Imaginary Exceptions: On the Powers and Limits of Thought Experiment

A thesis presented

by

Tamar Szabó Gendler

to

The Department of Philosophy

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Philosophy

Harvard University

Cambridge, Massachusetts

August 1996

Abstract

Thought experiment is one of the most widely-used and least understood techniques in philosophy. A thought experiment is 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. The aim of my dissertation is to show that both the powers and the limits of this methodology can be traced to the fact that when the contemplation of an imaginary scenario brings us to new knowledge, it does so by forcing us to make sense of *exceptional cases*.

The dissertation has five chapters: an introduction, three case studies and a conclusion. My main contention is that certain patterns of features which coincide only fortuitously may nonetheless play a central role in the organization of our concepts, and that to the extent that imaginary scenarios involve disruptions of these patterns, our first-order judgments about them are often distorted or even inverted.

In the introduction and conclusion, I discuss the role of imaginary exceptions in the acquisition of new knowledge. I argue that appeal to imaginary cases is unavoidable because the world is neither maximally replete nor effortlessly navigable, and that appeal to exceptional cases is indispensable if we wish to avoid mistaking accidental regularities for regularities that reflect deeper truths about the world.

In the first case study, I discuss a famous thought experiment of Galileo's, and I try to show that the guided contemplation of an imaginary scenario can provide us with new scientific knowledge in a way that argument alone cannot. In the second

case study, I try to show that standard interpretations of the puzzle of the Ship of Theseus founder because they ignore the importance of the background norms against which we can make sense of local instances of extrinsically-determined identity. And in the third case study, I try to show that thought experiments in the personal identity literature are inconclusive because they disregard the explanatory role played by contingent facts about the ways human beings come into existence.

It is good to know something of the customs of various peoples, in order to judge our own more objectively, and so that we do not make the mistake of the untraveled in supposing that everything contrary to our customs is ridiculous and irrational. But when one spends too much time traveling, one becomes at last a stranger at home…In addition, fiction makes us imagine a number of events as possible which are really impossible…Thus it happens that those who regulate their behavior by the examples they find in books are apt to fall into the extravagances of the knights of romances, and undertake projects which it is beyond their ability to complete.

--René Descartes, *Discourse on Method*, Part I

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

Suppose…that a diamond could be crystallized in the midst of a cushion of soft cotton, and should remain there until it was finally burned up. Would it be false to say that the diamond was soft? This seems a foolish question, and would be so, in fact, except in the realm of logic. There such questions are often of the greatest utility as serving to bring logical principles into sharper relief than real discussions ever could. In studying logic we must…consider them with attentive care, in order to make out the principles involved…[T]he question of what would occur under circumstances which do not actually arise is not a question of fact, but only of the most perspicuous arrangement of them.

--C.S. Peirce, “How to Make Our Ideas Clear,” Section III

Table of Contents

[*Note added 10/06/06 upon conversion to .pdf*: pagination in the Table of Contents refers to the original dissertation. In the process of conversion to .pdf, some sections may have shifted by 1 or 2 pages.]

Acknowledgments

My first debt is to the three members of my committee—Robert Nozick, Derek Parfit and Hilary Putnam—each of whom has been an exemplary teacher, mentor and advisor at many levels. Each has been my teacher in a graduate seminar, my mentor as a teaching assistant, and my advisor as I wrote my dissertation. But each has also been teacher, mentor and advisor in a deeper way as well.

Hilary Putnam's writings have influenced my work for many years, and his support and encouragement, both philosophically and personally, have been invaluable throughout my graduate career. Derek Parfit, whose seminar on personal identity originally sparked my interest in the topic of thought experiments, has been exorbitantly generous with his time, offering copious written comments on early drafts, and spending hours on the phone long-distance from England as the dissertation neared completion. And Robert Nozick, who has known me since the day I was born, has been—from my first-year seminar and my second-year paper, through my tenth-round e-mail message and eleventh-hour defense—an incomparable source of criticism, support and enthusiasm. I doubt any graduate student has had three advisors who are such *fun* to talk philosophy with; I always come away from our conversations enlivened, provoked, energized, encouraged, inspired, impressed and—most of all—*very* glad to have chosen a career as a philosopher.

This dissertation owes its existence to many others besides. During my graduate career, I received funding from the Mellon Foundation (1989-90, 1990-91 and 1994-95) and the National Science Foundation (1991-92, 1992-93 and 1993-94). I am grateful to both of

these organizations for their financial support, and I hope both will have the resources to fund many others in the future.

My initial training in philosophy took place at Yale University, and I spent my first year of graduate school at UC Berkeley. Among the teachers I had at these places, I owe special thanks to Karsten Harries, Jonathan Lear and Ruth Barcan Marcus from the Yale Philosophy Department; Neil Ribe from the Yale Geology Department; Louis Dupré, Cyrus Hamlin and Ralph Hexter from the Yale Humanities major; Hannah Ginsborg and Lisa Lloyd from the UC Berkeley Philosophy Department; and John Dupré from the Stanford Philosophy Department.

Versions of (what have ended up as parts of) chapters 1, 2 and 4 have been presented as papers before audiences at the Second Annual Harvard-MIT graduate student conference (chapter 1); at the MEPHISTOS Conference in the Methodology, Philosophy, History and Sociology of Science and at the Harvard University Nth Year Seminar (chapter 2); and at the Departments of Philosophy at Harvard, Cornell and Syracuse Universities (chapter 4). For comments on those earlier versions, I am particularly grateful to Andrew Botterell and Steven Gross (chapter 1); Michael Glanzberg, Steven Gross, Ned Hall, Thomas Kuhn, John Murdoch, W.V. Quine, Sherri Roush, Simon Saunders and Alison Simmons (chapter 2); and Richard Boyd, Warren Goldfarb, Harold Hodes, Terence Irwin, Norman Kretzmann, Scott MacDonald, Elijah Millgram, Richard Miller, Ian Proops, Sydney Shoemaker, Susanna Siegel, Jason Stanley, Nicholas Sturgeon and Jennifer Whiting (chapter 4). In addition, I owe thanks to Myles Burnyeat, John Murdoch, Eileen O'Neill and Eleanor Stump for conversations which helped me to put the three central chapters in historical perspective.

During the fall term of 1994, I had the good fortune to attend Roger Shepard's William James lectures at Harvard University, which were on the topic of thought experiment. From these lectures ensued many hours of fruitful conversation with Roger, as well as an introduction to the world of cognitive psychology. It is through Roger that I came to know, in person and in writing, Susan Carey, Carol Krumhansl, and Liz Spelke; the influence of these connections are apparent throughout the dissertation.

Other interdisciplinary connections have also been of tremendous benefit. I have learned much from intellectual conversations, both formal and informal, with Peter Berkowitz (political theory), Neal Dolan (English), Rainer Forst (political theory), Naomi Gendler (law), Caroline Guindon (musicology), and John Henriksen (comparative literature). Each has, in his or her way, helped me to gain perspective on my own work, and on the enterprise of philosophy itself.

Over the years, I have spent many hours discussing philosophy with fellow graduate students at Harvard, and with counterparts at MIT, Princeton and Cornell. At the risk of leaving out the names of many who equally deserve acknowledgment, I would in particular like to thank Scott Brewer, Michael Glanzberg, and Steven Gross from Harvard; Andrew Botterell and Daniel Stoljar from MIT; Garrett Deckel and David Sosa from Princeton; and Jennifer Dworkin and Susanna Siegel from Cornell. Two young faculty members have been especially generous with their time; I am grateful to Elijah Millgram (Princeton) and Simon Saunders (Harvard) for many hours of general philosophical conversation and advice.

Six other people deserve special thanks. Throughout graduate school, I have been lucky to have maintained philosophical friendships with two people who, had things gone very slightly otherwise for any of us, would have been my classmates: Jenann Ismael and

Teresa Robertson. To each of them, I am grateful for late-night phone calls and early-morning e-mails, for words of encouragement and words of honesty, and for comments on first drafts of papers that never had final drafts (and on final drafts of papers that never had first drafts). Finally, four Harvard classmates stand out as long-term partners in philosophical conversation, people whom I look forward to having as life-long philosophical colleagues: Tal Brewer, Tony Laden, Tamar Schapiro and Sherri Roush. To Tal and Tony and Tammy I owe my constant awareness that there are deep questions in moral and political philosophy with which any philosopher needs to grapple; to Sherri and Jenann and Teresa I owe my awareness that this is true of metaphysics and epistemology as well.

I completed this dissertation in Ithaca, New York, and I am grateful to members of the Cornell Philosophy Department for their hospitality. My post as a part-time visiting lecturer in the department provided me with access to the university's resources, and allowed me to audit seminars with Carl Ginet (on *a priori* knowledge) and Jennifer Whiting (on self and psychopathology). My affiliation also gave me the chance to participate in reading groups that included (among others) Carl Ginet, Karen Jones, Scott MacDonald, Richard Miller, Sydney Shoemaker, Jason Stanley, Zoltán Gendler Szabó, Jennifer Whiting and Allen Wood; I have benefited from conversations with them all. During my time in Ithaca, I spent many fruitful hours in philosophical conversation with Jason Stanley and Jennifer Whiting, each of whom, in ways that he or she will surely recognize, has shaped aspects of this dissertation, and whom I am glad to call a friend.

My most gracious host in Ithaca has been Norman Kretzmann, whose generosity began with his allowing me to share his office and his library carrel, and continued throughout the year in countless ways. Norman read drafts of chapters 2 and 4, suggested bibliographic leads, and provided words of encouragement at the times when they were most needed.

Finally, I am extremely grateful to the members of my family, who have been invaluable sources both of intellectual and emotional support throughout the dissertation process. My parents, Mary and Everett Gendler, provided me with a home in which learning was valued and discussion was encouraged; throughout the years of graduate school, they were encouraging and supportive, and I am grateful to both of them. While I was at Harvard, I saw my sister Naomi Gendler daily; after moving to Ithaca, I spoke to her on the phone at least twice that often. Throughout, she has been an incomparable source of good sense, sympathy, editorial advice, and support.

Last of all, I would like to thank Zoltán Gendler Szabó. This dissertation could not have been written without his help—philosophical, editorial, and emotional. Nearly every page bears the imprint of his advice, and those which do not are the worse for it. I am unspeakably fortunate to have him as a philosophical companion, and even more fortunate to have him as a husband.

New Haven, Connecticut
23 August 1996

1. Introduction

Thinking about imaginary cases can help us learn new things about the world. This simple fact is both a commonplace, and a puzzle. It is a commonplace because it is undeniable that imaginary test cases play a central role in our investigation of the world—in legal reasoning, in linguistic theorizing, in philosophical inquiry, in scientific exploration, and in ordinary conversation. And it is a puzzle because it is *prima facie* surprising that thinking about what there isn't and how things aren't should help us to learn about what there is and how things are.

The goal of this dissertation is to show that this simple fact is both less of a commonplace and less of a puzzle than it might initially seem. It is less of a commonplace because the failures of this technique are far more prevalent than its enthusiasts would admit. And it is less of a puzzle because what accounts for its success is far less mysterious than its detractors would concede. Moreover, the explanation for both its success and its failure is the same. Both the powers and the limits of the technique can be traced to the fact that when the contemplation of an imaginary scenario brings us to new knowledge, it does so by forcing us to make sense of an *exceptional case*.

1.1 Exceptional Cases

1.1.1 Characterization of "Exceptional Case"

What I mean by "exceptional case" can be characterized more precisely as follows. (I will give an example of what I mean in the next paragraph.) Suppose that there is some theory which, in general, (i) accounts for entities or situations that are identified—by users of the theory—under more than one description or on the basis of more than one characteristic. Now take two of these descriptions such that: (ii) It is not conceptually necessary that these descriptions co-vary, and (iii) one description has privileged status within the theory, such that what the theory purportedly aims to do is to say something about those entities or situations under that description. Let us call this the "privileged description" and call the corresponding characteristic the "privileged characteristic." Suppose that (iv) in a majority of cases, the entities or situations in question can also be picked out by some proxy characteristic which is generally a concomitant of the privileged characteristic. Then (v) An exceptional case is then one where there is some entity or situation which falls under the purview of the theory, but which lacks the proxy characteristic[27].

As an example, take an informal theory of speed which concerns itself with cars and people when they participate in races[28]. (i) Users of the theory identify the faster participant (that is, the participant which has the greatest average speed) by

[27] I do not mean to suggest that this is the only way in which a case might be exceptional. Many of the scientific examples I go on to discuss in the text below are exceptional in a much more mundane sense; they offer instances of phenomena that are *atypical* because they manifest an unusual* configuration of properties. The characterization of "exceptional" that I have just offered is a special case of this more general phenomenon. A discussion of the ways in which exceptional cases of this more special sort are connected to exceptional cases in general can be found in the concluding chapter.

> *Obviously, what is usual or unusual is so with respect to some set of background assumptions; nothing is unusual *simpliciter*.

[28] A more detailed discussion of this example can be found in Appendix A in section **A.1**.

Continues on next page…

examining how the participants look when they in are motion (perceptual blurriness), by checking to see which participant reaches the finish line first (goal-reaching), and by calculating the participant's mean speed over the course. (ii) It is not conceptually necessary that these criteria coincide[29]: the "blurriest" participant will not always reach the finish line first, nor will the blurriest participant or the participant who reaches the finish line first always be the participant who completes a given course in the shortest amount of time. (iii) What the theory really concerns is this last characteristic, namely, average velocity; this is the "privileged characteristic." (iv) In most cases, the participant which has the greatest average velocity will also be the participant which is "blurriest," as well as being the participant which reaches the finish line first. Hence either of these characteristics can serve a proxy for the privileged characteristic. (v) An exceptional case is one where the participant which is "blurriest," or the participant which reaches the goal first is not in fact the participant with the greatest average velocity.

Here is a second example. A theory of sex (i) accounts for persons who are identified—by users of the theory—on the basis of more than one characteristic; for instance, women are identified on the basis of reproductive capacities, chromosomal structure, manner of dress, body shape, social role, and emotional tendencies. (ii) It is not conceptually necessary that these descriptions co-vary: a person may have a uterus and wear army boots, or she may work as a truck driver and have a close and

[29] As it would be, for instance, if the theory concerned closed planar figures and the two characteristics were *trilaterality* and *triangularity*.

caring emotional relationship with her children. (iii) One description has privileged status within the theory, such that what the theory purportedly aims to do is to say something about persons under that description. In this case, the "privileged description" is the biological description, and the corresponding "privileged characteristics" are, let us say, those that concern chromosomal structure. (iv) In a majority of cases, persons can be picked out by some proxy characteristic which is generally a concomitant of the privileged characteristic. So, for instance, a person wearing a skirt is likely to have the chromosomal structure in question, as is a person who is a nursery school teacher. (v) An exceptional case is then one where some person who falls under the purview of the theory as a result of having the requisite chromosomal structure lacks the proxy characteristics: for instance, someone who wears jeans and a flannel shirt, is romantically involved with a woman, works in construction, and has a violent temper.

The term "exceptional case" may also apply to cases within a theory that lacks characteristic (iii), that is, a theory where no description has privileged status. Instead, such a theory concerns entities that have some proper or improper subset of the cluster of characteristics that together typify the entities which the theory concerns. Of such a theory, we might say (iii') some cluster of properties characterizes entities described by the theory, such that what the theory purportedly aims to do is to say something about those entities or situations which exhibit some significant portion of these characteristics. We will call entities that exhibit the designated properties [30] "paradigm cases[31]." Following this adaptation of (iii), we might say here that (iv') in a majority of cases, the entities or situations in question

[30] Paradigm cases may exhibit all of the designated properties, or as many as normally possible if it is not normally the case (or not possible) that all of the characteristics in question are held conjointly by a single entity or occur conjointly in a single situation. My characterization here is intentionally sketchy, as I am hoping to remain neutral among the following sorts of cases. (The list that follows is not intended to be exhaustive.) Suppose Xs are characterized by having some (proper or improper) subset of the properties *a*, *b*, *c*, and *d*. It may be that:

- Paradigm X has all four of these properties.
- Paradigm X has some particular trio of these properties; no X does (or could) have all four.
- Paradigm X has some particular trio of these properties; some Xs do (or could) have all four.

[31] Note that speaking of a "paradigm" may be misleading in cases of the following sort:

Xs have properties *a* and *b*
Ys have properties *b* and *c*
Zs have properties *a* and *c*

Xs, Ys, and Zs are the only instances of the entities in question—call them Ns. Here we have two choices:

(1) we might say that there *is* a paradigm case of being an N, namely being an entity that has *a, b* and *c*; or

(2) we might say that there is no single paradigmatic N, but rather, that there are three prototypical Ns, namely: Xs, Ys, and Zs.

There are two problems with the first reply: (a) In some sense, the entity with *a, b* and *c*—even if it were to exist—would be extremely *atypical*, in that (i) it resembles no actual entity, and (ii) it differs from all actual entities in a straightforwardly specifiable way, namely, it has three of the properties that appear only pair-wise in all actual instances (and the latter might itself be what we seek). (b) There are certain problems which arise with imaginary cases that do not arise with actual cases, such that taking an imaginary case as a paradigm may lead to certain sorts of problems. For these reasons, it makes sense to go with the second reply.

One last remark: an analogue to this sort of case can be found in a particular species of birds which inhabit a circle of territory around the Arctic circle, of which some rough approximation of the following is true: Those in Greenland can breed with those in Norway and those on Baffin Island; those on Baffin Island can breed with those in Greenland and those in mainland Canada; those in mainland Canada can breed with those on the Baffin Island and those in Kamchatka; and those in

Continues on next page…

can be picked by many of the characteristics which together characterize paradigm cases. (v') An exceptional case is one where there is some entity or situation which falls under the purview of the theory, but which cannot be picked out in the way (iv') requires[32/33].

A theory of gender is a theory of this sort. (iii') Some cluster of properties characterizes entities described by the theory, such what the theory purportedly aims to do is to say something about those entities or situations which exhibit some significant portion of these characteristics. So, for instance, females are persons who occupy some particular set of social roles, exhibiting a range of characteristics which epitomize being female. Persons who exhibit some (maximal or near-maximal) subset[34] of the designated properties are paradigm instances of females. (iv') in a

Kamchatka can breed with those in mainland Canada and those in Norway—but these relations are not transitive. (I thank Dick Boyd for this example.)

That is:

Greenland

Baffin Island Norway

Canada Kamchatka

[32] Either because the entity has too few of the relevant features, or because it has too many.

[33] In cases of the sort described in footnote 31, we would then say: (iv") in a majority of cases, the entities or situations in question can be picked by some sufficiently large subset of the characteristics which together make up the disjoint set of characteristics which can be found among prototypical cases. (v") An exceptional case is one where there is some entity or situation which falls under the purview of the theory, but which cannot be picked out in the way (iv") requires.

[34] Indeed, one frequently-made critique of traditional gender roles has been that it is impossible for a woman to satisfy simultaneously all of the requirements of being a

Continues on next page…

majority of cases, females can picked out by many of the characteristics which together characterize paradigm cases. (v') An exceptional case is one where some person falls under the purview of the theory but cannot be picked out in the way (iv') requires, that is, some case where there is a person who is—according to the theory in question—a female, but who bears few of the attributes associated with paradigm instances of females[35].

1.1.2 Ways of Accounting for Exceptional Cases

There are two main ways a theory can account for exceptional cases; or, to speak a bit more precisely, there are two main sorts of strategies users of a theory can employ in accounting for exceptional cases. The first strategy is to use exceptional cases as a way of progressively narrowing the range of privileged characteristics. So, for instance, suppose entities accounted for by the theory in question generally have characteristics *a*, *b*, *c*, *d*, and *e*. Suppose further that as an exceptional case, some entity is found which falls within the purview of the theory, but which has only *b* and *d*. It then follows that no characteristics other than *b* and *d* can be privileged characteristics in the sense that they necessarily belong to any entity which falls within the purview of the theory.

"successful" female; she cannot be simultaneously delicate *and* capable of doing all the housework, morally upstanding *and* sexually available, deferential in decision-making *and* capable of taking charge, etc.

[35] It might also be that an individual who bears *all* the attributes would be exceptional, but the exceptionality would be of a different sort. (See also footnotes 27, 31 and 32.)

Continues on next page…

Such an attitude towards exceptional cases involves using them as test cases to ascertain necessary and sufficient conditions. The conclusion then drawn is that even in non-exceptional cases, the characteristics that really matter are those that are present in the exceptional cases as well. According to this strategy, one uses exceptional cases to ascertain the theory's *exception-driven norms*. The *exceptions drive* interpretation of the *norms*; what is taken to matter about normal cases is whatever it is that they have in common with exceptional cases. So if normal cases of Xs have *a*, *b*, *c*, *d*, and *e*, but exceptional cases have only *b* and *d*, then on this strategy, we would say that all that matters even in normal cases of X-ness is that *b* and *d* are present.

The second strategy is to view exceptional cases as evidence for the strength of the theory's core. So, for instance, suppose again that entities accounted for by the theory in question generally have characteristics *a*, *b*, *c*, *d*, and *e*, and suppose further that some entity is found which has only *b* and *d*, but which nonetheless seems to fall within the purview of the theory. According to the second strategy, the proper thing to say about the entity in question is that it falls within the purview of the theory, but only because it is similar in certain crucial ways to more typical instances of entities which the theory describes. Under this strategy, one uses exceptional cases to ascertain the theory's *norm-driven exceptions*. The *norms drive* interpretation of the *exceptions*. On the basis of this strategy, one concludes that what it is that allows the exceptional cases to be cases at all is that they have enough in common with the normal cases. The reason the entity with only *b* and *d* counts as an X is because in general, *b* and *d* are found only in association with *a*, *c,* and *e* as well.

So, for instance, suppose again that both women and females generally have characteristics *a, b*, *c*, *d*, and *e*, where these are: *a*: wears thus-and-such sorts of clothing; *b*: has thus-and-such chromosomal structure; *c*: occupies thus-and-such social role; *d*: has thus-and-such body parts; *e*: tends to respond to stressful situations in thus-and-such way. And suppose that some person is found who is held to be both female and a woman, but who has only *b* and *d*. That is, the person in question has the designated chromosomal structure and body parts, but none of the social characteristics ordinarily associated with being a woman or being female. According to the first strategy, such a case would show that all that matters about being a woman or being female is having the requisite genetic and bodily characteristics. According to the second strategy, such a case would show that the reason such genetic and bodily structures "count" in making one female or a woman is because, in general, they are associated with the social characteristics enumerated above.

1.1.3 Patterns of Accounting

It seems to me (as I hope it seems to my reader) that the first strategy (exception-driven norms) is appropriate when we are talking about what I am calling *women*[36], whereas the second strategy (norm-driven exceptions) is appropriate when

[36] For the sake of this discussion, I am not considering more radical critiques which suggest that the categories I am calling *man* and *woman*, that is, the biological

Continues on next page…

we are talking about what I am calling *females*. That is, the proper conclusion to draw from the fact that someone can be a woman if she has only *b* and *d* is this: even in instances where the person in question also has *a* and *c* and *e*, the factors which *really* matter in making her a woman are *b* and *d*. By contrast, the proper conclusion to draw from the fact that someone can be a female is she has only *b* and *d* is *not* that *b* and *d* are all that matter even in ordinary cases. Rather, the proper conclusion to draw is that it is only because *b* and *d* are generally found in conjunction with *a*, *c* and *e* that we can—by courtesy as it were—include the *b*-*d* individual within the purview of the theory.

What I want to ask in this section is: is there anything systematic which can be said about what makes the first strategy appropriate in the first case, and the second strategy appropriate in the second case? My suggestion is that there are two tests which can be applied.

