to indulge ourselves by automatically replacing any item that breaks or is lost, and since we treat the lost ticket case under that rubric, we are hesitant to buy another. And since in general it would make us unhappy to deny ourselves enjoyment in one sphere whenever something has gone wrong in another, and since we treat the lost $10-bill case under that rubric, we are prepared to spend the money. Even when the baseline similarity between the two cases is brought out to us, we may maintain that it makes good sense to treat them differently, since each is best understood as belonging to one of two classes of cases between which we make a justifiable distinction.

With this idea in place, let us return to the Williams case. The perplexity there, you will recall, is that our response to the case seems to be frame-dependent. On one way of telling the story, we are inclined to take it as evidence in favor of the hypothesis that personal identity over time is a matter of physical continuity; on another, we are inclined to see it as supporting the view that personal identity over time is a matter of continuity of psychology.

The discussion above suggests a diagnosis of the difference in response. In the first presentation, when we try to make sense of the story in light of our general assumptions about ordinary cases of personal identity, we focus on issues involving the body. And in ordinary cases, continuity of body assures continuity of personhood. So when the Williams story is framed in a way that foregrounds the bodily perspective, we take this feature to be sufficient for continuity of personhood. Likewise with the second presentation. Framed in a way that highlights the psychological continuity involved, the case is assimilated to ordinary cases under the following line of reasoning. We notice that there is a feature possessed by the B-body person—psychological continuity—which suffices for continuity of personhood in ordinary cases. And so we are inclined to take that feature as decisive in this case as well.

\textsuperscript{39} I take this analysis from Nozick 1993, chs. 1 and 2. Among other things, Nozick there suggests that the bringing about/allowing distinction in ethics might be viewed as an example of the sort of baseline effect which Kahneman and Tversky have discussed (see Nozick 1993: 69 n.). For two independently arrived-at workings-out of this idea, see Horowitz 1998 and Kamm 1998. [Added in 2009: More recently, versions of this suggestion have been explored in a number of essays, including Sinnott-Armstrong 2008c and Driver 2008a. I investigate a variant of this proposal in Ch. 6 below.]
We have, then, a story about how we come to treat the two presentations so differently: we make use of a process of reasoning that we seem to employ quite generally. That is, we assimilate a situation under one description to one class of cases, and we assimilate that same situation under another description to a different set of actions. But it could be that in so doing we are making a mistake. Just as one might defend a substantive global theory of rationality according to which one or the other of the attitudes towards the missing $10/ticket is seen as categorically correct, so too might one defend a theory of personal identity according to which one or the other of the views we have been considering actually captures the truth about the nature and importance of personal identity. Indeed, this is precisely what the sorts of thought experiments described in the previous section seek to establish; they offer carefully described scenarios in which relevant and irrelevant features can be separated out so that we can determine which are essential and which accidental. If I am right in my claim that such cases do not, in general, show what their advocates purport that they show, I need to offer further reasons.

4.2. Exceptions and Norms

I have suggested above that the structure of the concept person is such that it applies in central cases as the result of the correlative appearance of a set of frequently associated characteristics, and to other cases as the result of our assimilation of them to these central cases. In defending this view, I now describe two strategies that might be employed when confronted with exceptional cases.

The exception-as-scalpel strategy uses exceptional cases as a way of progressively narrowing the range of characteristics required for the application of a concept by allowing us to isolate the essential features for concept-application from those which are merely ordinarily correlative. So, for instance, suppose that entities falling under a certain concept generally have characteristics \(a, b, c, d,\) and \(e.\) Suppose further that we come upon some entity that falls under the concept, but that has only \(b\) and \(d.\) We are then entitled to conclude that of the five characteristics typically associated with entities of the type under discussion, at most \(b\) and \(d\) are required characteristics of any entity that falls within the purview of the theory.

By contrast, the exception-as-cantilever strategy views the category-membership of exceptional cases as essentially reliant on the ordinary instances against which they can be seen as exceptions. So, for instance, suppose again that entities that fall under the concept in question generally have characteristics \(a, b, c, d,\) and \(e,\) and suppose further that some entity is found that has only \(b\) and \(d,\) but that nonetheless falls under the concept. According to the second strategy, the proper thing to say about the entity in question is that it falls under
Part I. Thought Experiments

the concept only because it is similar in certain crucial ways to more typical instances of entities that the theory describes.