The first test employs the distinction made previously, between theories which attempt to enumerate necessary and sufficient conditions for being an X, and theories which can be properly said to characterize Xs by identifying characteristic instances of Xs and enumerating their features. In theories of the first sort, exceptional cases are appropriately accounted for by means of the first strategy; these are theories with exception-driven norms. In cases of the second sort, they are appropriately accounted for by means of the second strategy; these are theories with norm-driven exceptions.

categories by which classification is made throughout the animal kingdom, are

Continues on next page…

On reflection, this should not be surprising. A theory which characterizes Xs on the basis of necessary and sufficient conditions must account for *every* case which falls under its purview on the basis of these conditions. So if exceptional-X manifests only some of the characteristics associated with typical-X, it follows that even for typical-X, the only X-making characteristics can be those which are present in exceptional-X as well. By contrast, a theory which characterizes Ys on the basis of paradigm instances must have instances which are paradigmatic[37] for the theory to be a theory of anything at all. So if exceptional-Y manifests only some of the characteristics associated with typical-Y, it does *not* follow that the only Y-making characteristics are those which are present in both typical-Y and exceptional-Y. To the contrary, what follows is that even for exceptional-Y, the Y-making characteristics are those which are present in typical-Y.

The reason for this is brought out by the second test. The second test involves asking the following question: what would happen if all cases were like the exceptional case; would there still be Xs?[38] In theories with exception-driven norms,

themselves socially constructed.

[37] For the sake of simplicity, I am temporarily ignoring family-resemblance cases of the sort described above. (See footnote 31.)

[38] Cf. Kant: "I ought never to act except in such a way *that I can also will that my maxim should become a universal law*" (Kant 1785/1964, 18/70) As Kant notes, certain practices, if generalized, would undermine the possibility of their own existence. See, for instance, his discussion of lying in the *Groundwork**.

> * Kant writes: "I can by no means will a universal law of lying; for by such a law there could properly be no promises at all, since it would be futile to profess a will for future action to others who would not believe my profession, or who if they did so over-hastily, would pay me back in like coin; and

Continues on next page…

the answer is “yes;” in theories with norm-driven exceptions, the answer is “no.” So, for instance, to return to the example above: both women and females generally have characteristics *a b, c, d*, and *e*[39], and the exception we are considering is one where the person in question has only *b* and *d* (the biological characteristics), and lacks *a, c* and *e* (the social characteristics). We now ask: what if every case were like this one? And the answer is: there would still be women; but there would not be females[40].

In short, what we do when we characterize what it is to be a woman in the sense I am describing is to enumerate a set of necessary and sufficient characteristics. So even all cases were to manifest only the bare minimum of such characteristics, there would still be women. By contrast, what we do when we characterize what it is to be a female in the sense I am describing is to identify a set of attributes which are ordinarily correlated with one another. So if all cases were to manifest only the bare

consequently my maxim, as soon as it was made a universal law, would be bound to annul itself” (Kant 1785/1964, 29/71). Or again: “[T]he universality of a law that every one believing himself to be in need can make any promise he pleases with the intention not to keep it would make promising, and the very purpose of promising, itself impossible, since no one would believe he was being promised anything, but would laugh at utterances of this kind as empty shams” (Kant 1785/1964, 55/90).

[39] As before, these can be understood as follows: *a*: wears thus-and-such sorts of clothing; *b*: has thus-and-such chromosomal structure; *c*: occupies thus-and-such social role; *d*: has thus-and-such body parts; *e*: tends to respond to stressful situations in thus-and-such way.

[40] I may seem to be stacking the deck unfairly, in that I am taking *woman* to describe a biological category and *female* to describe a social category. One might say: surely if all persons had only *a* and *c*, there would not be women. My reply: Indeed—but if there were even *one* case of a person who had only *a* and *c* without *b* and *d*, that person would not be a woman. The question is: what would happen if all cases were

Continues on next page…

minimum of such characteristics, there would be no females. Or, more precisely, there would be no category "female" to which the persons who are currently classified as females would belong. With the breaking down of a pattern of correlated characteristics would come a breaking down of the concept which describes such a pattern.

1.1.4 Application to the Question at Hand

As the reader may recall, the reason I have been discussing exceptional cases is because they play a role in the central contention of the dissertation, which is that both the powers and the limits of the contemplation of imaginary scenarios can be traced to the fact that when such contemplation brings us to new knowledge, it does so by forcing us to make sense of an exceptional case. And I contended further that this diagnosis explains both the powers and the limits of this technique. The question that naturally arises is: why is this so? My answer is three-fold. First, thinking about exceptional cases can lead us to a reconfiguration of our conceptual commitments, allowing us to organize information in a way that renders it newly meaningful. Moreover, exceptional cases are good test cases; they help prevent us from mistaking accidental regularities for regularities which reflect a deeper truth about the world. But, third, exceptional cases are dangerous; if we fail to keep straight the distinction between theories with norm-driven exceptions and theories with exception-driven norms, we are likely to draw radically misguided conclusions.

like this exceptional case, where the exceptional case in question describes an entity which *falls under the purview of the theory*?

The last of the these claims, I think, explains the adage that hard cases make bad law. The reason hard cases make bad law (that is, bad normative law) is because laws aim to provide normative guidance in ordinary cases. But since hard cases are exceptional cases, taking them as paradigms will lead to a misfit between theory and everyday practice. If we are trying to determine the age at which children should enter kindergarten, we should not make our decision by looking at John Stuart Mill; if we are trying to decide penalties for theft, we should not focus on a case where a man steals medicine to save his dying wife. We can make sense of exceptional cases here, but only against a background of normalcy: John Stuart Mill may go to kindergarten, but if he does, he will be precocious; the medicine-stealer may be a thief, but if he is, he is a thief who faces extenuating circumstances.

It is the second of these claims that explains the (not-quite adage that) hard cases make great science[41]. The reason hard cases make great (descriptive) science[42] is because science aims to provide systematic theories[43] that explain not only all

[41] By "science" here, I mean things like plain old-fashioned descriptive physics, not (necessarily) things like classificatory biology or cultural anthropology.

[42] Of course, hard cases might be a terrible place to *start* if one is trying to develop a theory; this is a point about the epistemology of discovery and the capacities of human beings for pattern recognition. I am concerned with the status of hard cases as test cases; this is a point about the epistemology of justification and the standards which we assume scientific theories will meet.

[43]Cf. Aristotle: "In all disciplines in which there is systematic knowledge of things with principles, causes, or elements, it arises from a grasp of those: we think we have knowledge of a thing when we have found its primary causes and principles, and followed it back to its elements" (Aristotle, *Physics* 184^{a}11-15.)

actual, but also all physically possible phenomena within their purview[44]. If a mere inventory of what there is were the sole goal of scientific theories, naïve regularity theories[45] would capture what it is that we mean by physical law, but they do not[46]. Even if there is not enough radium on earth to fill a shoe-box the size of Alaska, we still want our theory to tell us how much radioactivity such an amount of radium would have. And if a theory of planetary motion fails to predict the motion of a new heavenly body that appears on the scene, so much the worse for the theory.

[44] Of course, a scientific theory might respond to apparent counterexamples by refining the range of cases which it purports to cover.

[45] A naïve regularity theory is a theory which picks out as laws of nature those statements which are (a) universally quantified; (b) true; (c) contingent; and (d) contain only logical connectives, logical quantifiers, and non-local empirical predicates. (See Armstrong 1983, 12; the characterization in question comes from Molnar 1969.) Following Armstrong, we might call the uniformities picked out by such a characterization *Humean uniformities* or *Humean regularities*.

[46] In identifying Humean uniformities with laws of nature, naïve regularity theories face a cluster of problems. (The following discussion is borrowed from Armstrong 1983, 12-13.) First, they face *extensional difficulties* in that they pick out uniformities which do not seem to be laws of nature, so they take as sufficient for being a law of nature a characteristic (being a Humean uniformity) which is in fact not sufficient. Second, they face *oversimplification difficulties* in that they identify the content of laws with the content of uniformities, whereas there are cases where it seems that a gap exists between a law and the manifestation of that law, that is, between the content of a law and the content of the uniformity which is its manifestation. Finally, they face *intensional difficulties,* in that even if it is assumed that the content of the law and the content of the uniformity are identical, it might still be that the law has properties which the manifestation lacks. For detailed arguments establishing the validity of these criticisms, see Armstrong 1983, pp. 1-73. Regardless of what one thinks of the positive characterization he offers of his own theory in the second part of the book, his negative arguments against naïve regularity theories are, I think, quite convincing.

Whether hard cases make good or bad philosophy is a much more difficult question. Like normative legal theories, philosophical theories are often concerned with providing normative guidance in ordinary cases. But like scientific theories (indeed, even more than scientific theories) philosophical theories aim to provide us with accounts not only of how things are, but also of how things can and cannot be[47]. This might seem to suggest a simple division: philosophical theories that are normative are likely to have difficulties with exceptional cases, whereas those that are descriptive are likely not to. Unfortunately, the distinction is not so simple.

1.2 Imaginary Cases

Because the world is neither maximally replete nor effortlessly navigable, *all* legal and scientific and philosophical investigation inevitably involves the contemplation of cases that are *imaginary*. Law school textbooks are brimming with hypothetical cases which are the central diet of legal education. Elementary physics texts are replete with postulated examples which ask the reader to imagine performing certain experiments, and to predict their outcomes. And elementary philosophy texts overflow with passages like this one, taken from the opening pages of Jonathan Dancy's introductory text on epistemology:

> [In thinking about what we mean by "knowledge"] we do not need to rely on actual mistakes in the past. For our purposes, possible mistakes will do just as well. This can be seen in the ethical case. An imaginary example can be so

[47] "Experience tells us…what is, but not that it must necessarily be so, and not otherwise. It therefore gives us no true universality; and reason, which is so insistent upon this kind of knowledge, is therefore more stimulated by it than satisfied" (Kant 1787/1929, B2/A2)

> described that I am willing to say that the action it recounts is good. And that judgment of mine is as much universalizable, as binding on my future judgments about relevantly similar cases, as if the example had been real rather than imaginary. Similarly, an imaginary case in which I would claim to know that *p*, but where *p* is false, will succeed in preventing me from claiming to know that *p* in a new case which is not relevantly (i.e., discernibly to me) different. So imaginary cases are as effective in the argument as actual ones. (Dancy 1985, 13-14).

That is, Dancy suggests that when an imaginary case is indiscernibly different from an actual case, our responses to that case are as relevant as our responses to an actual case would be[48].

I suggest the following spelling-out of Dancy's criterion: if, for all we know, the case presented to us might be actual, then our responses to that case are to be treated as we would treat responses to an actual case[49]. So, for instance, there is little if any difference between deciding whether an actual child raised as a Jehovah's witness should be excused from the classroom during the pledge of allegiance, and deciding whether, if there were a child raised as a Jehovah's Witness in the class, that child should be excused from the classroom during the pledge of allegiance. This is true of any realistic normative case, where the situation described may, for all we know, be actual, and the judgment in question is about what should be done or said.

[48] It is interesting to note that Dancy is one of the main contemporary opponents of the technique of thought experiment in ethics. (See Dancy 1985b, as well as his discussion of "switching arguments" in Dancy 1993, 64-66.)

[49] But cf. Nozick: "[I]f principles are only supposed to cover the cases that will, would, and could arise, then before the fact, if it is thought such a case is impossible (that the situation, motivations, or whatever that would lead to it could not arise or succeed), it might not be considered a *relevant* counterexample to that or any principle. But once it is discovered that [such and such] *can* [occur]—because it *did*—then the principle *P* that [failed to account for it] is refuted" (Nozick 1993, 38).

In such cases, whatever it is that explains the informativeness of thinking about actual situations in order to decide the proper application of principles or concepts explains the informativeness of thinking about imaginary situations.

Of course, *thinking* about a situation—actual or hypothetical—is not the same as *confronting* a situation, and it may be that there are issues raised by *reasoning* as such as a way of determining answers to normative questions. Moving the case from the world to the mind "cleans it up" in certain ways, and allows it to be isolated from the network of events of which it is a part. But this is equally true of cases that are drawn from experience and of cases that are merely described. Still, it might seem—even for realistically described normative cases—that the difference between an imaginary and an actual situation about which we are called to make a judgment is that the latter carries with it a certain sort of *urgency*: some state of affairs that is actually out there, and we need to make a decision about how it is properly to be evaluated. An imaginary situation, by contrast, is purely academic: failure to make a judgment about it, or to make a proper judgment about it, might leave us with less knowledge in some scientific sense, but it would not—at least not directly—leave us less able to negotiate the world as it actually is

Although this is a difference, I do not think it is a deep one—certainly not deep enough to drive a wedge between the *class* of cases that are actual and the cases of cases that, though for all we know might be actual, as a matter of fact are not. In the first place, judgments about imaginary cases may help us in making judgments about actual cases, and this possibility may impart an instrumental urgency to decision-making about imaginary cases. Secondly, even if, for two comparable cases, the actual case carries with it an urgency that the imaginary case does not, there may be other imaginary cases that carry more urgency than either of the two. Third, in many instances we do not *know* (or even have an opinion about) whether a case is actual or imaginary, and this does not seem to affect our ability to make judgments about the case, or the conviction with which we make them.

Moving to realistic descriptive cases, where the question asked is ‘what would happen in the following circumstances?’ the distinction between actual and merely imaginary cases again seems minimal. Answering the question ‘what would happen if I were to release this actual ball down this actual plane?’ and answering the question ‘what would happen if I were to release some ball with mass x down some plane with slope y?’ merit exactly the same reasoning process. It might seem that all sorts of things could get in the way of the accuracy of the first answer: the ball might be chipped, or the plane might be warped, or a breeze might be blowing. But this would be to change the case in question; so long as the circumstances described are the same, the reasoning process about the two cases will be as well.

But perhaps this is the salient difference: the imaginary case allows *idealization*, whereas the actual case does not. Again, if the issue is that the imaginary case describes an ideal case, whereas the actual case does not, then the cases are not the same; if the imaginary case involves a perfectly smooth plane, whereas the actual case involves one that is only approximately smooth, then the two cases differ, and we are back to our previous failed attempt to distinguish imaginary cases from actual ones[50]. However, we now have the suggestion of a systematic difference, namely, that imaginary cases have the *possibility* of being described in idealized ways, whereas actual cases do not. However, to the extent that *this* is what is at issue, then the difference between the cases is not that the one is, as a matter of coincidental fact, actual whereas the other is, as a matter of coincidental fact, imaginary, but rather that the case in question is *essentially* imaginary: we know *without looking at the world* that no actual state of affairs corresponds to the case described by the scenario.

One way to characterize this distinction is as follows. The imaginary scenario may be either *hypothetical* or *counterfactual.* By "hypothetical" I mean: for all we know, such a situation might indeed obtain; whereas by "counterfactual" I mean: without looking at the world, we know the situation is imaginary[51]. Then we can say that, to the extent that the contemplation of imaginary exceptional cases raises philosophical issues that the contemplation of *actual* exceptional cases does not, this is a consequence of the cases being, in the sense that I have just described, counterfactual[52].

[50] That is: the question we are concerned with is: what is the difference between a case which is imaginary and a case which is actual, where the cases have *exactly the same* characteristics.

[51] Of course this line is not clean. However, one might think of the distinction as follows: roughly speaking, it is the fact that the world is not effortlessly navigable that leads us to consider cases that are *hypothetical*, and the fact that it is not maximally replete that leads us to consider cases that are *counterfactual*.

[52] I am using these terms in a rather idiosyncratic way. A different distinction, though one not unrelated to the one I have been discussing, can be found in the works of Onora O'Neill. O'Neill suggests that when we are discussing the role of (imaginary) examples in ethics, a distinction can be drawn between examples which are

Continues on next page…

1.3 Thought Experiments

1.3.1 What Is a Thought Experiment?

It may seem remarkable that I have gotten this far in the introduction without mention of thought experiments. The omission is deliberate. The term "thought experiment" has broad and disputed application[53], and I wish as little of my argument as possible to depend upon one or another delineation of this expression. Nonetheless, I think any characterization of the term which aims to do justice to the way in which the expression is used will have to account for at least the following six cases, which I take to be paradigmatic instances of its proper usage. (For those unfamiliar with them, I provide brief summaries of the cases in the endnotes to this chapter):

- Plato's story of the Ring of Gyges, in which Glaucon tries to bring his listener to see that "those who practice justice do it unwillingly and because they lack the power to do injustice" (Plato c.380BC/1974, 359b)[(i)].

- Stevinus's discussion of the inclined plane, in which he seeks to establish that the force required to hold a ball in place along an inclined plane is inversely proportional to the length of the plane[(ii)].

hypothetical : "they consist of more or less specific principles of possible action (e.g. the principle of cheating a gullible customer...) whose moral significance can be determined...by applying the Categorical Imperative; [and] examples [which] are *ostensive*: they point out acts or persons or lives some of whose features are held to be morally significant (e.g. taking the life of Christ as a model for imitation)" (O'Neill 1986, 7). She continues: "hypothetical examples, being themselves principles of action, must evidently remain indeterminate, even when relatively specific, and so cannot fully determine an act. The acts or persons or lives which are pointed to in ostensive examples may, in themselves, be fully determinate. But their relevance to a case in hand must...be guided by some...understanding of the morally significant aspects of the example...*Judgment* is therefore always needed...Neither principles nor examples alone can guide action" (O'Neill 1986, 7-8).

[53] For a collection of recent attempts to define "thought experiment," see Appendix B.

- Einstein’s discussions of moving trains, in which he seeks to establish that a commitment to the view that the speed of light is constant brings with it a commitment to the relativity of simultaneity.

- Putnam’s discussion of Twin Earth, in which he seeks to establish that meaning is fixed externally (Putnam 1975a, 223-227)[(iii)].

- Nozick’s discussion of the Experience Machine, in which he seeks to establish that we care about more than hedonistic experience (Nozick 1974, 42-45)[(iv)].

- Thomson’s discussion of the ailing violinist, in which she seeks to establish that abortion may be morally justifiable, even if the fetus is a person (Thomson 1971)[(v)].

This characterization does run counter to the original use of the term, which reserved its use for the contemplation of imaginary cases eliciting physical intuitions[54].

George Bealer writes:

> In recent philosophy there has been an unfortunate blurring of traditional terminology. Rational intuitions about hypothetical cases are often being erroneously termed *thought experiments*. This deviates from traditional use, and blurs an important distinction which should be kept vividly in mind. Traditionally, in a thought experiment one usually elicits a physical intuition (not a rational intuition) about what would happen in a hypothetical situation in which physical, or natural, laws (whatever they happen to be) are held constant but physical conditions are in various other respects nonactual and often highly idealized…A classic example is Newton’s thought experiment about a rotating bucket in an otherwise empty space. Would water creep up the side of the bucket (assuming that the physical laws remained unchanged)? Rational intuition is silent about this sort of question. Rational intuitions concern such matters as whether a case is possible (logically or metaphysically), and about whether a concept applies to such cases…[T]o call [such cases] thought experiments is not only to invite confusion about philosophical method, but to destroy the utility of a once useful term.” (Bealer 1996, 9-10).

[54] The term *Gedankenexperiment* appears to have been coined by Ernst Mach in 1883; his paradigmatic example of a thought experiment is the case from Stevinus (described above). For Mach’s use of the term, see Mach 1933/1960 and Mach 1926/1976.

I think Bealer is right to point out that there is an important distinction to be drawn between rational and physical intuitions, and that there is an accompanying distinction to be made between the sorts of thought experiments conducted in pursuit of philosophical knowledge and those conducted in pursuit of scientific knowledge. Indeed, section 1.3.3 is devoted to an attempt to classify thought experiments along precisely these grounds. Nonetheless, I do not think it misleading to use the term to describe the contemplation of imaginary cases more generally. For I think there are certain structural features that hold in common across all thought experiments.

1.3.2 The Tri-partite Structure of Thought Experiments

For each of the examples above, the following characterization captures something important about its fundamental structure:

(1) An imaginary scenario is described .

(2) An argument is offered which attempts to establish the correct evaluation of the scenario.

(3) This evaluation of the imagined scenario is then taken to reveal something about cases beyond the scenario.

So, for example, in the Experience Machine example, the imaginary scenario is the existence of the machine as described, the evaluation of the scenario is that we would consider hooking up to such a machine to be undesirable, and the larger lesson is that “we learn that something matters to us in addition to experience by imagining an experience machine and then realizing that we would not use it” (Nozick 1974, 44). In the Twin Earth example, the imaginary scenario is the existence of the planet on which something qualitatively identical to water has the chemical structure *XYZ*, the evaluation of the scenario is that speakers of English and Twin-English refer to something different by their use of ‘water,’ and the larger lesson is that “‘meanings’ just ain’t in the *head*” (Putnam 1975a, 227). In the Stevinus example, the imaginary scenario is the presence of the balls along the inclined plane connected by the chain at the bottom, the evaluation of the scenario is that the balls will remain in equilibrium, and the larger lesson is that the force necessary to hold a ball in place along an inclined plane varies inversely with the length of the plane. That parallel analyses could be offered for the remaining examples is, I trust, clear[55].

Although this positive taxonomy is not especially interesting (its breadth of application comes in part from the imprecision of its categories) its negative counterpart is surprisingly useful. Corresponding to (1), (2) and (3) above are three criticisms:

[55] That characteristic (3) is the intention of thought experiment is strikingly apparent in the following quote from Strawson:

> In general, this form of question [that is, questions asked about wildly imaginary cases] may be seen as simply a convenient, if perhaps over-dramatic, way of raising more evidently legitimate types of question: questions not about hypothetical beings at all, but rather, for instance, about the extent to which, and the ways in which, *we* might find it possible to reinterpret, within a part of our experience, some of the most general conceptual elements in our handling of experience as a whole…My real concern is with our own scheme, and the models of this chapter are not constructed for the purpose of speculation about what would really happen in certain remote contingencies. Their object is different. They are models against which to test and strengthen our own reflective understanding of our own conceptual structure. Thus we may *suppose* such and such conditions; we may discuss what conceptual possibilities and requirements they can be seen by us as creating...In all this we need no more claim to be supposing real possibilities than one who, in stricter spheres of reasoning, supposes something self-contradictory and argues validly from it (Strawson 1959, 86).

> (1') Unimaginability: The scenario described is not (fully) imaginable[56].
>
> (2') Unsound Argument: The scenario described is imaginable, but the argument establishing the correct evaluation of the scenario is unsound[57]
>
> (3') Inapplicability: The scenario described is imaginable, and the argument establishing the correct evaluation of the scenario is sound, but the conclusion does not reveal about the actual world what it is that the author takes it to reveal.

These criticisms provide an extraordinarily fruitful way of classifying criticisms of thought experiments which amount to the objection that the scenario we are being asked to consider is *just too far-fetched*. Those who make unimaginability objections question the very coherence of the exceptional case described[58]. Those who make unsound argument objections question our ability to use normal modes of reasoning in thinking about the exceptional case[59]. And those who make inapplicability objections question the extent to which conclusions drawn on the basis of the

[56] This may be for one of two reasons: the scenario might be *incoherent*, or it might by *underdescribed*. If the latter, it might be either *resolvably* or *irresolvably* underdescribed. I discuss the distinctions among these sub-criticisms in Appendix C, in sections **C.2.1** and **C.2.2**.

[57] Again, this might be for one of two reasons. The argument might be *independently unsound* or it might be *situationally unsound*. An argument is independently unsound if it is unsound for reasons having nothing to do with the thought-experimental scenario. An argument is situationally unsound if its use depends on appeal to a principle that cannot justifiably be employed in the way that the author wishes because something about the imaginary scenario implies that the ordinary criteria for application of a particular concept are unavailable or the ordinary justification for appeal to a particular principle is missing. (The criticisms I offer in chapter 4 are criticisms of this sort.)

[58] That is, they contend that the case is either incoherent, or irresolvably underdescribed.

[59] That is, they contend that the argument is situationally unsound.

exceptional case can be applied (in the way that the author wishes) to normal cases as well. In all three cases, the objection is perhaps best understood as an expression of the legitimate concern that our beliefs, desires, and concepts are deeply tied to out views about which alternate possibilities are salient, so that the imagined disruptions of these patterns of saliency will leave us with too little to base our judgments on—either because we cannot fully imagine such a scenario, or because we would not know how to make sense of it, or because we would not know how to apply its lessons to our world[60].

So, for instance, someone might object that the idea of a machine which would perfectly simulate experiences is just too far-fetched for us to make sense of, and by that, they might mean one of three things:

[60] As a test of the fruitfulness of this analysis, consider the sense it makes of the following oft-quoted passage from Kathleen Wilkes. (The passage is cited in Martin 1991 and Oderberg 1993, among others; reference is made to similar passages in Baillie 1993, Hertzberg 1991, Madell 1991 and Rovane 1993) She writes that in thinking about wildly imaginary cases (what I would call clearly counterfactual cases):

> either (a) we picture them against the world as we know it, or (b) we picture them against some quite different background. If we choose the first, then we picture them against a background that deems them impossible…If we choose (b), then we have the realm of fantasy, and fantasy is fine to read; but it does not allow for philosophical conclusions to be drawn, because in a world determinately different we do not know what we would want to say about anything (Wilkes 1988, 46).

I would analyze her argument as follows:

> (2') *situationally unsound:* "[if] (a) we picture them against the world as we know it…then we picture them against a background that deems them impossible"
>
> (3) *inapplicable:* "or (b) we picture them against some quite different background… then we have the realm of fantasy, and fantasy is fine to read; but it does not allow for philosophical conclusions to be drawn"

Continues on next page…

> (1') Unimaginability: That we cannot really imagine there being such a machine.
>
> (2') Unsound Argument: That although we can imagine there being such a machine, if we try to reason about how we would react to the possibility of being hooked up to it, we will inevitably make judgments on the basis of assumptions which would not be relevant under such circumstances.
>
> (3') Inapplicability: That although we can imagine there being such a machine, and we can make informative judgments about what it would make sense to do and say if we were confronted with the possibility of being hooked up to it, the conclusions which can be drawn on the basis of those judgments do not establish the claim that, in general and in actual cases, "something matters to us in addition to experience."

This sort of objection is most powerful when it is made against thought experiments which are, in the sense described at the end of section 1.2, counterfactual. But even in the case of hypothetical thought experiments, objections of at least the latter two kinds can be made[61]. One might (albeit implausibly) object to Stevinus that his evaluation of the scenario he has described is incorrect, or that the case he has described is a special one, from which general conclusions cannot be drawn. Nonetheless, this asymmetry points to a certain distinction among types of thought experiments which this characterization does not capture, and which I describe in the next section.

1.3.3 Three Sorts of Thought Experiments

> (1') *irresolvably underdescribed:* "in a world determinately different we do not know what we would want to say about anything"

[61] See the detailed discussion of the Einstein-Bohr debate concerning Einstein's clock-in-a-box thought experiment in Norton 1991, 139-142. See also Allen I. Janis, "Can Thought Experiments Fail?" (Janis 1991).

Continues on next page…

In the last section, I stressed certain features which I think are common to all thought experiments. These features make it reasonable to describe thought experiment as a mode of reasoning. But there are crucial contrasts to be drawn among types of thought experiment, and in this section, I describe one such difference. What I want to suggest is that there are three basic sorts of questions which might be asked about an imaginary scenario:

(1) What would happen?

(2) How, given (1), should we describe what would happen?

(3) How, given (2), should we evaluate what would happen?

We might call the first type *factive* (concerning what we think the facts of a situation would be), the second *conceptual* (concerning what we take to be the proper application of concepts), the third *valuational* (concerning the proper moral or aesthetic response to a situation). These three questions correspond—*very roughly*—to the sorts of issues addressed by thought experiments in (1) science; (2) metaphysics and epistemology; and (3) ethics and aesthetics. Of course, this is only a generalization; scientific thought experiments may ask how something ought to be described, epistemological thought experiments may ask what ought to be, and ethical thought experiments may ask what would happen. But although it relies on an exaggeration (indeed, my argument for how it is that imagination can lead to new knowledge relies on showing that the line between the first and the second question rests on a false dichotomy) I think it is illuminating rather than misleading to speak in this case of the distinction among what is, what is said to be, and what ought to be. So, for instance, we have examples above of each of the three sorts of cases:

(1) Factive: What would happen?
E.g: *Einstein, Stevinus*

(2) Conceptual: How, given (1), should we describe what would happen?
E.g: *Gyges, Twin Earth*

(3) Valuational: How, given (2), should we evaluate what would happen?
E.g: *Experience Machine, Thomson's Violinist*

Each sort of case raises distinct philosophical puzzles.

In the case of factive thought experiments, the issue is: how is it that thinking about something in a new way can lead us to recognize something new about the physical world? (Consider, for example, Stevinus realizing that we already had enough information to figure out what was going on in the case of the inclined plane.) Before we start thinking about the scenario, we don't know what aspects of the situation are going to be relevant. Somehow, however, by imagining something *particular*, we get richness sufficient to answer the question. But why does *that* help? Obviously we are giving ourselves the image, so in some sense we must already *know* the thing we are trying to find out[62]. (Where else *could* the information be coming from, if not from the image we are contemplating?) So the puzzle is: what is enabling us to see this imagined particular as something about which we can make a judgment if we think hard enough? This is the problem which I address in chapter 2, where I discuss a famous thought experiment of Galileo's.

[62] No one expects to get rich by paying their right hand with their left! Why then do we expect to gain knowledge by such a technique?

For the second and third types, the issue is: what do we expect to learn about our concepts or values by trying to make sense of this imagined case? Why should thinking about a case that has not occurred or is not going to occur help us understand how we (should) evaluate actual cases? The reasons that we think about imaginary cases are the reasons I discussed in section **1.2**: not every way that things might be is a way things are, and even when things are some way, it may not be so easy for us to reach them and manipulate them. At the same time, one of the ways we are able to learn which features of a particular circumstance are relevant to our evaluation of it is by Mill's methods of comparison, varying the factors which contribute to a situation to see which of them plays which role.

In scientific cases, when we explore physical dependencies, we do (actual) physical experiments. In philosophical cases, when we explore conceptual dependencies, we do thought experiments. In both, we test various hypotheses by considering cases in which we systematically vary the possible contributing factors. And in actual scientific experiments, the circumstances which we modify are circumstances in the world, and the knowledge we gain is empirical knowledge. But in philosophical thought experiments, the circumstances we modify are circumstances which we ourselves have contrived: we conjure up some situation, conjure up modifications of the situation in light of what we wish we knew about the world, and suddenly declare ourselves to have learned something new. But unlike the factive cases, in conceptual and valuational cases we already know *what* would happen; we are worried here about how we should *describe* or *evaluate* it. So the puzzle here is: what could possibly be guiding us in making such a judgment? This is the problem which I address in chapters 3 and 4, in my discussions of the Ship of Theseus (chapter 3) and of personal identity (chapter 4).

And with these preliminaries in place, it is time to begin.

Endnotes to Chapter 1

[i] For a discussion of this example, see Appendix A

[ii] Stevinus considered the question of how much force would be necessary to prevent a ball from sliding down an inclined plane. He imagined a triangular prism on top of which is laid a circular string of fourteen balls:

Since assuming otherwise would commit us to the position that perpetual motion is possible, Stevinus concludes that the balls are in a state of equilibrium, that is, that the chain moves neither to the right nor to the left. He next imagines cutting the string at the two lower corners, such that two balls remain along the side with the sharper incline, and four along the side with the shallower incline. Since the balls were in equilibrium prior to the cutting, they remain so afterwards; so the shorter and the longer string of balls are in balance. From this Stevinus concludes that the force required to hold a ball in place along an inclined plane is inversely proportional to the length of the plane.

[iii] Putnam asks his reader to imagine a planet, Twin Earth, which is "*exactly* like Earth" except that on Twin Earth, "the liquid called 'water' is not H_2O but a different chemical whose chemical formula…[will be abbreviated] simply as *XYZ*." According to the story, "*XYZ* is indistinguishable from water at normal temperatures and pressures," so that when speakers on Twin Earth refer to "water," they refer to something that gives them exactly the same sorts of experiences (in terms of taste, quenching-of-thirst, wetness, cleansing power, etc.) that water (that is, H_2O) gives t speakers who use the term on Earth. So even though the speakers of English and the speakers of Twin-English use the term "water" in exactly the same circumstances, and even though they may be in exactly the same psychological states as one another when they do so, the extension of the term (that is, what it is that the term picks out in the world) differs in the two cases; "the extension of 'water' in the sense of water$_E$ is the set of all wholes consisting of H_2O molecules, or something like that; the extension of water$_{TE}$ is the set of all wholes consisting of *XYZ* molecules, or something like that." On the basis of this and other examples, Putnam takes himself to have established that "'meanings' just ain't in the *head*" (Putnam 1975a, 223-227).

(iv) Nozick asks his reader to imagine a machine which allows “superduper neurologists” to “stimulate your brain so that you would think and feel that you were writing a great novel, or making a friend, or reading an interesting book” where really “ [a]ll the time you would be floating in a tank, with electrodes attached to your brain.” Importantly, *from the inside* there would be no way to tell that you were ‘hooked-up’: “while in the tank you won’t know that you’re there; you’ll think it’s all actually happening.” Nozick contends that we would consider hooking up to a machine to be undesirable, concluding: “[w]e learn that something matters to us in addition to experience by imagining an experience machine and then realizing that we would not use it.” (See Nozick 1974, 42-45)

(v) Thomson tells a story in which while you were asleep, a Society of Music Lovers has connected you to an ailing violinist, so that, for the next nine months, your kidneys will process the fluids he will need for survival. Thomson argues that even if disconnecting him will result in your death, you might still be morally justified in doing so, and she suggests that this, along with other cases, shows that even if we assume that the fetus is a person, abortion might be morally justifiable. (See Thomson 1971)

2. Galileo

In this chapter, I discuss one of the most famous thought experiments in the history of science. Through a careful examination of this *factive* thought experiment, I try to show that the guided contemplation of an imaginary scenario can provide us with new scientific knowledge in a way that argument alone cannot.

2.1 Argumentative Reconstruction

It has been widely argued that thought experiments in science are eliminable, and that whatever demonstrative force they have is the result of their being sound arguments dressed up in heuristically appealing clothing. According to this view, a scientific thought experiment's justificatory force comes from the fact that it can be reconstructed as an argument with explicit premises that make no reference to imaginary particulars. In this section (**2.1**), I present this thesis, describe the philosophical rationale behind it, and then distinguish between a weaker and a stronger version of what is being claimed.

2.1.1 The Elimination Thesis

The view that thought experiments lead to justified conclusions because they are arguments finds clear articulation and powerful defense in a recent paper of John Norton's entitled "Thought Experiments in Einstein's Work." In the paper, Norton puts forth a hypothesis about thought experiments that he calls the *elimination thesis*. With a few carefully placed ellipses, Norton's thesis is this: "any conclusion reached by a good thought experiment will also be demonstrable by an argument which...is not a thought experiment" (Norton 1991, 131) [63]. What Norton means by "an argument which is not a thought experiment" is explained by the following. Norton thinks thought experiments *are* arguments, but arguments with a few distinct characteristics; namely, they "are arguments which posit hypothetical or counterfactual states of affairs, and which invoke particulars irrelevant to the generality of the conclusion" (Norton 1991, 129). An "argument which is not a thought experiment" is an argument which does *not* do these things. So what the elimination thesis says is that any good scientific thought experiment can be transformed into a non-thought-experimental argument in such a way that the demonstrative force of the thought experiment is not lost.

Below I will suggest that as initially formulated, the thesis is ambiguous; it is compatible with both a weaker reading, which I will call the *dispensability thesis*, and a stronger reading, which I will call the *derivativity thesis*. But in order to formulate these versions, I first need to clarify what the terms in the thesis mean, and how such a position might be justified.

2.1.2 Clarification of Terminology

[63] I should point out that my reconstruction of Norton's quite subtle position (developed primarily to account for thought experiments in twentieth-century physics) is somewhat free, and that he is fully cognizant of the extent to which such eliminations may be difficult in practice. The passage quoted above continues: "Whilst thought experiments may be eliminable in principle, it does not follow that this elimination will be easy. In fact, one can almost guarantee that it will not. For thought experiments are usually introduced when the straight argument would be difficult to develop." (Norton 1991, 131). As will become apparent below, my main

Continues on next page…

How would one go about transforming a thought experiment in keeping with the elimination thesis? Since thought experiments are generally presented in the form of narratives or of invitations to the reader to imagine a particular scenario, the first step of the elimination will generally be to transform this apparently non-argumentative passage into an argument that freely makes reference to hypotheticals, counterfactuals, and particulars of whatever sort[64]. Thereby the unregimented thought experiment will be transformed into a thought-experimental argument. The second step of the elimination will be to remove from this argumentative reconstruction all reference to non-actual states of affairs, and all reference to particulars not mentioned in the conclusion. Thereby the thought-experimental argument will be transformed into an argument that is not thought-experimental. So what an elimination will involve is first a process of argumentative reconstruction in which the narrative presentation is replaced by a series of explicit premises sufficient to establish the desired result, and then a process in which those premises that make reference to hypotheticals, counterfactuals, and particulars are replaced by premises in which no such reference is made. If the elimination thesis is correct, such a process will preserve completely the thought experiment's demonstrative force.

What is meant by "demonstrative force"? I will suggest two problematic readings and then one that I will endorse. If the claim that "any conclusion reached by a good thought experiment will also be demonstrable by an argument which...is not a

dispute with Norton concerns not the effectiveness of thought experiments, but the issue of what explains their justificatory force.

thought experiment" means no more than that in the reconstruction of a mature science, the conclusions that were (as a matter of fact) reached by thought experiments can be derived from more fundamental principles by means of inference schemes licensed within the science, then the thesis is trivially true. Even if the development of Newtonian mechanics relied on a series of crucial thought experiments, its textbook presentation might well establish particular conclusions on the basis of more conventional forms of argument. At the same time, there is a reading of "demonstrative force" according to which the thesis is trivially false. If the claim is taken to mean that—as a matter of psychological fact—any conclusion which was reached by a good thought experiment might also have been demonstrated to the person who reached the conclusion by means of a non-thought-experimental argument, then the elimination thesis is certainly false. None would doubt the important heuristic and illustrative role played by thought experiments in scientific exploration, and the crucial tasks they play in instruction and persuasion.

The proper reading of "demonstrative force" makes the elimination thesis epistemologically interesting. On this reading, demonstrative force concerns the role that thought experiments play in living bodies of knowledge: after the moment of discovery and before the end of inquiry. It concerns whether a particular conclusion based on a particular process of reasoning (thought experiment) is thereby justified—whether if such a process leads to true beliefs, those beliefs should count as knowledge. So the issue raised by the elimination thesis is this: can reasoning about

[64] Note that I am here using "hypothetical" and "counterfactual" in their standard

Continues on next page…

(reasonably) specific entities within the context of an imaginary scenario lead to rationally justified conclusions that—given the same initial information—would not be rationally justifiable on the basis of a straightforward argument?

2.1.3 The Negative Argument and the Positive Argument

The elimination thesis is defended with two arguments, one negative, one positive. Norton's version of the negative argument runs as follows: thought experiments must be arguments because there is nothing else for them to be. "Thought experiments in physics provide or purport to provide us information about the physical world. Since they are *thought* experiments rather than *physical* experiments, this information does not come from the reporting of new physical data. Thus there is only one non-controversial source from which the information can come: it is elicited from information we already have by an identifiable argument" (Norton 1991, 129). Norton considers this position almost trivial: "the alternative," he writes, "is to suppose that thought experiments provide some new and even mysterious route to knowledge of the physical world" (Norton 1991, 129). That is, the negative argument contends that if we have obtained new information about the empirical world without having obtained new *empirical* information about the empirical world, the only way we *could* have done so is by means of an argument.

The positive reason Norton thinks that thought experiments are just arguments in disguise is this: "the *analysis* and *appraisal* of a thought experiment will involve

senses, and not in the more special sense introduced in section **1.2** above.

reconstructing it explicitly as an argument," so that "a good thought experiment is a good argument, a bad thought experiment is a bad argument" (Norton 1991, 131). That is, the positive argument amounts to saying that if a thought experiment can be *reconstructed* as an argument, then what it was all along *was* an argument. So even if the reason I come to know something by contemplating a thought-experimental scenario doesn't *seem* to be because there is an argument into which the thought experiment can be reconstructed, it is. The reason my belief is *justified* is because, in the end, the thought experiment *was* a disguised argument all along.

2.1.4 The Dispensability Thesis and the Derivativity Thesis

We are now in a position to recognize that the elimination thesis as originally formulated and defended actually involves two distinct claims. These might be stated as follows:

> **The Dispensability Thesis**: Any good scientific thought experiment can be replaced, without loss of demonstrative force, by a non-thought-experimental argument.
>
> **The Derivativity Thesis**: The justificatory force of any good scientific thought experiment can only be explained by the fact that it can be replaced, without loss of demonstrative force, by a non-thought-experimental argument.

Loosely put, the dispensability thesis says that we can always get from here to there without appeal to a thought experiment. If a thought experiment legitimately transports us from one state of belief to another, a non-thought-experimental argument could too. Thought experiments may be convenient and efficient ways of reaching conclusions about the physical world, but they have only the advantage that a car has over walking; they get us where we want to go much more quickly, but they don't get us anywhere we couldn't reach by more pedestrian means.

The derivativity thesis says that not only can any good scientific thought experiment be replaced, without loss of demonstrative force, by a non-thought-experimental argument, but that to the extent that a good scientific thought experiment has demonstrative force, it is *because*, deep down, the thought experiment *is* an argument. We may be misled by the surface features of the case to think that something non-argumentational is doing justificatory work, but we are wrong. The *reason* the dispensability thesis is true is that all that was *ever* justificatorally at play was something argumentative. What looked like a car turned out to be propelled by foot power all along (like a child's go-car, or one of the vehicles on *The Flintstones*). So the dispensability thesis says we *can* get by without what we commonly call thought experiments; the derivativity thesis tells us that we already *do*.

In the next section, I challenge the dispensability thesis by showing that it does not hold true of a widely-acclaimed thought experiment of Galileo's. I do so for two reasons. First, since this particular example is generally treated as the paradigm of an effective thought experiment; diagnosing the source of its success is in itself a worthwhile endeavor. Second, challenging the dispensability thesis in this way allows me to shed light on the derivativity thesis as well. Obviously, if the dispensability thesis is false, the derivativity thesis is too; the more interesting question is whether some alternative explanation can be offered of the thought experiment's success. I try to say something positive about this question in section **2.4**.

2.2 Galileo's Thought Experiment and its Reconstruction

2.2.1 Galileo's Thought Experiment

Perhaps the most famous thought experiment in the history of western science is the thought experiment with which Galileo is credited with having refuted the Aristotelian view that the speed with which a body falls is directly proportional to its weight[65]. The thought experiment appears in his last and most mature work, the *Discourse Concerning Two New Sciences*[66], in the context of a more general discussion of the possibility and nature of motion in a void. Galileo's goal in the section as a whole is to establish that "if one were to remove entirely the resistance of the medium, all materials would descend with equal speed" (Galileo 1638/1914, 116); the thought experiment in question leads to the weaker conclusion that "both great and small bodies, *of the same* [*material*], are moved with like speeds" (Galileo 1638/1914, 109, italics added).

The view that Galileo is challenging is that "movables differing in heaviness are moved in the same medium with unequal speeds, which maintain to one another

[65] Challenges to the Aristotelian thesis—both empirical and conceptual—had appeared in a number of mid- and late-sixteenth century works. (For relevant passages, see [Cardan] Cooper 1934, 74-77; Damerow et al. 1992, 365; [Tartaglia] Drake and Drabkin 1969, 63-143, esp. 120ff; Damerow et al. 1992, 378; [Benedetti] Drake and Drabkin 1969, 147-237, esp. 206; Drake and Drabkin 1969, 31-41; Dijksterhuis 1961, 269-271; Drake 1989, 27-30; [Stevin] Cooper 1934, 77-80; Dijksterhuis 1961, 324-329.) Galileo himself had produced a less conclusive version of the famous thought experiment as early as 1590 in an unfinished and unpublished dialogue on motion; see Cooper 1934, 80-90; Drake and Drabkin 1969, 331-377. Since my primary purpose in this paper is not historical, I will focus only on Galileo's 1638 presentation of the refutation, bracketing the interesting question (a question not without philosophical interest) of why it was that such a simple and obvious mistake remained part of the West's scientific world view for nearly 2000 years.

[66] Page references will be to the page number in the National Edition.

the same ratio as their weights [*gravità*]" (Galileo 1638/1914, 106). That is, he is challenging the view that heavier bodies fall faster than lighter ones, and that they do so in direct proportion to their heaviness. On the version Galileo takes himself to be opposing, the proportionality is linear; "a moveable ten times as heavy as another is moved ten times as fast" as the other (Galileo 1638/1914, 106)[67].

The famous thought experiment, rephrased slightly, is the following: Imagine that a heavy and a light body are strapped together and dropped from a significant height[68]. What would the Aristotelian expect to be the natural speed of their combination? On the one hand, the lighter body should slow down the heavier one while the heavier body speeds up the lighter one, so their combination should fall with a speed that lies between the natural speeds of its components. (That is, if the heavy body falls at a rate of 8, and the light body at a rate of 4, then their combination should fall at a rate between the two[69].) On the other hand, since the weight of the two bodies combined is greater than the weight of the heavy body alone, their

[67] See Aristotle *Physics* 215a24-216a21; *De Caelo* 301b.

[68] Note that in the remarks that follow, all references to bodies should be understood as referring to bodies of the same material. For the purposes of my discussion, this constraint is irrelevant.

[69] The *Discourse* presents a four-day conversation among three characters: Simplicio, who represents the Aristotelian, Sagredo, who represents the intelligent lay-person, and Salivati, who represents Galileo himself. The relevant passage reads as follows: Salviati: "Then if we had two movables whose natural speeds were unequal, it is evident that were we to connect the slower to the faster, the latter would be partly retarded by the slower, and [the slower] would be partly [hastened] by the faster. Do you not agree with me in this opinion?" he asks Simplicio, who responds, "You are unquestionably right" (Galileo 1638/1914, 107). Note that Salviati need not be taken as disingenuous here. If the natural speed of all bodies is the same, then the initial clause of the conditional is never fulfilled (there are not "two bodies whose natural speeds are different"), so the condition of the consequent is met vacuously.

combination should fall with a natural speed greater than that of the heavy body. (That is, if the heavy body falls at a rate of 8 and the light body with a rate of 4, their combination should fall at a rate greater than 8.) But then the combined body is predicted to fall both more quickly, and more slowly, than the heavy body alone[70]. The way out of this paradox is to assume that the natural speed with which a body falls is independent of its weight: “large and small bodies move with the same speed provided they are of the same specific gravity” (Galileo 1638/1914, 109).

2.2.2 Reconstruction of the Galileo Case

Transformed into an argument that conforms to the strictures of the elimination thesis, Galileo's reasoning can be reconstructed as follows. The first claim of the Aristotelian is that:

(1) Natural speed is mediative.

That is, natural speed is a property such that if a body A has natural speed s_1, and a body B has natural speed s_2, the natural speed of the combined body A-B will fall between s_1 and s_2.

The second premise of the reconstruction is that:

(2) Weight is additive.

[70] Salviati: “But if this is so, and if it is also true that a large stone is moved with eight degrees of speed, for example, and a smaller one with four [degrees], then joining both together, their composite will be moved with a speed less than eight degrees. But the two stones joined together make a larger stone than the first one that was moved with eight degrees of speed; therefore this greater stone [*composite*] is moved less swiftly than the lesser one. But this is contrary to your assumption. So you see how, from the supposition that the heavier body is move more slowly than the less heavy, I conclude that the heavier moves less swiftly” (Galileo 1638/1914, 107-108).

That is, weight is a property such that if body A has weight w_1, and body B has weight w_2, the weight of the combined body A-B will be equal to the sum of w_1 and w_2.

From these two premises (plus the assumption that not all weights and natural speeds are either zero or infinite), it follows that:

(3) Natural speed is not directly proportional to weight.

For the first is a mediative property, whereas the second is an additive property, and a mediative property cannot be directly proportional to one that is additive. Furthermore, the only way to maintain (1), (2) and (3) simultaneously is to assume that all natural speeds are the same. Then weight might be additive and natural speed (in a vacuous sense) mediative, with no contradiction thereby implied. Thus natural speed is shown to be independent of weight.

2.2.3 Four Ways out for the Aristotelian

If the dispensability thesis is true, then Galileo’s thought experiment should be replaceable by some non-thought-experimental argument without loss of demonstrative force. My goal in the next two sections (**2.2.3** and **2.2.4**) is to show that the standard reconstruction presented in section **2.2.2** is not such an argument.