Applying the exception-as-scalpel strategy to the Williams case, we discover the following. The first presentation reveals that physical continuity is not necessary for diachronic personal identity, since it seems that there are cases where a person’s future self might share none of her original physical matter. The second presentation reveals that psychological continuity is not necessary for diachronic personal identity, since it seems that there are cases where a person’s future self might share none of her original psychological characteristics. Moreover, neither feature could be sufficient, unless the pre-operation individual is supposed to be identical to both her continuers in the second scenario. So neither physical nor psychological continuity is either necessary or sufficient for diachronic personal identity. This suggests that our concept of person is not organized around a set of necessary and sufficient conditions that play a role in how we identify candidates as falling under that concept. The exception-as-scalpel strategy cuts away too much.

By contrast, consider the exception-as-cantilever strategy. On this strategy, the Williams case is to be understood as follows. The first presentation reveals that there are imaginable cases in which we would be inclined to attribute diachronic identity in the absence of any sort of psychological continuity; the second presentation reveals the same about physical continuity. But rather than concluding something about the (lack of) necessary and sufficient conditions for the application of the concept person, the exception-as-cantilever strategy tells us to conclude this about our classification of these exceptional cases as cases where diachronic personal identity obtains: our decisions about these cases are justified by the rational permissibility of assimilating them to ordinary cases. In the first scenario, we focus on the similarity that concerns physical continuity; in the second, we focus on psychology. But in both cases, we are

---

40 [40] Alternatively, one might follow Parfit in saying that prudential concern need not track identity, in which case thinking about these scenarios tells us nothing decisive about personal identity (since the evidence they used in favor of the attribution of identity was: projected prudential concern). Depending on whether one thinks that prudential concern generally tracks (identity, and depending on which sorts of cases “generally” is supposed to cover—all normal cases? all actual cases? all (technologically, biologically, physically, metaphysically, logically) possible cases?—our projected judgments about a particular scenario will turn out to be very, somewhat, or not at all relevant to determining the identity-conditions for personhood over time. But this only helps to make my point: if judgments about prudential concern are relevant to making such determinations, then I refer the reader to my argument in the paragraph to which this footnote is attached. If judgments about prudential concern are not relevant to making such determinations, then it is unclear how contemplation of the sorts of imaginary scenarios described in the literature is supposed to give us the right sort of information (since it seems that our primary basis for judgments about identity is projection of prudential concern).

41 [41] Cf. Johnston 1987 for related discussion of these points. [Added in 2009: I return to this issue in Ch. 4 below.]
cantilevering out from the set of generally obtaining correlations that characterize ordinary cases.\textsuperscript{42} The persons in these far-flung stories are persons by courtesy only.

4.3. Conclusion

Thinking about actual and imaginary exceptional cases is indispensable if we wish to avoid mistaking accidental regularities for regularities that reflect a deeper truth about the world. And because the world does not provide us with easily accessible instances of all the combinations there might be, thought experiment—the contemplation of a well-described imaginary scenario in order to answer a specific question about some non-imaginary situation—is an indispensable technique: in philosophy, in science, and in ordinary reasoning. At the same time, critics of this methodology have correctly pointed out that it can be misused. Most have suggested that constraints be imposed on the sort of possibility involved in the imaginary scenario, that the more far-fetched the case, the less likely it is to be informative.

I have tried to show that this analysis locates the problem in the wrong place: the risk of misusing thought experiment arises not from the outlandishness of the scenarios, but from the structure of the concept that the thought experiment is intended to explore. Concepts structured in certain ways (those organized around a set of necessary and sufficient conditions that play a role in our identification of instances of that concept) can be clarified by means of the exception-as-scalpel strategy described in the last section; those structured in other ways may require us to treat exceptions-as-cantilevers.

I have argued that the concept of person, and with it, the concept of personal identity, is a concept of the latter sort. Diachronic personal identity is a matter of the continued coincidence of enough of the factors that ordinarily allow us to persist over time: psychological, physical, and perhaps even social factors play a role. We can make sense of cases where one or another of these features is absent. But this does not mean that our evaluations of them will be reliable guides to what matters in ordinary cases. For, as I have argued throughout this piece, our ability to make sense of exceptional situations far outruns our ability to make reliable judgments about them.\textsuperscript{43}

\textsuperscript{42} To use a slightly different metaphor: the “persons” in these stories come in through the back gate—but the only reason there is a back gate for them to come in through is because ordinary cases form a certain sort of fence.

\textsuperscript{43} [*] Acknowledgements: For comments on earlier material from which this chapter is descended, some of which (particularly material in §4) appeared in my 1996 dissertation, I am grateful to Robert Nozick, Derek Parfit, and Hilary Putnam. For discussion of the more recent material, I thank Shaun Gallagher, John Hawthorne, Zolán Gendler Seabo, and two anonymous referees for the \textit{Journal of Consciousness Studies}. 

73