I begin my case by pointing out that there are a number of “ways out” for the defender of the view that natural speed is directly correlated with weight—a view which, for the sake of convenience, I will call the Aristotelian view. These ways out involve denying premises (1) and (2) by proposing a series of alternative hypotheses about the physical properties of strapped-bodies, that is, bodies of the sort described by the thought experiment. The point of talking about these ways out is to show that there are ways to maintain the negation of (3) by adopting alternatives to (1) and (2), and adopting these may well be less disruptive to the Aristotelian picture than giving up (3). What I will suggest below is that these ways out, though logically available, run counter to certain tacit knowledge about the physical world. It is for this reason that, when the case is presented as a thought experiment, they do not even occur to us. To block them as moves in a straight argument, however, requires metaphysical commitments that seem not to be at play in the thought experiment itself. What these commitments are, and what role I think they *actually* play in Galileo's reasoning is a point I will turn to after presenting the four ways out.

The first ways out would be for the Aristotelian to deny that the properties in question are *determinate* for strapped-together bodies in one of the following two ways.

(4) Natural speed is not physically determinate for strapped-bodies[71].

(5) Weight is not physically determinate for strapped-bodies.

That is, she might reject (1) or (2) on the grounds that they presuppose that natural speed and weight are properties that apply universally, even to bodies that are in some way monstrous[72]. Since strapped-bodies are odd entities, she might say, they

[71] For a version of this "way out," see Koyré 1968, 51.

[72] Galileo preemptively deals with this way out, by getting a concession from Simplicio straight away. Simplicio: "each falling body acquires a definite speed fixed by nature, a velocity which cannot be increased or diminished except by the use of force [*violenza*] or resistance" (Galileo 1638/1914, 107).

need not be governed by the sorts of laws that govern ordinary objects. In particular, they need not have determinate natural speeds or weights.

The third way out for the Aristotelian would be to avoid the conflict between (1) and (2) by saying that there *is* a fact of the matter about whether a strapped-body is one body or two, and that its physical properties in falling will depend on the answer to this question. She might say:

> (6) Natural speed and weight are mediative for strapped-bodies that are *united.* Natural speed and weight are additive for strapped-bodies that are *unified.*

That is, sometimes when two bodies are strapped together, they are merely *united* and remain, as a matter of fact, two objects; sometimes, when they are strapped together, they are *unified* and form, as a matter of fact, a single object. In the first case, both weight and speed will be mediative; so that the combined body will have a weight intermediate between those of the two original bodies, and fall with a natural speed that lies between the two original speeds. In the second case, both properties will be additive; so that the weight of the unified body will be equal to the sum of the weights of its component parts, and its natural speed correspondingly equal to the natural speeds of the two combined. Since the mediativity of the properties holds only with respect to united pairs of objects, and the additivity only with respect to unified single objects, there is no way that the Aristotelian can be forced to a contradiction. What she *is* forced to accept, however, are radical discontinuities in nature. A body, united, might be falling steadily at a rate of, say, six, and suddenly, should its parts happen to become unified, begin falling at a rate of, say, twelve.

But the Aristotelian can avoid the problem of discontinuity. A fourth way would be for her to say that given two bodies that fall together there is a fact of the matter about their degree of connectedness, and that this determines their physical properties when falling. The claim would be:

> (7) Natural speed and weight for strapped-bodies are determined by *degree of connectedness* (C) such that the speed/weight of B_1-strapped-to-B_2 where B_1 has w_1 and B_2 has w_2 will be: $(C)(w_1+w_2) + (1-C)((w_1+w_2)/2$[73].

We let C measure the degree of connectedness between the two bodies; that is, we let it be a number between zero and one that corresponds to the degree to which two bodies that fall together are unified: if the bodies are completely unified, C will take a value of one; if the bodies are completely disunified (that is, united), C will take a value of zero. For intermediate cases, the value will be between these two, and the speed and weight of the combined body will lie between the mean and the sum of the two initial values. So if the two bodies are completely unified, the additive law will apply completely; if the bodies are merely unified, the mediative law will apply throughout; and for intermediate cases, some proportional average will be found between them. Thus the assumption that natural speed is a function of weight can be maintained, and it can be maintained without violation of continuity, under the assumption that degree of connectedness is a relevant physical property. That this

[73] I make a number of simplifying assumptions here to keep the equation manageable. I assume that the units for measuring weight and natural speed correspond so that the number representing an object’s weight is the same as the number representing its natural speed; and I assume that the natural speed of two merely unified bodies is the mean of their individual natural speeds.

way-out too seems not to be a live option brings us to the point where I will suggest my alternative explanation of what is going on.

2.2.4 What the Reconstruction Misses

The Galilean argument can be saved by appeal to two broad, defeasible, tacit assumptions, each of which captures an important feature of our representation of experienced reality. One is that, for any body that one might encounter, there is a determinate fact concerning its weight and natural speed. That is:

(8) Natural speed and weight are physically determined.

The other is that there is *no* determinate fact whether strapped-bodies are one object or two. That is:

(9) Entification is not physically determined.

What (8) says is that a particular question about natural properties has a determinate answer. Any body, no matter how oddly shaped, will have a particular weight and a particular natural speed that are fixed by the world. What (9) says is that a particular question of entification has an indeterminate answer. Whether we consider a strapped-together body to be a single object, or two objects held together by a strap, or indefinitely many objects held together by internal forces, is merely a question of the aspect under which we choose to view that object. The answer to the question "how many objects?" does not follow from any *physical* property we might discover; it is a question about our words, not a question about the world[74].

These two premises are sufficient to eliminate the "ways out" enumerated above. If (8) holds, then (4) and (5) are not available as lines of escape; if (9) holds, then neither (6) nor (7) can be appealed to as a means of avoiding the Galilean conclusion. And the *way* in which they eliminate (4)-(7) is very different from the way that a simple reassertion of (1) and (2) would. They show not only *that* there is something wrong with (4)-(7) as descriptions of the way the world is, but *why* there is something wrong with them. They show *what* about our tacit understanding of physical reality, and of our instincts concerning plausible candidates for physically relevant and irrelevant properties, is missed by someone who appeals to (4) or (5) or (6) or even (7).

So the initial reconstruction of the Galilean thought experiment fails to capture what is really doing the work in the case. As hypotheses about the ways strapped-bodies might behave in fall, (4), (5), (6) and especially (7) are in principle available as alternatives to (1) and (2); just as natural speed might be mediative and weight additive, it might be that both natural speed and weight depend on the degree to which the two bodies are connected. So if (1)-(3) were truly capturing what is going on in the Galilean thought experiment, there would be ways out for the Aristotelian that would allow her at least to shift the burden of proof back to the Galilean.

[74] I discuss some of the implications of this fact in the case of artifacts in chapter 3

Continues on next page…

That these ways out do not seem available when the thought experiment is presented in its unreconstructed form shows that this eliminative reconstruction has failed to capture its original demonstrative force. (What has been lost is the way in which, by evoking tacit knowledge about the how falling bodies actually behave, the thought experiment preemptively precludes such ways out.) Accordingly, I tried to come up with a reconstruction sufficiently strong to rule out (4)-(7) in a similarly categorical and decisive manner. This involved appeal to two rather comprehensive and metaphysical-sounding principles, namely (8) and (9). (8) and (9) give background support to (1) and (2) and thereby help to establish (3).

But just as (1) and (2) are too weak to capture the way in which alternative hypotheses concerning fall are ruled out by the thought experiment, (8) and (9) are too strong. They represent approximate articulations of defeasible assumptions about the physical world. But as they stand, they articulate principles that have less certainty than the conclusion they are taken to support. Prior to contemplation of the case Galileo describes, the Aristotelian may be committed to *something* like (8) and *something* like (9), but he is certainly not committed to them unmodified. To reconstruct the case as an explicit argument with some version of (8) and (9) among the premises would require enumerating outright their defeasability conditions. But this is something he does not know how to do. Contemplation of the case Galileo describes *brings him to see* that these principles are not defeated in *this* case. And it is this recognition that serves as the basis for the case's power. No austere

below.

argumentative reconstruction will be able to this, because part of the thought experiment's function is to bring the Aristotelian to accept certain *premises*. In the next section, I will discuss what makes belief in these premises *new*, and what makes it *justified.*

2.3 Denying the Dispensability and Derivativity Theses

2.3.1 Rejecting Reconstruction: What the Thought Experiment Does

If the dispensability thesis were correct, then the conclusion established by Galileo's thought experiment—that "both great and small bodies...are moved with like speeds" (Galileo 1638/1914, 109) —*should* be demonstrable by means of a non-thought experimental argument. "Demonstrable" here means: rationally justifiable on the basis of the same background conditions. So let us spell out what the background conditions in question are.

For the Aristotelian, daily experience seems to confirm the theory that heavier bodies fall faster than lighter ones. Just as the Galilean sees cases where lighter bodies fall more slowly than heavy ones as exceptional, so the Aristotelian sees as crying out for explanation those cases where the rate of fall is (nearly) simultaneous. That is, when gold beaten into a very thin leaf reaches the ground more slowly than a solid lump of the same material[75], the Galilean must posit some factor that explains the divergence of this result from the generally expected outcome. Similarly, when the Aristotelian sees two stones of very different weights fall to the ground with like speeds, the circumstance requires diagnosis and explanation.

So far all I have pointed out is the possibility of maintaining theoretical commitments in the face of apparent counter-evidence. This can be done by appealing to additional principles that explain away anomalous data by showing that the phenomena at issue are subject to the fundamental principle in question, but that the world's complexity has prevented them from manifesting this. So the Galilean might appeal to air resistance, the Aristotelian to the fact that the bodies have not been dropped from a height sufficiently great[76]. Such explaining-away of recalcitrant exceptions is not a desperate move by a failing paradigm; it is a fundamental element of doing normal science in a non-ideal world such as the one we inhabit[77].

The point of this discussion is to give a better sense of the background conditions under which the argumentative reconstruction must be demonstratively

[75] See Galileo 1638/1914, 109.

[76] In Galileo's dialogue, when Salviati points out that a rock of two pounds and a rock of twenty pounds will strike the ground nearly simultaneously when dropped from a height of one- or two-hundred feet, Simplicio retorts that this may just be a result of not having given the objects enough falling-time for the differences to become apparent (see Galileo 1638/1914, 109-110). Simplicio says: "Perhaps from very great heights, of thousands of *bracchia [58.4 cm*] [discrepancies] would follow which [are] not seen from lesser heights." That is, Simplicio suggests that perhaps the reason we do not observe the sorts of differences which Aristotle's theory predicts is that we are making observations under non-idealized circumstances, and that were we to eliminate distorting elements--like the fact that the fall is only a few hundred feet--the true phenomena would reveal themselves. Galileo provides Salviati with a rather sharp retort. See Galileo 1638/1914, 100 and Galileo 1638/1974 rev. 1989, 69, footnote 44.

[77] Of course, it is a commonplace in the history of science that at a certain point the burden of positing literal or metaphoric epicycles becomes too great, and the theory—especially when there is a simpler alternative available—collapses under the weight of its own internal complexity. But it is a similarly well-established commonplace in philosophy that theories are underdetermined by evidence, and that extra-scientific considerations do some of the work in determining theory choice.

forceful if the dispensability thesis is to be shown to hold in this case. The thesis tells us that some non-thought experimental reconstruction of the case Galileo presents should be able to do the same thing that the non-argumentative version does: lead the Aristotelian from the same background assumptions to the same rationally justified conclusions. What I have argued above in some detail is that the standard reconstruction fails in this regard. The thought experiment does not take the Aristotelian from commitments to (1) and (2) to the conclusion (3); nor does it take the Aristotelian from (1) strengthened by (8) and (2) strengthened by (9) to the conclusion that (3). Prior to the thought experiment, the Aristotelian is explicitly committed to the *negation* of (3); this commitment serves as a filter through which apparently contrary evidence is reinterpreted. The thought experiment must be sufficiently forceful to undermine this framework-shaping assumption. What I have argued is that neither the original nor the revised reconstruction comes close to this: the former is subject to challenges such as the four ways out; the latter rests on shaky metaphysical principles that have less certainty than the conclusion they are taken to support.

Suppose, however, that somehow it were possible to come up with an argumentative reconstruction that almost exactly captures the strength and limits I have attributed to Galileo's thought experiment[78]. Would we then be justified in accepting the deeper methodological claim put forth in the derivativity thesis: that the *justificatory force* of whatever beliefs I hold via thought experiment is a function of the thought experiment's argumentational essence? In the next section (**2.3.2**), I argue that the answer to this question is "no."

2.3.2 Rejecting the Positive Argument: What Makes these Beliefs *New*?

The derivativity thesis says that the justificatory force of any good scientific thought experiment can only be explained by the fact that it can be replaced, without loss of demonstrative force, by a non-thought-experimental argument. So rejection of the thesis can take two forms: denying that the justificatory force of a particular scientific thought experiment *can* be explained this way, and denying that it can *only* be explained this way. The simplest way of doing the former, of course, would be to show that the dispensability thesis is not true of some thought experiment; obviously, if there is no non-thought-experimental argument with which the thought experiment can be replaced without loss of demonstrative force, than the justificatory force of the thought experiment cannot be explained by the possibility of such replacement. But for the sake of argument, I am supposing that the dispensability thesis is (at least approximately) true. This leaves two avenues for denying the derivativity thesis: denying that the argumentative reconstruction explains the justificatory force of some thought experiment *at all*, and denying that it explains such justificatory force *entirely*.

As I discussed above (**2.1.3**), two sorts of defense are offered for the derivativity thesis. The negative argument contends that thought experiments are

[78] Suppose, for instance, we were to enumerate the defeasibility conditions of (8) and (9), an we included these modified versions as premises alongside (1) and (2).

arguments because there is nothing else for them to *be*; the positive argument contends that thought experiments are arguments because their "analysis and appraisal" involves explicit argumentative reconstruction. In this section I will address the positive thesis, suggesting that it fails to get at what is most interesting about thought experimental reasoning; in the final section, I offer some thoughts about what sorts of alternative justifications are available such that the negative thesis too is untenable.

My contention throughout this chapter has been that it is a mistake to think that the *reason* conclusions to thought experiments are justified is because thought experiments have argumentational analogues (if indeed they do). Rather, I want to suggest that the Aristotelian comes to have novel justified true beliefs about the empirical world not because he has (whether he knows it or not) followed along the path of a recognized argument form, but rather because he has performed an act of introspection that beings to light heretofore inarticulated and (because he lacked a theoretical framework in which to make sense of them) heretofore implausible tacit beliefs. There are two things that I need to show myself able to explain. The first is how it is that knowledge has been *gained*. In what way is it that the Aristotelian has come to believe something *new*? The second is how it is that *knowledge* has been gained. In what way is it that the Aristotelian has come to believe something *justified*? I will answer the first in the remainder of this section, and the second in section **2.3.3**.

In addressing the issue of novelty, I will, begin with brief remarks about what I think is *not* at issue, and then say something about where I think the important questions rest. One might say that the beliefs are not new since, in some sense, the Aristotelian had access to them before. After all, he has acquired no new knowledge of the external world; all he has done is reshuffle partly tacit beliefs he already held. But this view of what makes knowledge new is surely too stringent; it would, among other things, rule out all mathematical reasoning as a potential source of new knowledge. On the other hand, it seems too weak to say that a belief is new if it merely results from putting together two explicitly held beliefs that have not, for the individual in question, been previously connected. If I believe that snow is white and I believe that crows are black, but I have never thought about the two at the same time, it seems wrong to say that the belief that snow-is-white-and-crows-are-black should count as a *new* belief for me[79]. In any case, without spelling out precisely what it is that makes some beliefs new and others mere implications, there is a simple reason to think the Aristotelian's belief that the speed at which a body falls is independent of its weight should count as a new belief for him: Until recently he was explicitly committed to the truth of its negation. This alone suggests that—whatever the implicatory relation between his prior commitments and this view about natural speed—the belief should count as new.

[79] Except under very odd circumstances. Suppose I believe that crows are black because I live in a village where there are crows, and I have seen many of them. I also believe that snow is white, because I have read about it in books. In my village, it is taboo to think about black things and white things at the same time, because this is thought to allow the evil spirit access to the soul. One day, I leave my village for the north, and I observe a crow circling above a field of snow. I find the image aesthetically striking, and I ask myself why this is so. In analyzing my response to the visual experience I am having, I realize with a start: "Snow is white and crows are black." In such a case, it seems plausible to suggest that this is a new belief.

Continues on next page…

But there is a deeper and more interesting way that the belief is new, and that is the following. The thought experiment that Galileo presents leads the Aristotelian to a reconfiguration of his conceptual commitments of a kind that lets him see familiar phenomena in a novel way. What the Galilean does is provide the Aristotelian with conceptual space for a new notion of the *kind of thing* natural speed might be: an independently ascertainable constant rather than a function of something more primitive (that is, rather than as a function of weight). It is in this way, by allowing the Aristotelian to make sense of a previously incomprehensible concept, that the thought experiment has led him to a belief that is properly taken as *new*.

What this suggests is that the derivativity thesis is *missing the point* of what makes the Galileo case work as it does. The recognition that natural speed is independent of weight comes not from tracing the implications of antecedent commitments to (1) and (2), which, after all, lead to the denial of a position to which the Aristotelian is explicitly committed (and thence to retreats such as the four ways out). The recognition comes from the sudden realization, on the part of the Aristotelian, of the conceptual possibility of a certain sort of physical property. Prior to contemplation of the case, there was no room on the Aristotelian picture for the thought that natural speed might be constant, not varying—that it might be dependent not on some specific features of the body in question, but only on the fact that it is a body at all[80]. After contemplation of the case, there seems to be no conceptual space for the view that it might be variable[81].

This is not the place to discuss general issues of incommensurability across theory-change; my only purpose is to give some sense of the *mechanism* whereby novelty arises in this case. What is important for our purposes is this: one of the things that enables this rather striking shift in the representation of physical reality is that the Aristotelian recognizes that there are experientially possible objects—strapped-together bodies—for which the defeasability conditions of (8) and (9) are not met (that is, objects of which (8) and (9) hold true), and that these are objects for which his old notion of natural speed simply *does not make sense*. If entification is arbitrary and natural speed and weight are fixed by the world, then a feature-dependent notion of natural speed is just plain incoherent. So one way of thinking about how the thought experiment works is this: it brings the Aristotelian to recognize the inadequacy of his conceptual framework for dealing with phenomena

[80] To get a sense of how odd this transition is, try thinking about *weight* as something dependent not on the specifics of the body in question, but as something constant for all bodies. (That is, to ascertain a body's weight, we would not need to know anything more about it that the simple fact that it is a body.) Clearly this would be a major conceptual readjustment; one might even be inclined to say that we aren't talking about *weight* anymore, since whatever sort of thing weight is, it is surely something that depends upon specific features of the bodies to which it applies. The analogy is not perfect, since part of what happens as a result of thinking about the Galileo case is that it becomes apparent that there is no physical application for the Aristotelian idea of natural speed; like phlogiston, it disappears into the ether of abandoned concepts.

[81] For other examples of this sort of conceptual shift, see Appendix A.

which—through the contemplation of this imaginary case—he comes to recognize as always having been part of his world.

What this suggests is that "the *analysis* and *appraisal* of a thought experiment" need not "involve reconstructing it explicitly as an argument," so that "a good thought experiment is a good argument, a bad thought experiment is a bad argument." After all, the argument from (1) and (2) to (3) is no better or worse than the argument from (not-3) to (not-1) or (not-2). Like an experiment, *part* of what makes a thought experiment good or bad is the validity of the procedure by which the same result can be repeatedly obtained. But another thing that distinguishes good thought experiments from bad is their ability to direct the reader's attention to inadequacies in her conceptual scheme that she herself recognizes immediately, as soon as they are pointed out to her. It is *this,* I want to suggest, that grounds her new beliefs. Of course, I have said nothing so far about what makes these beliefs *justified*. It is to this issue that I turn in the final section of this chapter.

2.3.3 Rejecting the Negative Argument: What Makes these Beliefs *Knowledge?*

Thought experiments work in a variety of ways. By describing appropriately selected imaginary scenarios, they provide contexts within which sense can be made of previously incomprehensible conceptual distinctions[82]. This happens when two features that are constantly conjoined in our representations of all actual cases are imaginatively separated in the thought-experimental scenario in a way that shows them to have been isolatable all along. By describing specific situations, thought experiments, like analogical reasoning in general, can justify conclusions about particular cases without explicit or implicit appeal to more general absolute principles[83]. Many of the higher-level principles by which we negotiate the world are *defeasible*, and the determination of their applicability to particular situations must be made on a case-by-case basis. By bringing the reader to focus on particulars, thought experiments can help the reader distinguish warranted from unwarranted applications of the principle in question[84].

All of this provides some illumination of why thought experiments are *heuristically* useful. But it tells us nothing about what makes them *justificatory*. After

[82] By this I mean they do something like what the answer to a riddle does in making suddenly intelligible what previously appeared to be a nonsensical description.

[83] For an articulation and defense of this view of analogical reasoning, see Brewer 1994: “A legal reasoner uses the resources of analogy in a ‘context of doubt’ in order to build and maintain confidence in her judgment about how that doubt is to be resolved. One of the most important ways in which *legal* analogists seek to build and maintain confidence is by showing that show [*sic*] defeasibility is defeated in the target case because it was *likewise* defeated in the source case. One keeps one's eyes on the shared characteristics of source and target, that is, because one knows the source case managed to defeat defeasibility” (Brewer 1994, 61; Brewer credits Robert Nozick with joint responsibility for the idea, see Brewer 1994, 61 footnote 153). Or again: “[A]rgument by analogy consists not just of a narrow argumentative process of inferring the truth or probable truth of some propositions from the truth or probable truth of others. It involves also the abductive of *discovering* the rules to be applied, of making sense of *patterns* of characteristics, of putting characteristics into rule-like patterns” (Brewer 1994, 32). See also Sunstein 1993, as well as his Tanner Lectures (unpublished, delivered at Harvard University December 1994). Sunstein identifies four “different but overlapping features” of analogical reasoning in law: *principled consistency; a focus on particulars; incompletely theorized judgments;* and *principles operating at a low or intermediate level of abstraction*. (Sunstein 1993, 746); for criticism, see Brewer 1994, 19 footnote 55.

all, as Norton would ask, if the thought experiment is not an argument, why should we put faith in its conclusion? What I want to suggest in the remainder of this chapter is that *constructive participation* gives part of the answer: the justificatory force of the thought experiment actually comes from the fact that it calls upon the reader to perform an *experiment in thought*.

An experiment in thought is an *actual* experiment; the person conducting the experiment asks herself: “What would I say/judge/expect were I to encounter circumstances XYZ?” and then *finds out* the (apparent) answer. To make the technique seem less odd, let me point out two instances where the use of such methodology is widespread and fruitful. First, experiment in thought is one of the primary techniques of contemporary linguistics: “when linguists want to test hypotheses about the structure of a particular language, their methodology crucially involves...real experiments carried out in thinking. This peculiar kind of experiment...[is] experiment by introspection” (Thomason 1991, 247). Examples of experiment in thought include, then, judgments concerning the grammaticality of certain sentences, or the meanings of certain phrases. Experiment in thought also lies at the heart of moral reasoning: “it is precisely our moral views about examples, stories, and cases which constitute...data for moral theorizing” (Thomson 1986, 257). Ascertaining whether we consider a particular course of action to be morally permissible may well involve an experiment in thought.

[84] This often happens when the particulars are sufficiently well-sketched to evoke practical as well as theoretical responses.

How does this connect with the Galileo case? What kind of experiment in thought plays a role there? Answer: by thinking about the case in question, we discover what sorts of motions and objects we think are possible in the world. Do we think objects can be strapped together? Yes, we do. Do we think objects fall with radical discontinuities in speed? No, we think they do not. Do we think entification is something that is fixed by the world? No, we do not. Do we think weight and natural speed are fixed by the world? Yes, we do. We *come to recognize* that we have these beliefs by contemplating such cases, *and*--and this is where the justificatory work comes in--the fact that we have these beliefs gives us *prima facie* warrant to think that they are true.

In the *Science of Mechanics*, a few pages after coining the expression *Gedankenexperiment*, Mach writes:

> Everything which we observe imprints itself *uncomprehended* and *unanalyzed* in our percepts and ideas, which then, in their turn, mimic the process of nature in their most general and most striking features. In these accumulated experiences we posses a treasure-store which is ever close at hand, and of which only the smallest portion is embodied in clear articulate thought. The circumstance that it is far easier to resort to these experiences than it is to nature herself, and that they are, notwithstanding this, free, in the sense indicted, from all subjectivity, invests them with high value (Mach 1942, 36).

This, of course, is the beginning not the end of an answer to the question. But it is sufficient for the purposes of this chapter. What I have been trying to show is that *something* besides argument might give justificatory force to thought experimental reasoning. The alternative I have proposed has been this. By focusing on imaginary scenarios and making reference to particulars, thought experiments can provide a fulcrum for the reorganization of conceptual commitments; this explains the way in which they can provide us with novel *information* without empirical input. And by bringing the reader to perform experiments in thought, thought experiments can lead us to reject shaky (and ultimately false) theoretical commitments in light of newly systematized but previously inarticulable *knowledge* about the way the world is. The justificatory force of thought experiments is thus parasitic on the extent to which the messy twisted web of background beliefs that underpin our navigation of the world are rightly considered knowledge. To establish this, on coherentist or evolutionary or empiricist grounds, is not my aim here. What I have succeeded in doing, I hope, is in showing what it is that the argumentative reconstruction picture misses.

2.4 Conclusion

In this chapter, I have argued that a certain view about thought experiments in science is false. The view is that any scientific thought experiment can be reconstructed as a non-thought-experimental argument without loss of demonstrative force. In the first part of the chapter, I explained the philosophical motivations for adopting such a view, and distinguished two versions of the position. The first—the dispensability thesis—concerns the replacability of thought experiments; the second—the derivativity thesis—concerns their justificatory force. In the remainder of the chapter, I tried to refute both of these theses. Through a detailed discussion of a thought experiment of Galileo's, I tried to show that the standard argumentative reconstruction of the case fails to capture its justificatory power, and I suggested reasons to think that any other argumentative reconstruction would fail in similar ways. I then argued that even if one were to provide an argumentative reconstruction that did almost perfectly capture the thought experiment's demonstrative force, this would not show that the *reason* the thought experiment is successful is because, deep down, it is nothing more than an argument in disguise. I suggested that, to the contrary, the success of the thought experiment is a result of the way in which it invites the reader's constructive participation, describes particulars in ways that make manifest practical knowledge, and describes an imaginary scenario wherein relevant features can be separated from those that are inessential to the question at issue.

The final task is to connect this discussion to the broader goals of the dissertation. One such goal is to establish that guided contemplation of imaginary cases can lead to new knowledge of the world. Showing that this happens in the Galileo case provides me with an example that helps me to establish this conclusion. But there is also a second aspect to the dissertation's broader project. In chapters 3 and 4, I will focus more explicitly on the role that exceptional cases should play in the development of theories. And in chapter 4, where I will contend that something goes methodologically wrong in the far-fetched cases such as those described in the personal identity literature, one of my aims is to pin-point exactly *where* that is. What I have shown so far is that in some cases, constructive participation in the guided contemplation of imaginary scenarios *can* lead to new knowledge. So although in certain cases this sort of guided contemplation seems to lead us astray, what I have tried to show in these pages is that imagination *as such* is not the problem.

3. The Ship of Theseus

3.1 Conceptual Thought Experiments

Having examined an example of a *factive* thought experiment (one where the question being asked was: "what would happen under thus-and-such circumstances?") we now turn to an example of what I call a *conceptual* thought experiment. Here, the question being asked is: "how, given that we know what would happen under thus-and-such circumstances, should we *describe* what happens?" That is, how should the case we are confronted with be accounted for? In particular, this chapter will address instances of how we develop or ascertain or establish criteria for what it is to apply (correctly) the expression "*x* is the same F as *y*" when the candidates in question are separated in space or in time. We are concerned here with *conceptual* cases, cases where all of the observational data have been agreed upon so that, at least in one sense, there is no dispute about what there is[85/86]

As before, I will suggest that what gives such thought experiments their justificatory force is the fact that they call upon the reader to perform an *experiment*

[85] For a particularly engaging presentation of these issues, see Chisholm (1971), 4-5. See also (among others) Shoemaker (1963), 3-5; Perry (1975), 8-10.

[86] Note that I am using the term "conceptual" in way which differs from the way it is used by Parfit (see section **3.5.2** below). The sort of case I am considering is one where all observational data are agreed upon, but where there might still be issues about what sorts of underlying facts about the world are producing these phenomena. As Parfit points out, there is a second sense in which we might speak of a question as being conceptual, namely, in cases where there is full agreement about all the data *and* full agreement about the proper description of the data in terms that all parties agree are the only causally relevant ones; remaining disputes concern nothing more than what all parties would agree are merely issues of *labeling*.

in thought. She asks herself: “what would I say about thus-and-such a case?”, and then, by observing her own response, she finds out the answer to this question. The answer would be something like: “I would judge that I was confronted with an instance of so-and-so” or “I would feel some pull to say P and some pull to say not-P” or “I’d be pretty confused and wouldn’t really know what to say.” Here again, performing an experiment in thought lets us come to recognize that we have certain beliefs, beliefs which might have remained hidden were they not brought to the fore by the contemplation of this imaginary exceptional case.

In addition, this chapter aims to establish a number of specific claims, the most important of which is this: Standard interpretations of the puzzle of the Ship of Theseus have drawn the wrong sorts of general conclusions from the particular case by assuming that the proper way to understand artifact identity is to allow the exceptions to drive the norms, rather than allowing the norms to drive the exceptions. That is, they attempt to resolve the puzzle by making modifications to our pre-reflective identity criteria for objects in general.

Cases where identity is extrinsically determined are cases where whether *x* is identical with *y* depends on something other than the spatio-temporal and causal relations between *x* and *y*. In such cases, our ordinary criteria of identity pick out more than one candidate; they cease working as criteria of identity, and instead become criteria of *identity-candidacy*. But, I will argue, it is only against a background norm of intrinsically-determined identity that we can make sense of local instances of extrinsically-determined identity. Most resolutions of the puzzle misidentify the implications of such cases by locating the extrinsic determination in the identity criteria. What I will argue instead is that the best way to account for such cases is to locate the extrinsic determination in the *processes* by which identity is generally preserved.

3.2 The Story

The imaginary scenario which will concern us in this chapter comes in three versions, which I will unimaginatively refer to as "the first version," "the second version," and "the third version," respectively.

First Version: There was once a thirty-oared ship belonging to Theseus, which during its seaworthy years underwent gradual repair. Over the years, one by one, each of its original planks was replaced with a new plank of the same size and shape and material, and the old planks were gathered in a barn on the shore, where they rested, piled in a heap. Eventually none of the original planks remained as pieces of the seaworthy vessel, though its appearance was unchanged and its duties unaltered. And although the sophists "of Athens were wont to dispute [whether] after all the planks were changed, [it was] the same numerical ship as it was at the beginning" (Hobbes, *De Corpore* II: 11), Theseus himself faced no difficulties, either practical or metaphysical. He felt no need to (re)christen the craft, no worries about which of the vessels at the dock was his, and no puzzles about whether, strictly speaking, his current ship is the same ship as the ship he commissioned some years back.

Or, to contrast that first version with a *second version*: There was once a thirty-oared ship belonging to Theseus, which during its seaworthy years was disassembled during the winter season. Each fall, when the frosts came, the planks of the ship were pried apart, and stacked in a barn on the shore, where they rested, piled in a heap. Each spring, the pieces were reassembled in a form identical to that of the original ship, so that its appearance remained unchanged and its duties unaltered. And although sophists were wont to dispute its status, arguing that the assembled ship was a different ship than the ship that plied the seas last season, Theseus himself faced no difficulties. He felt no need to (re)christen the craft, no worries about which of the vessels at the dock was his, and no puzzles about whether, strictly speaking, his current ship is the same ship as the ship he commissioned some years back

Or, to tell the story in a *third version* which is philosophically puzzling: There was once a thirty-oared ship belonging to Theseus, which during its seaworthy years underwent gradual repair. Over the years, one by one, each of its original planks was replaced with a new plank of the same size and shape and material, and the old planks were gathered in a barn on the shore, where they rested, piled in a heap. Eventually none of the original planks remained as pieces of the seaworthy vessel, though its appearance was unchanged and its duties unaltered. One fine afternoon, Theseus collected the planks from the barn, and nailed them together in a form identical to that of the original ship, a form shared by the repaired ship which continued to ply the waters. Suddenly, Theseus faced difficulties, both practical and metaphysical. There were two ships before him: Did one (or both) require rechristening? Which of them was the fine ship of Theseus? Was either of them identical with the ship he had commissioned some years back?

3.3 The Puzzle

The perplexities raised by this ancient story are a consequence of the fact that in a world of change, criteria for identity over time permit disruptions in spatio-temporal continuity. In particular,

(1) objects can survive disassembly and subsequent reconstruction, and

(2) objects can survive the gradual replacement of component parts over time[87]

Moreover, it seems that objects can straightforwardly survive *both* (2) (replacement of parts) *and* (1) (full disassembly and reconstruction)–so long as the two processes take place in the proper sequence. Suppose, for instance, that an object is first disassembled, then reconstructed, and then has its parts gradually replaced. So long as each of these processes is identity-preserving on its own, no *extra* problem arises from the processes taking place successively. So the problem we face in thinking about the Ship of Theseus is not that the entity has undergone *too many* ordinarily identity-preserving processes to have survived the ordeal; the worrisome feature is not that, with this last process of alteration, the line between sameness and difference has suddenly been crossed.

Rather, the puzzle seems to arise as the result of a particular pattern of events: the two processes are intertwined in a particular way, such that the part-replacement takes place *while* the object is being disassembled. And here, even though the only

[87] Cf. Wiggins (1980), 91. Cf. also Hirsch's distinction between what he calls the "sortal rule" (Hirsch 1982, 34-64; cf. Wiggins 1980, 35ff) and the "compositional rule" (Hirsch 1982, 64-71). These are discussed in more detail in section **3.6.1** below. Cf. also the distinction between the principle of *spatio-temporal continuity of form*

Continues on next page…

sorts of events that have occurred are events whose ordinarily result is the (mere) preservation of certain objects, in this particular case, the result of the combination of processes is that at least one new object has been created. This, then, is the puzzle that will concern us regarding the ship of Theseus, namely, *that a process which is ordinarily identity-preserving is in this instance entity-creating*.

Another way to put the problem is this: we need to be able to differentiate cases where the process in question is identity-preserving from cases where it is entity-creating. If we cannot, then given that such a process may be *clearly* entity-creating (in that it might result in the creation of more than one equally viable candidate), why should we be so sure that it is identity-preserving in the first case? That is, if an intrinsically-specified process could *ever* be entity-creating, why should we assume that it is *ever* identity-preserving[88]?

It is precisely this line of reasoning that I wish to block. The Theseus case is exceptional because it describes a circumstance in which an ordinarily identity-preserving process is instead entity-creating. But to conclude from this that a particular sort of process is never identity-preserving, or that identity-criteria in general are somehow suspect, is a mistake. Rather, the correct response to such a case, as I will argue in the final section (**3.7**) is to locate the problem in the

and the principle of *identity of parts* (or *identity of matter*). See, for instance, Smart (1972) and Smart (1973), Scaltsas (1980).

[88] Cf. Nozick (1980): There is a principle that “if there could be another thing so that then there would not be identity, then there isn’t identity, even if that other thing does not actually exist” (32). (Nozick believes that this principle is false.)

specification of the process. But in order to get there, I first need to say more about what I mean by identity over time[89].

3.4 Is the Ship of Theseus an Exceptional Case?

3.4.1 Automatic and Specially-Secured Identity

Following Mackie, we might distinguish between "thing-concepts which are such that their ordinary identity-conditions automatically ensure conformity to the logic of identity and those such that special clauses are needed to secure that conformity" (Mackie 1976, 149). The logic of identity, of course, requires reflexivity, symmetry, and transitivity. Everything is identical to itself; to anything which is identical to it; and to anything identical to anything identical to it.

This means, among other things, that if an object at one time is (strictly) identical to an object at a later time, then anything else that the earlier object is identical to is also identical to the later object. Suppose for the moment that we take the two candidates in the third version of the story to be equally good candidates for being identical with the original Ship of Theseus, and, in an effort to be equitable, we say that the original Ship of Theseus is identical both to the continuously-repaired

[89] Throughout my discussion in this chapter, I am ignoring four-dimensionalist positions (that is, positions which hold that objects are extended in time as well as space). Because I think philosophy should try to make as much sense of our ordinary ways of thinking about the world as possible, I consider such views to be non-starters. This is not to deny that four-dimensionalist positions offer elegant ways of resolving perplexities that confront other views. This chapter represents an attempt to make sense of what seem to me serious conceptual commitments that we have concerning the status of artifacts as real entities that persist through time. It may be that these

Continues on next page…

ship and to the original-planks ship. Since we are using "(strictly) identical" to mean "related by a relation which conforms to the logic of identity," then these two ships must be identical to one another. But, one might continue, since it is evident that they are not, then the original ship cannot be identical to both of them[90].

Turning next to the first and second versions of the story, we note that in each of these stories the original ship *was* identical to some ship produced by a process qualitatively identical to one of the processes described in the third version. However, it is fully compatible with (for instance) the first version of the story that, long after the tale has been told, the discarded planks could be reassembled in the form of the original ship. So, Mackie would say, identity in this case is "specially-secured;" we need an explicit no-competitors clause to guarantee that the criterion of identity which the ship meets (continuous replacement of parts over time) actually serves here as a criterion of *identity* and not as a criterion of *candidacy-for-identity*.

Mackie argues that such a no-competitors clause is required for any (non-abstract) entity whose continued (diachronic) existence requires more than the persistence of some simple substrate[91]; the ordinary identity-conditions for organisms and artifacts—including ships and trees and human beings—do not offer the *guarantee* that transitivity will hold. The problem can be illuminated as follows.

commitments are ultimately incoherent and need to be abandoned. But for the purposes of my discussion, I will take them as fixed.

[90] Assuming there was only one original ship to begin with.

[91] See Mackie 1976, 150.

Take (Mackie's reading of) Locke's general theory of identity through time: that "*x*-occurrences at t_1 and t_2 are occurrences of the same *x* if and only if there is a continuous *x*-history linking them" (Mackie 1976, 149). That is, two trees at t_1 and t_2 are the same tree only if there is a continuous tree-history linking them, two stereo systems at t_1 and t_2 are the same stereo-system only if there is a continuous stereo-system-history linking them, and so on. The problem that arises is this: tree-histories and stereo-system histories allow for the possibility that more than one tree or stereo at t_2 might be tree-identical or stereo-identical with a single tree or stereo at t_1 . For instance, suppose the ordinary identity conditions for trees are that a tree at t_1 is tree-identical with a tree at t_2 if and only if the two bear such-and-such relations. But now suppose that when tree A was a young sapling, it was uprooted, carefully divided into two equal parts, and the two parts—call them B and C—were transplanted to spots equidistant from the tree's initial rooting. And suppose that both B and C bear such-and-such relations to A. Then by tree-identity, B is tree-identical with A, and so is C. But clearly B is not tree-identical with C. So transitivity fails[92]. A similar story could be told for the stereo, or, indeed, for any organism or artifact whose identity conditions allow for certain sorts of replacements of parts over time, or for survival under conditions of loss of matter. For organisms and artifacts, then, it is in general

[92] Alternatively, one might think of this as being a case where symmetry fails. We might say: A is identical with B but B is not identical with A, thereby preserving the transitivity of tree-identity. In either case (that is, if symmetry or transitivity is not satisfied) tree-identity fails as a strict condition of identity over time. I think locating the problem in transitivity better captures the intuitive sense of what is going on.

possible that there be two spatially distinct simultaneously existing entities, each meeting the requirements for *x*-identity with some earlier entity.

The case we are concerned with is one where it is possible that two (or more) distinct individuals would each be adequate candidates for identity with some earlier individual, but for the existence of the other competitor(s). Mackie's terminology provides a convenient way to describe the difference between cases where such problems might arise, and cases where they cannot. The identity of objects whose identity-conditions are such as to rule out the possibility of *x*-identity holding without strict identity also holding is *automatic*. The identity of objects whose identity-conditions *do* allow the sort of multiple-candidacy described above is *specially secured.*

3.4.2 Organisms, Artifacts, and Exceptional Cases

With this distinction in place, let us return to the issue which concerns us in this chapter. I have suggested that it is a mistake to allow our normal ideas of identity to be driven by exceptional cases. But hasn't Mackie's argument just shown that the Ship of Theseus case is anything but exceptional? Hasn't he established that it fits a pattern which is followed when we enumerate identity conditions for any artifact or organism? What could be more normal than this?

I have two responses to offer here. The first concerns organisms, the second concerns artifacts. First: it is important to notice that in the case of organisms, the only sorts of cases where identity-criteria will permit multiple-candidacy are cases where some small proportion of the original mass of the entity (half or less) serves to support the organism's continued existence[93]. That is, it is only under cases of severe

[93] I am assuming that organisms cannot survive full disassembly and subsequent reconstruction. Van Inwagen concurs: "Note that there is no tendency to identify a 'reassembled' *organism* with the 'original.' If God were to 'reassemble' the atoms that composed me ten years ago, the resulting organism would certainly not be *me*" (van Inwagen 1990, 140). This is not to say that organisms cannot survive operations in which their organs are removed and replaced, only that they cannot survive being fully disassembled into particle-sized bits, and then reassembled from those bits.

Does this concession raise the possibility of the following Theseus-like case? Over time, my body parts are replaced one by one, and each of the discarded organs is connected up to some sort of machine which enables it to remain functional. After the process of removal is complete, the discarded organs are reassembled in the form of a human being. Is that (a candidate for being) me? Here, even though the relevant pieces are pretty large, I am inclined to say that the reassembly process is more like God collecting the atoms that composed me ten years ago than like the surgeon removing my kidney, cleaning it, and replacing it in my otherwise intact body. That is, the process is *not* identity-preserving. And it is not identity-preserving precisely because it involves full disassembly and subsequent reassembly.

A more difficult version of the case (due to Robert Nozick) would be one in which the machine to which the disassembled parts are connected up is an artificial body, whose parts are gradually replaced over time by the parts which originally belonged to the original human body. Here, the disassembled parts do not need to be *re*assembled; they are assembled as they are connected up to the machine that preserves them. To the extent that we would be inclined to call the person who ultimately results from this process (a candidate for identity with) the original person, it seems that we *are* committed to the idea that organisms can survive full disassembly and subsequent reconstruction.

My inclination is to treat this case as a fission case. Up to a certain point, it resembles the surgeon removing and cleaning the kidney: the original person remains with the original body, while some of his parts have been moved over to a machine which is keeping them functional. At a certain point, however, we come to have a case where a human being has been divided in two: part of (what used to be) him is over here, and part of (what used to be) him is over there. (What we should say about cases like this is something I discuss in Chapter 4.) Eventually, however, these two halves are reunited at the location which was originally occupied by the artificial body. (This sort of situation is generally referred to as "fusion.") So, I think, this is ultimately a case of the sort that I describe in the main text, where some small proportion of the original mass of the entity serves to support the organism's continued existence.

mutilation[94] that criteria for being-the-same-*x* (where *x* is an organism) fail to entail strict identity[95]. In all other cases, if A is the same *x* (organism) as B, then it follows that B is identical to itself; to anything which is identical to it (in particular, it is identical to A); and to anything identical to anything identical to it (in particular, it is identical to anything identical to A). That is, while identity for organisms may be extrinsically determined, such that we need to look at something besides A and B and the causal connections between them to determine whether A is the same as B, it is an *intrinsic* (local) matter whether this is so. We can tell *by looking at the organisms themselves* whether we need to look elsewhere to ascertain identity. Whether we have

[94] But what about cases involving cells which are part of a larger organism, and cases involving entities such as amoebae and worms, all of which regenerate by splitting and for which it seems odd to speak of such cases as "mutilation"?

Replies: (a) Cells can be neglected for the time being on the grounds that they are not strictly speaking *organisms*, but only *parts* of organisms. (And to the extent that they are organisms, the answer I give in (b) applies to them.) (b) Worms and amoebae are organisms, and about them I suppose the right thing to say is this. In part *because* the way in which such organisms reproduce, identity criteria for such entities are unclear, if we apply normal criteria of identity*. So it is not true for amoeba and worms that "it is only under cases of severe mutilation that criteria for being-the-same-*x*…fail to entail strict identity," but the *reason* is not that being cut in half is not severe mutilation for such creatures. Rather, it is because *in general* we do not know what to say about identity over time for such beings, unless we develop special principles meant to apply to entities of those sorts. See also Wiggins 1980, 73 note 20, concerning "*wave, volume of fluid, worm, garden, crystal, piece of string, word-token, machine*" (italics in original).

> * Cf. Wiggins: "One amoeba becomes two amoebas, but 'becomes' receives an analysis making it correspond to ordinary 'becomes' as constitutive 'is' corresponds to the ordinary 'is' of predication and identity. The matter of the original amoeba—the 'it'—is the fusion, or the matter, of the two new ones taken together. There is matter such that first *a* was constituted of it, and then *b* and *c* were constituted of it" (Wiggins 1980, 72 note 18).

[95] That is, they fail to entail that '…is the same *x* as…' is reflexive, transitive and symmetric.

Continues on next page…

to look beyond A and B to ascertain identity is something we can determine *just by looking at A and B*. So rather than saying, as Mackie does, that identity for organisms is specially-secured, we might say instead that identity for severely mutilated organisms is specially-secured[96].

But doesn't this violate conditions of simplicity? Why should we have one set of criteria for applying the concept being-the-same-*x*-as, and another set of criteria for applying the concept being-the-same-mutilated-*x*-as[97]? My reply: Note first that the difference between these two sets of criteria is trivial; all that the second set has that the first set does not is the clause "so long as there are no equally good competitors." Indeed, in some sense the proposal is *simpler* than Mackie's; he suggests adding this caveat to *every* enumeration of identity-criteria for organisms; I propose a way of systematically restricting the need for this qualification. Second, Mackie himself acknowledges that any proposal of this sort will seem somewhat *ad hoc*. Special-securement, after all, is a way of connecting a not-so-strict concept to a strict one; it allows us to acknowledge the ways in which entities that we consider to be individuals over time can grow and change in size and shape and composition, without giving up the applicability of strict identity, which demands both symmetry and transitivity. But the lump under the rug needs to go somewhere: in Mackie's

[96] I discuss such cases (insofar as they concern persons) in the next chapter.

[97] Note that I am *not* claiming that a mutilated *x* is not an *x*. So the criteria for being-the-same-mutilated-*x*-as will, in some sense, be the same as the criteria for being-the-same-*x*-as. What I am suggesting is that in the second case, we need an extra proviso

Continues on next page…

words, “When the concept of one thing of a certain sort is relaxed by allowing growth and so on, it needs to be somewhat arbitrarily restricted elsewhere in order that the resulting ‘thing’ should conform to the strict logic of identity” (Mackie 1976, 171). I contend that my proposed restriction is less arbitrary than Mackie’s.

Let us turn now to the issue of artifacts. Here, in addition to cases of survival-under-mutilation which parallel the case of organisms and which can be accounted for in the same way, we have the additional possibility that two (or more) competing principles of identity may be at play, such that one of the principles is well met by one candidate, and the other principle by another. Clearly, the Theseus case is a case of this sort: the original-planks ship gains candidacy through the maxim that objects can survive disassembly and subsequent reconstruction, and the continuously-repaired ship gains its candidacy through the maxim that objects can survive the gradual replacement of component parts over time. And here, we cannot employ the strategy employed above; it is indeed normal that artifacts undergo the replacement of parts over time[98], and that they are (occasionally) disassembled and reassembled. So it will not do to carve out a special class of artifacts (as we have for organisms) and

to allow us to *apply* the criteria properly. With this clarification in place, the looseness of expression in the text itself should be less confusing.

[98] By “replacement of parts” I mean something over and above the sort of gradual component replacement that is normal for all physical entities whatsoever. Cf. Shoemaker: “People are sometimes become puzzled by the notion of personal identity on being told that during any seven-year period (or so) all the molecules in a human body are replaced by different ones. Clearly, anyone who is puzzled by the notion of personal identity for this reason should be equally puzzled by the identity of dogs and oak trees…[and] rivers, bicycles…and the like” (Shoemaker 1963, 5). For a fascinating discussion of the ways in which this problem troubled the Stoics, see Sedley 1982.

Continues on next page…

say: in those and only those cases do we need to look elsewhere to see whether we have a competitor. For in the case of artifacts, we cannot rule out the possibility of multiple-candidacy merely by observing that the artifact has not undergone fission.

We can, however, rule out the possibility that cases where identity will (as a matter of fact) be specially secured are the rule rather than the exception[99]. Indeed, it is precisely because we cannot tell by looking at an artifact whether we will have to look elsewhere to determine identity that cases where we *do* have to must be exceptional. For if they were not, then we would not in general be able to ascertain artifact identity for artifacts which (like nearly all artifacts) have undergone gradual part-replacement[100]. But I take it that this is something which we do all the time[101/102].

[99] I am thus contending that there are systematic reasons that we find such cases paradoxical. In this, I go one step further than Scaltsas (1980). He contends that, contrary to "the predominate opinion…the example of Theseus's ship as well as other similarly constructed cases are genuine paradoxes" (Scaltsas 1980, 152). But the reason he offers is practical rather than conceptual: "only the need for making such decisions in everyday life will force us to develop a functionally acceptable code of artifact reidentification" (Scaltsas 1980, 157).

[100] Cf. Note 98.

[101] In the sections that follow, I respond to the retort that we cannot, in fact, ascertain artifact identity, either because there are no artifacts (van Inwagen), or because there is no fact of the matter to be ascertained (Parfit).

[102] Cf. Hirsch: "It seems central to the way we think and speak about persistence that we should typically be able to pick out an object and go on to trace its career unambiguously along some space-time path" (Hirsch 1982, 39). And, one might add, that we should be able to do so without keeping our eyes constantly on the object.

It may seem, however, that I am letting my epistemology drive my ontology. What does what we can and cannot know have to do with what there is? That we are unable to observe events that take place outside our light-cone does not mean that they do not occur; that an arithmetical sum is too great to calculate in any human lifetime does not mean that there is no fact of the matter. But in the case of artifacts, the issue is more complicated. I will say more about this question in the pages that follow, but let me here note two things which may make the claim more defensible.

First, I take it as uncontroversial that artifacts are interest-dependent; they depend on the intentions of their maker and the use to which they are put by their employers[103]. Second, even leaving us out of the picture, there is something odd about a conception of the world which makes so much of what happens over-here depend on what happens over-there. It is surely true that Cambridge changes[104] sometimes matter in substantial ways. If I seek to be the tallest person in the room,

[103] Even those who reject the "real" existence of artifacts recognize something to this (Aristotelian) insight. See the discussion of van Inwagen and Rosenberg below (page 2).

[104] Cf. Geach 1972. A Cambridge change is a change which a thing undergoes as a result of satisfying a description at one time which it fails to satisfy at another, even though the thing has undergone no robust or substantial change. Every object undergoes infinitely many Cambridge changes at any given time. For instance, for each event *E* that occurs, every object that exists at the time of *E*'s occurrence becomes an object-that-existed-during-the-occurrence-of-*E*. Indeed, whenever any object moves or changes, all other objects undergo Cambridge changes in relation to that object. And if we allow that the changes in question can themselves be Cambridge changes, then each change ramifies infinitely. (Each Cambridge change in each object brings about a corresponding Cambridge change in every other object, each of which in turn brings about a corresponding Cambridge change in every other object, and so on.)

then your stepping out the door may allow me to satisfy that description even though I have grown no taller. If you double-fault in tennis, I gain the point even though I have done no work. And if *A* and *B* are both candidates for identity with *C*, then *A's* going out of existence may plausibly render *B* identical with *C*, under certain circumstances. But even if such non-local Cambridge-changes *sometimes* matter for identity, it cannot be true that they usually do[105]. Even if our view of the world is radically holistic, we nonetheless think that it is sometimes possible to make judgments about individual instances. But if each claim about here and now required a full inventory of the universe, or even a reasonably thorough cataloguing of some portion of the universe, we would never be able to make such judgments; holism would be tantamount to skepticism. Since I take it that whatever it is that we mean to speak of when we speak of identity over time, it is not something that requires this sort of commitment to a Leibnizian world where every bit of the world is mirrored in every other, I consider any position that would make extrinsic determination the norm to be untenable[106].

[105] Derek Parfit points out that on his view, *all* questions of identity are Cambridge questions. If we define non-Cambridge properties as properties which can play causal roles (in non-intentional contexts), then identity seems clearly to be a Cambridge property. Indeed, it was central to our discussion of the Galileo case (chapter 2) that questions of identity generally go unanswered by the world. I do not mean to deny this in my discussion here. The aspect of Cambridge changes with which I am concerned is their *non-locality*.

[106] This is not to deny that there is an important strand in philosophy (and certainly in religion as well) which stresses our insurmountable ignorance as finite beings.

The purpose of this sub-section has been to clarify what I mean when I say that the case of the Ship of Theseus is an exceptional case. With this in place, let us turn to a number of influential interpretations of the case's import. These interpretations can be fruitfully divided into two categories. Interpretations of the first type attempt to dissolve the problem by showing that although there *seemed* to be something deep at issue here, the puzzle actually rests on false assumptions about the relation of our concepts to the world. Interpretations of the second type attempt to solve the problem by coming up with general principles which can cover the case. After presenting the interpretations, I will try to show why they miss what I think is *really* at issue, and why this is a consequence of mis-generalization from an exceptional case.

3.5 Attempts to Dissolve the Problem

3.5.1 van Inwagen

In *Material Beings,* Peter van Inwagen advances a thesis which he calls "the Denial." The Denial says that "there are no tables or chairs or any other visible artifacts except living organisms" (van Inwagen 1990, 1)[107]. And "if there are no

[107] Or again: "My position vis-á-vis tables and other inanimate objects is simply that there *are* none. Tables are not defective objects or second-class citizens of the world; they are just not there at all…There are certain properties that a thing would have to have to be properly called a 'table' on anyone's understanding of the word, and nothing has all of these properties" (van Inwagen 1990, 99-100). Or yet again: "There are…no tables and chairs, and there are no other artifacts. Artisans do not create; not, at least, in the sense of causing things to exist. They rearrange objects in space and cause bonding relations to begin to hold or cease to hold…between objects" (van Inwagen 1990, 127).

Continues on next page…

artifacts, then there are no philosophical problems about artifacts." In particular, there are no problems with "identity and persistence through mereological change" (van Inwagen 1990, 128). So the puzzle of the Ship of Theseus[108] is easily solved; "there are no ships, and hence there are no puzzles about the identities of ships" (van Inwagen 1990, 128). "All that happens in the story is that planks[109] are rearranged, shuffled, brought into contact, separated, and stacked. But at no time do two or more of these planks compose anything, and no plank is ever a proper part of anything…[T]he story end[s]as it beg[ins]: with no ships at all" (van Inwagen 1990, 129).

Despite his hard-nosed attitude about what there *really* is, in the ordinary conduct of things, van Inwagen is perfectly willing to speak of artifacts[110]. What he suggests is that there is a certain *mismatch* of languages when, on the one hand, we refer to such entities as artifacts while, on the other hand, we insist on holding our talk to standards of precision appropriately reserved for philosophical conversation.

[108] Van Inwagen calls the puzzle of the Ship of Theseus "the greatest and most profound of the classical puzzles about the identity of artifacts" (van Inwagen 1990, 128).

[109] In telling the story, van Inwagen treats planks as "honorary simples." He suggests that one might speak more precisely of "simples arranged plankwise" (van Inwagen 1990, 128).

[110] Van Inwagen suggests that talk of such entities be understood on analogy with talk of the sun's moving. Just as it is compatible with a Copernican world-view to say : "It was cooler in the garden after the sun had moved behind the elms," so too is it compatible with the Denial to say: "Some of my chairs are being re-upholstered in eggplant-colored velveteen." (For further discussion of this analogy, see van Inwagen 1990, 1-2 and 98-107).

He writes: “I am just as willing as you are to use sentences containing ‘house’ or ‘ship’ in the ordinary business of life. But if you begin to insist that the things we are talking about strictly and in every respect conform to such general logical principles as the Law of Excluded Middle, then I shall insist that we have departed from the ordinary business of life, and I shall consequently insist that we adopt a language capable of bearing [such] weight…a language that refers to nothing besides simples and living organisms and abstract objects” (van Inwagen 1990, 131)[111].

But so far, this gets us nowhere in terms of the original puzzle. Presumably, van Inwagen would allow that common parlance permits us to speak of the continuously-repaired ship in the one-sided case (where the original planks are destroyed) as being the same ship as the original Ship of Theseus. And, presumably, common parlance allows us to speak of the disassembled and reconstructed ship in the one-sided case (where no repairs are made) as being the same ship as the original Ship of Theseus. So the question is: what does common parlance permit us to say in the case described by the third version? That is, simply pointing out that there *are* no ships does not help us with what I have identified above as the core of the puzzle, namely that the Theseus case presents us with an instance where a process that is

[111] Again, he offers an analogy: “If you get sticky about strict adherence to the Principle of Noncontradiction–if, that is, you insist that even the surface structure of my sentences never be of the form ‘p and not p’–then I shall stop saying ‘It is and it isn’t’ in response to ‘Is it raining’ and instead talk of mists. If you insist that every piece of language that has the syntactical form of a singular referring expression denote an object and that every predicate expression concatenated with such a referring expression express a possible property, I shall stop saying ‘The average father has 1.3 children’ and say that the number of children divided by the number of fathers is 1.3” (van Inwagen 1990, 131). See also van Inwagen 1990, 98-107.

Continues on next page…

ordinarily identity-preserving (where by this, we mean: identity-preserving in the common parlance sense) is in this instance entity-creating (where again, what we mean by this is: entity-creating in the common parlance sense). Pointing out that, strictly speaking, there are no ships does nothing to help us in making headway with this problem[112].

But where van Inwagen's suggestion is helpful is in freeing us from applying inappropriately strict standards to the answer that we give in this case. Perhaps when we speak of identity in the common-parlance sense, identity can be intransitive, so that the original Ship of Theseus is identical with each of the two final candidates, but they are not identical with one another. Or perhaps when we speak of identity in the common-parlance sense, identity can be partial, so that the original Ship of Theseus is somewhat identical with each of its successors, but not fully so. Or perhaps when we speak of identity in the common-parlance sense, identity can depend on interest, so that the reconstructed ship is identical to the original Ship of Theseus if one is interested in the question for antiquarian reasons, and identical to the repaired ship if one is interested in the question for marinarial reasons. Or perhaps when we speak of identity in the common-parlance sense, identity over time is subject to extrinsic constraints, such that what happens to C can play a role in determining whether A and

[112] The problem I am pointing out connects to a general difficulty faced by such ideal-language approaches. One lesson of positivism's legacy is that replacing natural language with some highly-constrained artificial language is problematic for at least two reasons: (a) that natural language does not seem to be the jumbled mess that radical critics take it to be, and (b) that if it were, it is unclear how we could come up with the conceptual resources necessary to replace it with something more precise.

B are identical, even going so far as to make A identical to B until the moment that C comes into existence.

3.5.1.1 Identity Under a Sortal

One popular way of dealing with the puzzle, which makes use of the insight we have been considering in the last paragraph, is to distinguish the two candidates on the basis of interest. Such a solution can be neutral with regard to the question of whether ships and so on *really* exist. For instance, van Inwagen writes: “Statements that are apparently about the continued persistence of artifacts make covert reference to the dispositions of intelligent beings to maintain certain arrangements of matter” (van Inwagen 1990, 134), while Rosenberg, in direct challenge to van Inwagen writes: “correct answers to questions regarding the identities of artifacts turn on facts regarding the intentional activities of persons” (Rosenberg 1993, 708). So regardless of our ontological commitments, we might say that the original ship is the same as the continuously-repaired ship insofar as the arrangement of matter is maintained in order to promote seaworthiness, and that it is the same as the reconstructed ship insofar as the arrangement of matter is maintained in order to promote antiquarian precision[113].

[113] Mackie suggests that this would be Locke’s solution. He writes: “Locke’s theory would lead us to say that the seaman’s ship [the ship which has resulted from the constant repairs] is the same *ship* as the original one, since there is a continuous ship-history linking them, whereas the antiquarian’s ship is only the same collection of ship components as the original, since what links them is a continuous history of what has for most of the time been a dispersed collection of ship components, and such a dispersed collection is not a ship–particularly when throughout much of this time a large part of this collection has formed part of something else that was undeniably a sea-going ship*” (Mackie 1976, 143).

Continues on next page…

In many ways, this answer is appealing: it seems to capture the reasons *why* we are inclined to take each of the ships in the third version of the story to be viable candidates for identity with the original ship. And it seems to show that the conflict results from a conflict *between* standards, not a conflict *within* a single standard. Consider the story by David Wiggins wherein a monument is to be erected in honor of Theseus, on the top of which will be placed his ship. Wiggins suggests that:

> Surely some people would say that the ship put together from discarded planks was the right one to raise up there. And a dispute might break out about this matter between priests who favoured the working ship and antiquarians who favoured the reconstruction. The difficulty is then a certain incomparability to their positions. It may seem that one party would be looking for an archaeological relic and the other for a functionally persistent continuant; and that the dispute was to be traced to a disagreement about what it is for something to be a sacred ship. The antiquarian who favours the reconstructed ship has a different interest, it might be said, from the priest who favours the continuously repaired continuant. Both are stuck with the qualification *ship*, but they have different interests. (Wiggins 1980, 93-94)[114].

So it seems that we have solved the puzzle: the repaired ship is the same sailing-ship as the original, and the reassembled ship is the same museum-ship, and that's that. The conflict was only apparent; metaphysics has been saved by a couple of carefully-placed hyphens.

*Note, however, that adding the qualification "particularly when…" is not as innocent as it initially seems.

[114] Wiggins does not think that this retelling provides a full solution to the puzzle. He continues: "we must not confuse the fact that it is in some sense a psychological matter whether *we* adopt the priests' or the archaeologists' view with it being a merely psychological matter which one is Theseus's ship–or an arbitrary matter" (Wiggins 1980, 94).

But on reflection, it is not so clear that we really have a solution. For suppose that we return to the first story, where the parts of the ship are replaced over time, and the original planks are somehow destroyed. And now suppose that a monument is to be built in honor of Theseus. Surely the antiquarians would not say: “What a hopeless endeavor! There will be no appropriate ship to put atop the pillar, for the brave ship of Theseus has been destroyed piecewise over the ages[115].” And the reason that they would not say this is *not* because they have, as a second meaning for ‘ship’, the meaning which the priests have. The reason is that even if one’s interests are antiquarian, artifacts can survive the gradual replacement of parts over time.

A similar argument applies in the other direction. Suppose that we return to the second story, where the planks of the ship are disassembled during the winter months. Suppose that some of them are even used as see-saws and balance beams at a large indoor playground. And suppose again that a monument is to be built in honor of Theseus. Surely the priests would not say: “But wait! What do you think you are doing? There’s nothing to put atop the monument. That old ship went to pieces many winters ago.” And, again, the reason they would not say this is *not* that they have, as a second meaning for ‘ship’, the meaning that the antiquarians have. The reason is that

[115] Note that were they committed to the view that the original *matter* had to be the same, then they would be unable to countenance the persistence of *any object whatever*, since all objects undergo a natural process of replacement of parts over time (cf. note 98). Presumably, however, one might insist that this process be a *natural* one.

even if one's interests are functional, artifacts can survive disassembly and reassembly[116].

3.5.1.2 Summary

Let me summarize where we stand so far. Van Inwagen has argued that, strictly speaking, there are no artifacts. When we speak of artifacts persisting, we are really speaking of the dispositions of intelligent beings to maintain certain arrangements of matter. I suggested that we can accept the core of this insight, that the persistence of artifacts is essentially tied to the interests of intelligent beings, while remaining neutral with respect to van Inwagen's ontology. One consequence of endorsing this aspect of van Inwagen's view is that there is a certain weakening of

[116] Nor is the problem that *ship* is somehow ambiguous, in the way that *passenger* is. We might imagine that the ship set sail to Delos in the morning with A, B, C and D aboard, and that it returned in the afternoon with A, C, E, F and G. And we might ask: "how many passengers were there on the Ship of Theseus today?" The answer to this question is *nine* under one construal of *passenger* (A, B, C, and D on the way out, and A, C, E, F, and G on the way back), and *seven* under another (A, B, C, D, E, F, and G).

Could this analogy help us with our two apparent readings of ship? I think not. The problem with the Ship of Theseus is not that, on one reading, ships can be counted on the basis of antiquarian considerations, and on another, they can be counted on the basis of functional considerations. Nor is the problem that we do not have a sufficiently precise idea of how to count ships*. The problem is that, within a certain range, even if we fully specify what we are interested in when we speak of ships, both criteria (disassembly-reassembly and the replacement of parts over time) seem to be perfectly good criteria for identity-preservation.

*Cf. Wiggins 1980, 73: "there is no universally applicable definite way of counting crowns. The Pope's crown is made of crowns, There is no definite

Continues on next page…

strictures concerning what can be said about objects. Among other things, it may be true that there can be two distinct *artifacts* in the same place at the same time, without that implying that single spaces can be multiply occupied by interest-independent entities. This opens up the possibility of thinking of the puzzle of the Ship of Theseus as being a puzzle that concerns an ambiguity in the sortal *ship*, such that on one reading, a ship is something that serves the needs of mariners, and on the other, a ship is something that serves the needs of curators. But I suggested that this is not a satisfactory solution to the puzzle, since even if one's interests are practical, it seems that artifacts can survive disassembly and subsequent reassembly, and even if one's interests are antiquarian, it seems that artifacts can survive the gradual replacement of parts over time.

3.5.2 Parfit

On the surface, it seems that nothing could differ more from van Inwagen's position than a position which grants existence to just about everything. Van Inwagen is extremely ontologically parsimonious, and the position I am about to describe represents an extreme in ontological profligacy. Nonetheless, I will contend, as far as the puzzle of Theseus is concerned, the two positions offer precisely the same (inadequate) resources for providing a satisfactory analysis of the puzzle.

Parfit contends that "if the existence of Xs just consist in the existence of Ys, Xs are not, in relation to Ys, separately existing entities" (Parfit 1994, 9). So, for

answer, when the Pope is wearing his crown, to the question 'how many crowns does he have on his head?'"

instance, suppose that we follow Parfit in saying that "the existence of a library just consists in the existence of a building and a set of books, and in the activities of various people" (Parfit 1994, 3). If we let the Xs be *libraries* and Ys be *buildings, books,* and *activities of various people,* then we have an instance where Xs (that is, libraries) are not separately existing entities in relation to Ys. In a parallel fashion, we might say that the existence of *ships* (Xs) just consist in the existence of *planks (arranged in a certain fashion and used for a certain purpose)* (Ys); so ships, in relation to planks, are not separately existing entities. For the sake of convenience, let us speak of the Ys as the *basis* for the Xs.

Parfit also discusses a category of entities which he calls *notional*. Notional entities are entities which are *mind-dependent*, entities which depend for their existence *as entities* on our thinking of them as such. So, for instance, the constellation *Ursa Major* is mind-dependent in this sense. Even though the stars which serve as its basis would exist regardless of our activities, and even though they would be arranged in the pattern which they are now arranged, the existence of the *constellation* depends upon our thinking of those stars as forming that pattern. What it is for something to be a constellation is for there to be some arrangement of stars in the heavens to which we assign a label; so the very idea of constellation has built into it some sort of observer-dependence, even though the configuration of stars which makes up the constellation does not[117]. Parfit contrasts this sort of mind-dependence

[117] Of course, the stars themselves may be (in this sense) observer-dependent; it may be that we see them as discrete only as a result of our concepts. But I take it as uncontroversial that stars are (in some sense) *less* observer-dependent than

Continues on next page…

with the mind-dependence of entities such as fictional characters. Fictional characters, he suggests, are *brought into existence* by our conceptual activities. It is not a matter of our giving a label to some substrate that exists regardless of our interaction with it; when we invent fictional characters, we create the Xs from scratch. So, he concludes, most 'notional' (that is, mind-dependent) beings are not "brought into existence by our conceptual activities. They would exist whatever concepts we employed. But their existence might be claimed to be, in another sense, conceptual" (Parfit 1994, 10). Ships and forests[118] and routes[119] are notional entities in this sense[120]. Parfit is perfectly willing to grant existence to any notional entity

constellations.

[118] "Suppose I already know that several trees are growing together on some hill. I then learn that because that is true, there is a copse on this hill. That would not be new factual information. I would merely have learnt that such a group of trees can be called a 'copse'. That is a conceptual fact. And it provides a sense in which, in relation to the trees, the copse is a 'notional being'. Though copses do not exist only because of the way we talk, when we learn that copses exist, we may only be learning a fact about how we talk" (Parfit 1994, 10).

[119] Cf. Chisholm 1971, 4-6: "[T]here is no dispute about any observational data. You have agreed about what it is that is called 'Elm Street', about what it is that is called 'Route 42', about the number of lanes in the various places, and about what parts are composed of what. Your dispute, then, has to do with criteria for *constituting the same road*."

[120] There is, of course, a second way in which ships are interest-dependent and forests are not, namely that ships are constructed artifacts. That there are planks assembled in this particular ship-wise pattern is the result of some intentional agent having arranged them that way. I am bracketing this issue for the moment, along with the vexed question of what might be called *found artifacts*, that is, previously existing entities which are put to a particular use by some intentional agent.

whose basis exists. As far as he is concerned, there *really are* ships and forests and routes.

He goes on to point out that in certain cases, even though we are fully cognizant of the underlying facts, we might still ask questions about notional entities. Suppose, for instance, that I gaze out the window and ask: “Is there a copse on that hill?” I might be asking one of two questions. I might be wondering whether the green circle which I see on the hillside is a cluster of trees or a celebratory tent. Or I might be wondering whether those five or six trees which I see over there really constitute a copse. If my worry is the first, I am asking a *factual* question; if my worry is the second, I am asking a *conceptual* question[121].

Now, what Parfit would presumably say about the puzzle of the Ship of Theseus is that the question “which of these ships is the Ship of Theseus?” is a conceptual question. We already know everything there is to know about the underlying facts: we know what happened to each of the original planks, and we know how each of the two competitors came into being. Our worry is not about which of two possibilities will be realized. Our worry is about how we should describe a state of affairs given that we already know everything there is to know about the basis.

But at this point, we find ourselves exactly where we were when we were told by van Inwagen that the worry that confronts us in the case of the Ship of Theseus is

[121] “Even when a question is about reality, the answer depends not only on reality but also on our concepts. If we already know about the former, we must be asking about the latter” (Parfit 1994, 11).

a worry about common parlance. Our rules for answering conceptual questions about ships permit us to speak of the continuously-repaired ship in the one-sided case (where the original planks are destroyed) as being the same ship as the original Ship of Theseus. And, presumably, our rules for answering conceptual questions about ships allows us to speak of the disassembled and reconstructed ship in the one-sided case (where no repairs are made) as being the same ship as the original Ship of Theseus. So the question is: what do our rules for answering conceptual questions about ships allow us to say in the case described by the third version? And simply pointing out that questions about ships are questions about the application of our concepts does not help us with what I have identified above as the core of the puzzle, namely that the Theseus case presents us with an instance where a process that is ordinarily identity-preserving (in these sense that our concepts would lead us to say that an object has persisted) is in this instance entity-creating (in the sense that our concepts would lead us to say that an object has been brought into existence). And from here, the argument continues as in the last section. Even if the question at hand is only a question about our words, it remains puzzling how it is that a practice that is ordinarily such that it allows us to track identity might sometimes allow us only to track identity-candidacy.

3.6 Attempts to Solve the Problem

3.6.1 A Traditional Solution: Hirsch

Traditional characterizations of the puzzle of the Ship of Theseus are characterizations which suggest that the puzzle arises when two criteria of identity are pitted against one another. In particular, the standard presentation suggests that the criterion of continuity of form is pitted against the criterion of continuity of matter, such that our intuitions pull in both directions. Traditional solutions to the puzzle of the Ship of Theseus are solutions which come down in favor of one or another of these two principles, concluding that one or the other (or neither) of the two resulting ships *is* the Ship of Theseus, on the grounds that one or the other (or neither) of the competing principles takes primacy in this case.

So, for instance, Eli Hirsch's interim solution to the puzzle falls into this category[122]/[123]. Hirsch, suggests that the puzzle arises when "considerations of compositional similarity and continuity yield incompatible identity judgments" (Hirsch 1982, 67), and he comes down, albeit somewhat tentatively, in favor of the constantly-repaired ship.

In order to understand the remainder of Hirsch's discussion, we will need two pieces of terminology. Hirsch's sortal rule (sufficient condition version) says:

[122] I discuss his considered solution in section **3.7** below.

[123] For other examples of discussions that take place within this tradition, see Smart (1972), Dauer (1972), Davis (1973), Smart (1973), and Scaltsas (1980).

> *The Sortal Rule*: A sufficient condition for the succession S of object-stages to correspond to stages in the career of a single persisting object is that:
> (1) S is spatiotemporally continuous[124]; *and*
> (2) S is qualitatively continuous; *and*
> (3) There is a sortal term F such that S is a succession of F-stages
> (Hirsch 1982, 36)

And his compositional criterion (second formulation) says:

> *The Compositional Criterion:* Where *x* is an object that exists at time t_1 and *y* is an object that exists at a later time t_2, a sufficient condition for *x* to be identical with *y* is that the same sortal is predicatively true of both *x* at t_1 and *y* at t_2, and some set of objects comprises a major portion of both *x* at t_1 and *y* at t_2, and this set of objects is similarly arranged in both *x* at t_1 and *y* at t_2 (Hirsch 1982, 65).

Informally, the Sortal Rule says that an object can persist through changes so long as the spatiotemporal continuity is not disrupted, and so long as the changes it undergoes are characteristic of entities of a certain sort, of which this object is one. The Compositional Criterion says that an object can persist through disassembly and reconstruction, so long as the earlier and later candidates are both composed of roughly the same matter arranged in roughly the same way. So the Sortal Rule can be seen as a more precise expression of what we have been calling the second principle, and the Compositional Criterion can be seen as a more precise expression of what we have been calling the first principle.

After presenting the story of the Ship of Theseus, Hirsch writes: “My own somewhat ambivalent inclination when reflecting upon this case is to judge that [the original ship] is identical with the [constantly-repaired ship] and not with the” reconstructed ship (Hirsch 1982, 69). He continues: “If this intuition is generally shared there would be two related ways to explain it. It will be noted that the judgment that [the original ship] is identical with [the constantly-repaired] follows from the sortal rule as well as from the compositional criterion, whereas the judgment that [the original ship] is identical with [the reconstructed ship] follows only from the compositional criterion…We might say that…where the sortal rule conflicts with the compositional criterion the former rule, which we know to be primary, takes precedence. Or, forgetting about primacy, we might simply say that the sortal rule in conjunction with the compositional criterion outweighs the latter standing alone" (Hirsch 1982, 69). So Hirsch's tentative solution to the puzzle is to grant identity with the original ship to the constantly-repaired ship, and to do so on the grounds that the criteria by which the latter can be said to be identical with the former are criteria which more precisely characterize the grounds on which identity over time can be said to obtain[125].

How successful is this solution, given our initial question? The initial question, you will recall, is what should be concluded from the fact that the Ship of Theseus story seems to present us with the story where a process which usually identity-preserving is in this case entity-creating? What the traditional answer says is this: "Yes, such a process *is* usually identity-preserving. But it is not identity-preserving here. And the reason it is not identity-preserving is that there is another,

[124] A characterization of this notion can be found at Hirsch 1982, 15-22.

[125] Cf. Davis (1973): "We are left with the conclusion that spatio-temporal continuity of form…is the primary condition for identity of ships, and perhaps for all spatio-temporal objects" (Davis 1973, 110).

better process which gives us another, better candidate." So the traditional answer is tantamount to saying that in this case, the compositional criterion is not really a criterion of identity; it is merely a criterion of identity-candidacy, a second-best way of being identical over time.

Such a solution is unsatisfying because it is either too general or too *ad hoc*. It is too general if what the solution says is that whenever an object persists by means of the compositional criterion, all we can conclude is that we have an persistence-candidate, for this is surely false[126]. But if the solution says that there is something about *this* case of conflict which gives the sortal principle criterion primacy over the compositional, then it is too *ad hoc*, for there is nothing explanatory about such a diagnosis. And such a criticism will apply to any solution which gives priority to one or the other principle without identifying some dynamic relation between them; either it will falsely denigrate the status of the principle which fails to receive priority, or it will fail to explain why *this* case is a case where the principles in question conflict.

But as before, there is an aspect of this proposed solution which gets at something deep about the puzzle of the Ship of Theseus. It is true that there is something odd going on when the two principles come in conflict with each other. And it seems right to say that but for the candidate presented by the second principle, the process identified by the first principle would be identity-preserving. So there is

[126] And it is also too general if the solution says that whenever the two rules conflict, we should choose the sortal rule, for then the position is subject to the sorts of objections that I make against the closest continuer theory (see section **3.6.2.2** below).

something in the interaction between them that is causing the conflict. To this extent, the traditional solution is after something deep. But it does not yet offer the full answer.

3.6.2 A Meta-Solution: Nozick

In *Philosophical Explanations,* Nozick contends that "to be something later is to be that thing's closest continuer." What this means is that to be some (physical) thing later is to be the entity that most closely meets the set of criteria by which the original entity may be generally said to continue over time. So to be E at t_2 is to be the entity which best matches those characteristics which, at t_1, would be expected to hold of an entity which followed the natural trajectory of the (ideal) E from t_1 to t_2, and which matches those characteristics for reasons that depend causally on E[127].

So, for instance, let us suppose that the Vienna Circle in the mid-1930s had twenty members, who met weekly in Vienna to discuss philosophy[128]. According to the natural trajectory of the ideal Circle, we might expect that in 1945 it would have the same twenty members, that these members would live in Vienna, and that the group would have met weekly to discuss philosophy during the intervening decade. Instead, however, what we find in 1945 are the following candidates for identity with

[127] "To say that something is a continuer of x is not merely to say its properties are qualitatively the same as x's, or resemble them. Rather it is to say they grow out of x's properties, are causally produced by them, are to be explained by x's earlier having had its properties, and so forth" (Nozick 1980, 35).

[128] The example is from Nozick (1980), 32-33; I have added certain details.

the original Circle: (a) a trio of members who spent the War in Istanbul, meeting weekly; (b) two members who spent the War hiding in Vienna, who met only twice during the period from 1939 to 1945; and (c) nine members who spent the War in the United States, meeting twice each month.

Each of these candidates meets some of the initial criteria; in particular, (a) met weekly during the decade from 1935 to 1945, and (b) has members who live in Vienna. Nonetheless, it seems that (c) has the best claim to the title “Vienna Circle.” The group, which was composed of a good proportion of the original members, met regularly to discuss philosophy; in this light, the Vienna criterion seems to hold less weight. However, as Nozick points out, it also seems right to say that, were (c) not to exist, (a) would be the Vienna Circle. “If no other group exists, the Istanbul group is the closest continuer; but if the group in the United States exists, *it* is the closest continuer” (Nozick 1980, 33). Since what it is to be the Vienna Circle at t_2 is to be the closest continuer of the-Vienna-Circle-at-t, then the Istanbul group would, but for the existence of its New World competitor, *be* the Vienna Circle.

Let us look first at how this theory applies to the case of the Ship of Theseus, where it seems initially to account for out intuitions extraordinarily well, and then examine some of its counterintuitive implications. Since the closest continuer theory is a *schema*, it does not tell us which of the two competitors in the third version *is* Ship of Theseus. What is does instead is to provide us with a framework for making sense of the puzzle[129]. The theory explains why we should expect to encounter cases with the structure of the Theseus case, that is, cases where information about the existence of some object C leads us to deny the status of continuer-of-A to some wholly other object B. Nozick writes:

> In the case of the ships, there are two relevant properties: spatiotemporal continuity with continuity of parts, and being composed of the very same parts (in the same configuration). If these have equal weight, there is a tie in closeness of continuation. Neither, then, is the closest continuer, so neither is the original ship. However, even when the two properties receive equal weight, if there had actually been one ship existing without the other, then it, as the closest continuer, would be the original ship…[So] the closest continuer schema does fit and explain our response to this puzzle. When we hear the first story…we are not puzzled or led to deny that it really is the same ship. Only when we learn of the reconstituted ship are we thrown into puzzlement, not only about its status, but about the earlier product of gradual rebuilding. *It is only when we learn of another candidate for closest (or equally close) continuer that we come to doubt that the gradually altered ship is the same ship as the original one* (Nozick 1980, 33-34, italics added).

There is something extremely appealing in this diagnosis. It *does* seem that it is only when we learn of the existence of the second ship that we come to doubt the status of the first. Moreover, such a theory helps refine the sortal-relative solution presented above in the context of Wiggins's story of the priests and the antiquarians[130]. As I pointed out above, the problem with the diagnosis that suggests there are two different concepts of *ship*, one antiquarian and one practical, is that the diagnosis cannot explain why subscribers to either concept would be inclined to

[129] "The closest continuer theory does not, by itself, answer the question…[but] it helps to sort out and structure the issues" (Nozick 1980, 33).

[130] The reader will recall that the priests are those who believe that the Ship of Theseus is the repaired ship, on the grounds that they are concerned with sailing-

Continues on next page…

accept the validity of the alternative. But the closest continuer theory can, and it can do so while preserving the insight that part of what is going on is that there are various interests at play. What the closest continuer theory can say about the priests and the antiquarians is that they have different *rankings* of the criteria for continuity. Each side recognizes that the other's principle has *some* force, but the one side thinks reassembly trumps replacement, while the other thinks the opposite.

But despite its considerable appeal in this case, the closest continuer theory has at least three extraordinarily counterintuitive implications. And these counterintuitive implications are, I think, are sufficient to override its otherwise alluring aspects.

First, on the closest continuer theory, identity over time is not transitive: A at $t_1$may be identical with B at t_2 and with C at t_3, but B and C may not be identical. Suppose, for instance, that A is our original entity. And suppose that, at t_2, B is its closest continuer. So at t_2, B *is* A. Suppose, however, that at t_3, *C'* is A's closest continuer, but that the closest continuer of B is B'. Then at t_3: C' is A, and B' is B. But B' is not A[131].

An example may help to make this clearer:

t_1: Let A at t_1= {Albert, Bernadette, Camille, Dorothy, Ella, and Fred}.

t_2: Let B at t_2 = {Albert, Bernadette, Camille}

ships, and that the antiquarians are those who believe that the Ship of Theseus is the reassembled ship, on the grounds that they are concerned with museum pieces.

[131] For another example of this sort (which differs in that A has no continuer at t_3), see Nozick 1980, 659, note 10.

Let C at t_2 = {Dorothy, Ella}
Let D at t_2 = {Fred}.

t_3: Let B' at t_3 = {Albert, Bernadette, George}
Let C' at t_2 = {Camille, Dorothy, Ella}
Let D' at t_3 = {Fred}

At t_2, B is A. It is a good enough continuer of A. And it is the closest continuer of A, since it has three of A’s original members, while C and D have only two and one respectively. At t_3, however, *C'* is A. It is a good enough continuer of A. And it is the closest continuer of A, since like B, it has three of A’s original members. But C' is *not* the closest continuer of B; it shares only one member with B (Camille), while B' shares two (Albert and Bernadette). So we have the following relations

At t_2 : B = A
At t_3 : B '= B
At t_3 : C' = A
At t_3 : C' ≠ B

That is, at t_3, C' is not identical to B, but it is identical to A, which is (at t_2) identical to B[132].

The second counterintuitive implication of the closest continuer theory is that objects may move instantaneously from one location to another, without being anywhere in between: again, A at $t_1$may be identical with B at t_2 and with C at t_3, where B and C are spatially distant from one another. Continuing with the previous example, let us label *A* “The Club”, and let us suppose that, at t_3, The Club is in Idaho, where we find Camille, Dorothy and Ella. At t_4, the three die tragically in a

[132] Nozick, of course, is well aware of this implication. He writes: “this…relation need not be transitive,” and goes on to present four ways that “a view of identity

Continues on next page…

boating accident, and B' (that is, the group: Albert, Bernadette and George) *becomes* The Club. But B', let us suppose, is in Indiana. So at the moment of the boating accident, The Club moves from Idaho to Indiana, instantly and without being anywhere in between.[133/134].

Third, according to the closest continuer theory, events in the future may change events in the past[135]: if, at t_3, B becomes identical with A, then, for instance, it becomes true of B that B (as A) came into existence at the time A came into existence, that B in its past underwent experiences that A underwent, and so on. And

[might] cope with these nontransitivities of…closest continuer" (Nozick 1980, 42-43).

[133] As before, Nozick is well aware of this implication. He writes: "It [does] seem[] strange that at a certain time, without any (physical) change taking place in it, the new [entity] could become the [original] when the [original] expires. However, once we have become used to the idea that whether y at t_2 is (identical with) x at t_1 does not depend only on the properties and relations of x and y, but depends also on whether there exists a z of a certain sort (which more closely continues x), then perhaps we can swallow this consequence as well" (Nozick 1980, 43-44). He adds: "This instantaneous movement of a person from one place to another does not violate special relativity's constraint[s]" (Nozick 1980, 660, note 13).

[134] One might put the situation less tendentiously by saying: "At t_1, A was at location L, and at t_2 A was at location M, and A was nowhere in between." That is, one need not speak (except out of habit) of A having *moved* from L to M; one might take the consequence of the theory to be that things might appear at places without having *moved* to those places.

[135] Note that there is a trivial sense in which even ordinary theories of identity over time allow for the possibility that events in the future might change events in the past; anything that happens to A-later makes it true of A-earlier that A-earlier was an entity to which such events will happen. For instance, suppose that A-later undergoes P at t_2. Then this event (which happens at t_2), makes it true of A-earlier that it will at some time be an entity which has undergone P. And this also makes it true of A-later that it is an entity with a predecessor (A-earlier) that has the property being-something-that-

Continues on next page…

this may turn out to be true of B because of something that happens to *C.* So, for instance, suppose that at t_3, B' (that is, Albert, Bernadette and George) realizes that all it has to do to be The Club is somehow to get rid of the Idaho contingent. So it arranges for the group to set out in a leaky boat, which, as we learned in the last paragraph, sinks at t_4. This allows the Indiana contingent to hoist the flag high and declare themselves The Club. But by doing so, the trio changes the past. They *make it true* of themselves that they are a club which has existed *since t_1* (the time that the original six came together), and they do with no substantial changes to themselves; everything that happens (in the ordinary sense of the term) happens to poor C[136/137].

Since one of the primary appeals of the closest continuer theory is its ability to account for our ordinary intuitions about identity over time, the fact that it produces such counterintuitive results is worrisome. If saying that the Istanbul group would be the Vienna Circle if the United States group had not existed means saying that

will-at-some-time-undergo-P. As will be clear from the example in the text, this is not the sort of case I am thinking of.

[136] Mark Johnston describes another such case. Following Williams, he supposes that there is a machine which has "read" A's psychology into B's body, and which is starting to "read" the psychology into C's. The B-body person, who is walking around thinking: "I am A. I did not just come into existence," sees the machine implanting the information in the C-body person, and realizes that if the process is completed, the C-body person will be a better continuer of A, and will, by the closest continuer theory, *be* A*. So he shuts off the transmitting machine, and thereby *makes it true* that "he is A and has existed before the operation of the machine as A. But…surely our intuition is that the B-body person's thought, 'I am A. I did not just come into existence,' is made true or false by what has happened up to and including the time at which that thought occurs. Surely no subsequent act by the B-body person can make this true or false" (Johnston 1987, 68-69).

*See Nozick's discussion of "Overlap" at Nozick 1980, 43-47 and 660 note 13.

identity over time is intransitive, that objects can move instantaneously from one place to another, and that events in the future can change events in the past, then perhaps it makes more sense to change some of our views about the Vienna Circle[138]. For if *that's* what it means to say that to be something later is to be that thing's closest continuer, then maybe we weren't committed to such a view after all[139].

Still, the counterintuitive implications of the view might be tempered somewhat if we combine the closest continuer theory with something like van Inwagen's or Parfit's view, where we recognize that the claims we are making about identity are claims about common parlance, or about how we apply our concepts. So, for instance, if the question we are asking is what Parfit would call a *conceptual question* (see section **3.5.2** above), then perhaps it is not so remarkable that our descriptions permit the sort of jumping about and intransitivity that we have been discussing in the last few pages. Again, if the question we are asking is not about what there is, but about how we speak about what there is, then perhaps it is not so remarkable that our descriptions permit us to attribute qualities to entities in the past on the basis of what happens to other entities in the future. But if there is some way of

[137] See also the discussion of Cambridge changes above (section **3.4.2**).

[138] Note, however, that there is a complicated interplay among the counterintuitive implications that I have enumerated, such that one might be able to block the problem of changing the past by strengthening the force of non-transitivity. (That is, one might say that it simply isn't true of B that things that are true of A are true of it, even if B is (the closest continuer of) A.)

[139] Nozick's solution seems to be to say that we could "get used to" such implications, once we recognize that they follow from the extrinsic determination of identity (see, for instance, Nozick 1980, 43-44). But since I think there is a better

Continues on next page…

capturing what is true in the closest continuer theory without having to countenance such implications, then we will be better off.

3.7 The Proposed Diagnosis

Although I think the puzzle of the Ship of Theseus is a deep and interesting one, I think it is ultimately tractable. And I think the source of the tractability lies in recognizing something about where the extrinsic determination is properly located in this case. What the puzzle shows is not that *identity* is extrinsically determined (at least not in the local sense that proposals like the closest continuer theory suggest), but that the *processes by which we trace identity over time* are extrinsically determined. What I will argue is that the process of disassembly and subsequent reconstruction in the second version of the story and the process of disassembly and subsequent reconstruction in the third version of the story are *different* processes—even though their intrinsic character is the same. Once we recognize this, we can help ourselves to the insights of each of the positions described in sections **3.5** and **3.6**, while avoiding their drawbacks. To see why this is so, we will need to reexamine the two principles which together lead to the perplexity. I will suggest that there are features internal to the principles themselves which will guide us in properly understanding their interrelations.

3.7.1 A Messier Puzzle

place to locate the extrinsic determination, I think we can avoid the need to get used

Continues on next page…

Imagine the following Theseus-like case, where the entity in question is a pickle-barrel which I have inherited from my beloved grandfather. Each year, in order to insure that the barrel lasts a long time, I replace one of the wooden planks from which it is made with a new one. I hire the same company every year, and unbeknownst to me, they collect the planks I have given them and store them in their warehouse. One year, I replace the last of the original planks: as far as I'm concerned, I still have my grandfather's pickle barrel—the replacement of parts has been gradual, and I have done nothing to the barrel that he would not have done himself. Indeed, during the years that he owned the barrel, he himself replaced all of the planks several times over. He hired the same company every year, and unbeknownst to him, they collected all the planks he gave them over the years and stored them in their warehouse.

One day, while practicing his clarinet, the son of the owner of the plank-collecting company comes up with the marvelous idea of making pickle barrels out of the planks languishing in the company's warehouses. Starting with the oldest planks, he assembles three pickle-barrels, each with a form exactly like the form of the pickle barrel that had belonged to my grandfather. But the similarity does not stop there. The first pickle barrel he assembles has exactly the same planks and exactly the same shape as the first pickle barrel my grandfather had ever owned, back in 1921 when Gendler's Groceries first opened its doors to the booming community of Chariton, Iowa. The second pickle barrel has exactly the same planks and exactly the same

to these consequences, while preserving the central insight.

shape as my grandfather's pickle barrel in 1947, which he rolled out of the store and into his kitchen before the store was sold to some fancy investors from Des Moines. And the final pickle barrel he assembles has exactly the same planks and exactly the same shape as the pickle barrel that I inherited from my grandfather in 1977, when the preconditions required for the concept of inheritance to apply were, alas, made manifest.

Knowing that I am an aficionado of such things, the plank-collector's son comes to me with these three reconstructed vintage pickle barrels, and asks me whether I want to buy any of them. I do, and in the course of negotiations inquire as to the provenance of the materials used to construct them. As you might imagine, the consternation with which I greet his reply is profound.

The story I have told, of course, is a version of classic story of the Ship of Theseus, with a small twist: this version is, one might say, double-barreled. In the tale I have told, there are *four* plausible candidates for the title "my grandfather's pickle barrel:"

(a) the barrel I had in my house before the arrival of the plank-collector's son

(b) the barrel reconstructed from the planks that made up the barrel I originally inherited in 1977

(c) the barrel reconstructed from the planks that made up the barrel my grandfather rolled out from his store in 1947

(d) the barrel reconstructed from the planks that made up the pickle barrel my grandfather purchased in 1921.

What I mean by “candidate” is this: but for the existence of the other three, each of the pickle barrels I have described might plausibly have been called “my grandfather’s pickle barrel.”

The two principles of survival over time which together are sufficient to create this sort of puzzle case are the familiar principles we have been discussing all along:

(1) objects can survive disassembly and subsequent reconstruction

(2) objects can survive the gradual replacement of component parts over time

Barrels (b), (c) and (d)—the reconstructed barrels—are candidates for identity with the original only if principle (1) applies to them; barrels (a), (b) and (c)—the barrels that have components that were not part of the original barrel—are candidates for identity with it only if principle (2) applies to them. So barrel (a)’s candidacy makes appeal only to principle (2), barrel (d)’s candidacy only to principle (1), and barrels (b)’s and (c)’s candidacies both to principles (1) and (2). Suppose now that barrels (a) and (d) are destroyed, and we are left only with barrels (b) and (c). It seems that we have no grounds for choosing between them. Moreover, it seems clear that there could be arbitrarily many such candidates of such a kind. So if such a process is indeed identity-preserving in the absence of competitors, it allows for cases far messier than the original Theseus case. If we maintain unbridled commitment to both principles, then to ascertain whether some later entity is identical with some earlier one, we would need to locate every piece of matter that had even been part of the original entity, and ensure that it is not part of a reconstructed object made from pieces that were at some time simultaneously part of some entity identical to the original.

3.7.2 Process-Dependency and the Two Principles

Since the puzzle seems to arise as the result of a certain intertwining of the principles, let us try to figure out whether the principles as stated actually capture the commitments we wish to be making. I will consider each principle in turn.

3.7.2.1 The First Principle

As stated, principle (1) is a *process-independent* principle. It says that objects produced by configuring the original component parts of some entity in a shape identical to the original entity can properly be said to be identical with the original entity, regardless of the process by which the disassembly and reassembly take place. Principle (1) makes appeal to two sorts of considerations:

(a) *formal*: the entity in question must be made up from matter arranged in thus-and-such way

(b) *material*: the entity in question must be made up from thus-and-such matter

Neither (a) nor (b) alone is sufficient for identity: a structurally-identical object composed of entirely new matter is not a live candidate for identity with the original object, nor is an entity which is composed of the same matter arranged in an entirely different way. (The relation of this to principle (2) will be discussed in a moment.) But (a) and (b) together suggest a *prima facie* plausible standard for identity over time of a complex object:

> Weaker principle: any object composed of the *same matter* as the original object arranged to have the *same structure* as the original object is the *same entity* as the original object.

Put this way, the principle seems extremely plausible. What else, after all, could be relevant to the situation? Here we have the same stuff in the same shape; we must have the same entity.

Despite its plausibility, the principle fails both as a necessary and as a sufficient condition of identity over time for middle-sized objects. It fails as a necessary condition because slight changes in form and slight changes in matter do not automatically disqualify an entity at t_2 as a proper candidate for identity with some object at t_1, so long as a number of other constraints are met. And it fails as a sufficient condition because it is possible that the same matter in the same form might fail to constitute the same entity over time, either because it does not make sense to speak of that collection of matter in that form as constituting an entity, or because the causal relations that connect the two are insufficient[(i)].

Although the weaker principle is not a sufficient condition for identity over time, a slightly strengthened version of the principle expresses the ideal for identity over time in a static world.

> Stronger principle: any object composed of the *same matter* as the original object, arranged to have the *same structure* as the original object, whose history is *spatiotemporally continuous* with the original object, is the *same entity* as the original object.

The stronger principle is certainly sufficient for identity[140]. But as a characterization with aspirations for application to normal cases, it is utterly inadequate. Actual entities are responsive to their environments: they grow; they decay; they interact with other entities. The stronger principle describes what identity-conditions might look like in a static world. But criteria of identity over time are attempts to provide us with guidelines for navigating a world of flux. As it stands, the first principle is not such a criterion.

3.7.2.2 The Second Principle

Unlike the first principle, which looks only at the beginning- and end-states to ascertain whether the specified criteria have been met, the second principle is concerned with the sorts of *processes* to which an object can be subjected while maintaining identity. In particular, it is concerned with the process that involves the

[140] Assuming, that is, that objects are such that they persist in a spatiotemporally continuous way. If we consider grue-like entities, this is not the case. Let us define the (grue-like) *Red House* as follows:

> The *Red House* is that entity which is the weekday-portions of the White House, and the weekend-portions of the Kremlin*.

The stronger principle is *not* a sufficient condition for identity for the Red House. To see this, let us take as our "original object" the Red House on Sunday (translation: the Kremlin). On Monday, the object which meets the identity-criteria specified by the stronger principle is the Kremlin. But despite meeting the criteria, the Kremlin is *not* identical with the earlier object. (That honor goes to the Red House on Monday, translation: the White House.)

As should be evident from my discussions throughout this chapter, I am not attempting to offer a theory of identity which would cover such gerrymandered entities. (For further examples of entities of these sorts, see the opening chapters of Hirsch 1982)

> *In both cases, we'll use Greenwich Mean Time to determine the start of the day.

form-preserving gradual replacement of parts over time[141]. What the second principle says is that so long as the replacement is gradual enough, so long as the replacement respects the internal structures of the entity in question, gradual changes in matter might be made with no loss of identity over time[142].

This principle rests on the assumption that there are certain normal processes which entities of one or another kind can undergo, processes which respect certain patterns in the world which we pick out as significant, and for which we recognize certain sorts of transformations as normal and identity-preserving and others as abnormal and identity-destroying. The specific features of the pattern in question depend on what sort of entity we are considering[143]; a lump of clay can survive changes that a statue cannot, a polygon can gain sides that would destroy a rectangle.

[141] I am here concerned with replacement of parts over and above that discussed in footnote 98 (that the exchange of matter every seven years is a process undergone by every physical object). One might express principle (2) more precisely as: "objects can survive the gradual replacement of parts over time, over and above that which occurs at the micro-structural level as a result of sub-molecular exchanges of particles."

[142] For a discussion of these matters as they relate to issues of personal identity, see Unger 1990, 123-125.

[143] Cf. Aristotle: "Some things are characterized by the mode of composition of their matter, e.g. the things formed by mixture, such as honey-water; and others by being bound together, e.g. a bundle; and others by being glued together, e.g. a book; and others by being nailed together, e.g. a casket; and others in more than one of these ways; and others by position, e.g. the threshold and the lintel (for these differ by being placed in a certain way); and others by time, e.g. dinner and breakfast; and others by place, e.g. the winds; and others by the affections proper to sensible things, e.g. hardness and softness, density and rarity, dryness and wetness; and some things by some of these qualities, others by them all, and in general some by excess and

Continues on next page…

But what is crucial about the second principle as far as our argument is concerned is that it brings out the ways in which the identity criteria for object *A* over time depends on all sorts of facts about objects other than *A*. Identity criteria over time depend on there being certain sorts of salient *patterns* by which we classify entities[144], and by which we make judgments about what sorts of changes entities of a certain sort can undergo[145]. Identity in this sense *is* extrinsically determined; there would be no such thing as being the same ship unless *ship* were something an entity might be. So although it is true that identity for particular objects cannot be, as a matter of course, extrinsically determined (for the reasons discussed above in section **3.4**), it is *not* true that the *processes* by which identity is ascertained cannot themselves be extrinsically determined.

3.7.3 Recapitulation

Let me recapitulate before going on. In section **3.7.1**, we learned that the two principles together actually permit puzzles much messier than the original puzzle of the Ship of Theseus; they permit cases where, once we eliminate entities whose candidacy depends on one or the other of the principles alone, there can be arbitrarily

some by defect. Clearly the word 'is' has just as many meanings" (Aristotle, Metaphysics 1042^b 15-25).

[144] For a fascinating discussion of these issues, see Peirce's series of 1878 essays, including "The Doctrine of Chances," "The Probability of Induction," "The Order of Nature," and "Deduction, Induction and Hypothesis." (See Peirce 1867-1893/1992, 142-199).

[145] Cf. Hirsch's Sortal Rule, as presented in section **3.6.1** above.

many candidates for identity with the original object. We then noted that the first principle, as formulated, is a process-independent principle; it specifies identity by considering how it is that a certain collection of matter is arranged at a certain point in time. And we noted that, as stated, the principle offers neither necessary nor sufficient conditions for identity. There is, however, a strengthening of the principle which *does* provide sufficient conditions. This stronger condition, however, is unable to account for the fact that we live in a world of flux and change.

We went on to note that the second principle is *not* process-independent; it specifies identity by looking at certain spatiotemporal and causal connections that link one object to another. We also noted that the second principle relies on there being certain sorts of patterns in the world, patterns upon which we depend for distinguishing between connections that "count" and connections that do not.

What I will argue is that this sort of process-dependency and extrinsic determination is true of the (correctly-stated) first principle as well. And once we recognize this, we will see that the (modified) principle does *not* countenance multiple-candidacy. And the same modification that allows us to block multiple-candidacy in the application of the first principle on its own allows us to block it in the case of the two principles intertwined. That is to say, it allows us to block the paradox that is posed by the case of the Ship of Theseus.

3.7.4 Conclusion

In presenting the proposed solution, let us return to the initial presentation of the puzzle in sections **3.2** and **3.3**. I suggested there that the story of the Ship of Theseus raises the puzzle of how a process that is ordinarily identity-preserving can in some cases be entity-creating. In particular, at least one of the two ordinarily identity-preserving processes we have been discussing in this chapter (disassembly-reassembly or gradual part-replacement) is, in the case of the Ship of Theseus, entity-creating.

But what should now be apparent is that to put the question this way is to presuppose that the process of part-replacement is the same process in the first version and the third version of the story, and that the process of disassembly-reassembly is the same process in the second and the third. And the obvious question to be asked is: why should we think that the process of part-replacement (or disassembly-reassembly) which is entity-preserving in the first (or second) version is the *same* process as that which takes place in the third version? That is, why should we think that processes are *intrinsically determined* in this sense? Why should what it is for a process to be a process of (say) disassembly and reassembly of the relevant sort be only a matter of what happens to the particular planks which are subject to being detached from one another, and then reattached to one another? Why might it not instead be that for such a process to be the relevant sort of process (that is, the same sort of process as the one described in the second version), certain things have to be true of *other* planks as well? In particular, perhaps to be the relevant sort of process, it must be true not only that the planks in question are detached from one another in a certain sort of gradual process, but also that no *other* planks are attached to these planks during the course of detachment? Or, to take the parallel case, perhaps to be the relevant sort of process, it must be true not only that the original planks are replaced sequentially by a set of planks of similar size and shape, arranged in the same manner, but also that the planks which are removed in the course of such a process be disposed of in some way which precludes reassembly[146]?

With this in place, let us build our solution with pieces culled from the various solutions discussed in sections **3.5** and **3.6**. From the attempts at dissolution discussed in sections **3.5.1** and **3.5.2**, we will help ourselves to views of identity over time for artifacts that need not conform to standards which are too strict for entities in our changing world. From the traditional solution of section **3.6.1**, we will adopt the insight that the second principle trumps the first; when there are two candidates for identity with an earlier object where one is produced by a process of gradual replacement of parts over time and the other produced by a process of disassembly and subsequent reassembly, the former will rightly be considered identical with the original object. And from section **3.6.2**, we will take on the *reason* this is so: it is the *interplay* between the two principles that explains why the second trumps the first; our commitments concerning identity over time do follow, to some extent, a closest continuer schema.

But our proposed solution differs from all of these in locating the source of the problem in the specification of the *processes* by which identity is maintained.

[146] Cf. Smart (1972) and Smart (1973).

When a part is removed from an object and not replaced by a new part, the part remains a candidate for later reconstitution of that same entity. This allows for identity-preserving disassembly and reassembly of watches and automobile engines of the sort we would like to countenance. But in other cases, when the process of disassembly takes place in a particular context, the subsequent reassembly of previous components will *not* produce a candidate for identity. That context is the context where a part that is removed is replaced by a new one.

In content, this solution is quite similar to that proposed by Hirsch. He suggests that the Compositional Criterion be modified to include the clause that “there is no object *z* such that *z* comes into existence at a time *t’* between *t1* and *t2*” which is itself a candidate for identity according to the Sortal Rule or the simple version of the Compositional Criterion (Hirsch 1982, 71). But Hirsch’s solution is *ad hoc* in the sense that it does not explain *why* (except as a means of ruling out multiple-candidacy) such a modification should be added.

To locate the extrinsicness in the process, however, is to help ourselves to a constraint which is already at play when we think about identity over time. Just as we assume certain background patterns when we apply the second principle to cases of part-replacement over time, so too do we assume certain background patterns when we apply the first principle to cases of disassembly and reassembly. And these background patterns include certain assumptions about what sorts of processes the entity from which the parts are removed will undergo. When these processes are exceptional, as they are in the story of the Ship of Theseus, the principles cannot be applied straightforwardly. But they can be applied in the modified manner described in this section.

Endnote to Chapter 3

[i] Locating the extrinsic determination here is far less disruptive than locating it in any of the places proposed above. And doing so allows us to treat the case of the Ship of Theseus as an exceptional case which does not license radical revisions in our conceptual scheme. The failure of the weaker principle as a sufficient condition can be seen from the following example. Suppose a child, Sylvia, has a set of 26 alphabet blocks, and on Monday she arranges the F-O-R-K blocks into a row on a template that holds them solidly in place. Suppose she likes this arrangement very much, carries it into her mother's study, and dubs it: "Look Mom—this is *Fork*!" she says proudly. Monday night, Sylvia puts the blocks back into their storage box, and they stay there until Sunday, when her two-year-old cousin Elmer comes to visit. Being generous, Sylvia lets her cousin play with the alphabet blocks and template, and while mucking around in his two-year-old sort of way, Elmer happens to put the blocks F-O-R-K in the template, in such a way that an object with the same form and same matter as Fork comes to occupy part of the living room. Is Fork back?

I think the answer is *no*, and for a very good reason: there is something missing from the weaker principle, namely some sense of *dependence*. Simply arranging matter that happened once upon a time to be part of a particular object into a configuration that happened once upon a time to be the structure of the same particular object is not enough to get us that object back. To "count as" the same object, the *reason* the new object has the matter it does in the shape it does must be *because* the old object was composed of that matter in that form.

Suppose instead of Elmer putting the F-O-R-K blocks in place on the template, Sylvia had done so. Would Fork be back? Here the answer seems less clear. Intuitively, if Sylvia glues the pieces in place, or leaves them there for several weeks, or develops some particular attachment to the sequence F-O-R-K, it does seem right to say that Fork is back—that is, that those particular alphabet blocks in that particular arrangement give us an object identical with the object originally dubbed *Fork*. And to the extent that one has doubts about saying whether Fork is back, it seems to be a function of there being doubts about whether Fork is an object at all. The worry seems to take the form: look, if those blocks spend most of their time jumbled up together in their storage crate, what's the point of calling some of them configured in some particular way an *object*? But so long as it seems to make sense of speaking of Fork as something that *might* exist at more than one time, it seems to make sense to think that Fork is back when Sylvia reconstructs him, but not when Elmer does. (Even if you do not share *this* intuition, the argument will work as long as you think it is a *cleaner* case to speak of Fork's being back when reconstructed by Sylvia than when reconstructed by Elmer. What I need for my argument to work is the intuition that *process* can affect rational evaluation of identity.)

So same matter-same form is not always enough to give us identity over time, and that this might be for one of two reasons. The first is that we are dealing with an object not stable enough to count as a candidate for identity over time. (Note that one logical extension of this idea is that there are, strictly speaking, no objects over time. See section **3.5.1** above.) The second is that the *process* by which the object came into being may affect whether we consider the later object to be identical with the earlier one. (Again, note that one logical extension of this idea is a four-dimensional view where what makes up Fork are the Monday part of F, O, R and K, and then the Sunday parts of them, etc. See footnote 89.) It is the second of these reasons which will matter for my argument.

4. Personal Identity

4.1 Introduction: The Facts of Life

Human beings come into existence through a well-known sequence of natural processes. Following some sort of manifestation of the facts of life, an egg fertilized by a sperm is implanted in the wall of a uterus, where it develops from a collection of a few cells into a progressively more complex entity. After nine months, the mother gives birth to the new human being, who then undergoes a long process of feeding and nurturing and interacting with other human beings who have taken on the responsibility for the development of this one. Eventually the child reaches adulthood, and as an autonomous agent, makes plans and commitments and decisions, and has experiences and relationships and ideas.

What I will argue in this chapter is that these well-known facts have not received the right kind of attention in recent philosophical discussions concerning the nature and value of personal identity. It is not that they are *forgotten*; even in the throes of debate, no one thinks that, as a matter of fact, human beings sometimes come into existence through teletransportation or fission or brain-state-transfer. It is that they are deliberately treated as *provincial* truths: "this may be the way we get persons off in our little corner of possibility space, but simply considering specimens from our village will not tell us everything we would want to know about what sort of things persons *are*. Many of the features that all of *us* share are just local color. And we wouldn't want to be such yokels as to mistake these coincidental features of persons-as-they-happen-to-be for the constitutive features of persons-as-they-truly-are[147]."

My goal in this chapter is to highlight a danger that plagues such attempts to overcome provincialism[148]: When we try to abstract away from contingent features of the way things happen to be, we tend to discount elements that actually play a role in explaining the appropriateness of our responses. So in our overzealous efforts not to confuse features that are accidental for features that are essential, we often end up treating something that is crucial as if it were merely coincidental.

The danger I will highlight is a general one, but I will spend most of my time trying to establish a fairly specific claim. What I will try to demonstrate is that a

[147] Cf. Shoemaker: "What Mackie and Perry have done is to indicate how personal identity (or copersonality, or psychological unity) are realized in *us*, i.e. in members of our own species. And this does not answer the question 'What does personal identity consist in?' at the level of abstractness at which we want it answered" (Shoemaker and Swinburne 1984, 127). Or again Unger: "[W]hy not stick only to actual cases?...The reason is that this extremely conservative methodology is apt to incur great costs...In attempting to ascribe beliefs to ourselves on the basis of quite limited data, we might wrongly describe our own attitudes" (Unger 1990, 11). Or Nozick: "We...are not so tied to our bodies that we find it impossible to imagine coming to inhabit another. We do not conceive of ourselves as (merely) our particular bodies, as inextricably tied to them" (Nozick 1981, 30).

[148] These attempts involve appeal to imaginary cases which, like experiments, are supposed to help us compensate for the often arbitrary way in which we tend to come upon information in the world. Experiments isolate sets of phenomena so that the relations among them are made manifest, and thereby reveal in an epistemically accessible way patterns that, in some sense, were already out there to be seen. In a parallel fashion, *thought experiments* are supposed to help us distinguish relevant from irrelevant features in actual cases by making manifest their relations in cleaner, non-actual, cases.

certain widely-accepted[149] argument of Derek Parfit's concerning the importance of personal identity is unsuccessful because it ignores the sort of facts I described in the first paragraph. Parfit tries to show that what ought rationally to matter to us when we care about survival and future well-being is not that *we ourselves* survive, but only that someone exist who is psychologically continuous with us in the right sort of ways. He describes an imaginary case where it seems clear that identity would not be what matters in this way, and argues that this result can be generalized to show that identity is *never* what matters[150].

What I will try to show is that a strategy that on its surface seems totally untenable can in fact be used to block Parfit's conclusion. I will argue that what explains the rationality of prudential concern is always identity, while accepting that in some cases rational prudential concern can hold in the absence of identity. The trick is to show that the feature that explains or justifies or makes rational a relation can be a different feature from the one that underpins it as a necessary condition. That is, the trick is to show that—in certain contexts—absent features can do explanatory work.

[149] Cf. Noonan 1989: "Parfit's argument...establishes that one cannot *both* regard personal identity as something which is of non-derivative importance *and* regard it as determinable by extrinsic facts" (Noonan, 196-197). Noonan goes on to reject the latter; he holds that personal identity is intrinsically determined. What I argue is that Parfit's argument fails to establish that there is such a dichotomy. I show that one can simultaneously hold personal identity to be something of non-derivative importance *and* regard it as extrinsically determinable.

[150] Parfit writes: "By considering these cases, we discover what we believe to be involved in our continued existence...Though our beliefs are revealed most clearly

Continues on next page…

After a brief discussion of the context into which Parfit’s case fits, I will present his argument and spell out what I take to be its crucial methodological assumptions. I will then try to show that the argument seems so compelling because of its tacit reliance on two ostensibly undeniable principles, the first of which concerns the ranking of preferences, the second of which concerns the assignment of explanatory force. I will demonstrate that neither of the principles applies to the fission case in the way Parfit’s argument requires, and show finally that this inapplicability can be traced to the ways in which the example ignores what I have been calling “the facts of life[151/152].”

4.2 Setting the Stage

4.2.1 A Context for Parfit’s Argument

Over the last forty years, a sizable literature has developed describing imaginary cases meant to illuminate questions concerning the metaphysics of personal identity and the rational grounding of our relations between our present and

when we consider imaginary cases, these beliefs also cover actual cases, and our own lives” (Parfit 1987, 200).

[151] My ideas on personal identity have been strongly influenced by the writings of Mark Johnston, especially his discussion of Parfit in “Reasons and Reductionism” (Johnston 1992b) and the forthcoming “Human Concerns without Superlative Selves” (Johnston 1992c). For those who know these works, my debts to Johnston should be obvious.

[152] For discussions of this methodological question, see, among others: Baillie 1993, esp. 200-205; Ethics 96 (July 1986), *passim*; Gale 1991, 299-303; Haksar 1991, 149-155; Hertzberg 1991, 153-155; Johnston 1987a, esp. 60-69, 80-83; Oderberg 1993, 32-36; Quine 1972, 490; Rovane 1994, *passim*; Unger 1990, chapter 1, esp. 7-15, 27-35; Wiggins 1980, chapter 6, esp. 169-175, 221; Wilkes 1988, chapter 1.

our future selves[153]. Such cases are of two kinds. The first kind exploit the assumption that two structurally identical individuals composed of qualitatively identical matter will manifest the same qualitative characteristics, both physically and in terms of non-physical states. Such cases describe situations where some individual (who may share none of the same matter as the ordinarily-produced person she comes to resemble) comes into existence by some reliable or unreliable process that produces an individual in all ways indistinguishable from an ordinary human being. Often these cases involve descriptions of procedures, such as teletransportation or brain zapping, which involve the microstructural reconfiguration of found matter in order to produce macrostructural results. Since in all actual instances, substantive and qualitative criteria of individuation and valuation coincide, these cases are meant to tease apart which of the features in question is actually doing the grounding work.

Other cases, following the assumption that whatever produces consciousness can be localized to some fairly small part of the human organism, in particular the brain, involve thinking about what would happen if the brain associated with some human being were transplanted into some other body, either intact, or, with the

[153] To the extent that they involve the contemplation of scenarios in which a human being is brought into existence who may be identical to no previously existing human being, these cases violate the predictable natural sequence of events that I reminded us of in the first paragraph. But unlike Greek myths and other people's religions, these cases are not invoked as poetic expressions of our deepest desires: that we should overcome our bodily encagement, that we should be able to create objects at will, that simply having enough knowledge about a person—knowing what she is like down to the last micromolecule—should be sufficient to *bring her back*. These are cases we consider, not as the production of some odd tribe of philosophers, but as examples about which the convergence of opinions is going to reveal our deepest commitments and beliefs about our nature as human beings. (See note 108.)

Continues on next page…

additional assumption that only part of the brain is necessary to support the array of psychological characteristics in question, in some attenuated form. The first of these is generally referred to as brain-transplant, the second as fission. These cases are meant to tease apart the extent to which the commitment to identity and value is bodily based, and the extent to which we require only psychological and qualitative similarity, along with (perhaps) some of the same matter and (perhaps) the right kind of cause.

4.2.2 Parfit's Argument

The case on which I will focus in this chapter falls into the second class, and is standardly referred to as *fission.* In Parfit's version[154], three triplets are involved in an accident in which the body of one—call him Brainy—is fatally injured, while the brains of his two brothers are totally destroyed. Brainy is such that the physical bases for his psychological characteristics happen to be realized in duplicate, one complete

[154] In Reasons and Persons, Parfit presents the case as follows:

> *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 (Parfit 1987, 254-5)

Parfit presents the vast majority of his examples in a first-person form (see Parfit 1987, 199ff). Although this convention certainly provides convenience of locution, I think it also reflects an attempt to bring the reader to imagine the cases "from the inside." My decision to use the third-person in discussing this case is on stylistic grounds alone. For a discussion of the issues involved in first-person as opposed to third-person thought experiments, see Shoemaker 1994. See also Madell 1991, esp. pp. 128-129.

set in each lobe. Following the accident, doctors divide his brain in half, and transplant the two hemispheres into the bodies of the two brothers.

In the first scenario, which Parfit calls the "one-sided case," only the left transplant takes, and the right hand transplant is destroyed. The resulting individual, whom we will call Lefty[155], has all of Brainy's memories and psychological states and a body almost indistinguishable from the one that Brainy had before the accident. Parfit holds, and for the sake of argument we will grant him that[156]:

[155] Note that "Lefty" is not a name; it is an abbreviation for the description "the individual who has Brainy's original left hemisphere." For a discussion of the sorts of confusion that arise from failure to recognize this, see Shoemaker 1984, 116-118.

[156] Parfit takes it as common ground both among supporters of the physical criterion of personal identity and among supporters of the psychological criterion of personal identity that person P_1 at t_1 is identical to person P_2 at t_2 if:

(i) P_2 has enough of P_1's brain to preserve psychological continuity*,
(ii) P_2 has all of P_1's psychological characteristics, and
(iii) the brain in question has been continuously and properly functional from t_1 to t_2
(iv) there are no other (equally good) candidates for identity with P_1

This position is rejected by Williams (see Williams 1973, *passim*), and more recently, by Thomson 1994, who endorse what might be called a "bodily criterion." But as Parfit has pointed out (personal communication), his fission argument might be made without appeal to a brain transplant. In this version, the one-sided case would involve Brainy's survival under conditions in which half his brain and body are destroyed, but the remainder (sufficient to support full psychological continuity) is kept functional by means of appropriate technology. The two-sided case would involve a corresponding division of Brainy's brain and body into two such medically viable psychological continuers.

I am not convinced that this reformulation would be sufficient to satisfy Williams and Thomson. The reason half-brain survival is sufficient to guarantee identity for holders of a psychological or brain criterion of identity is because (*ex hypothesi*) it is sufficient to support full psychological continuity. But it is not clear what the analogous critical mass would be in the case of bodily survival; that is, as sufficient brain-survival is to psychological continuity, so sufficient body-survival is to what? (*Medical viability* does not seem to be a clear enough criterion, since there is

Continues on next page…

(1) In the one-sided case, Lefty *is* Brainy.

Given this, it seems clear that were Brainy to know (before the accident) that there were some discomfort he could undergo in order to save Lefty greater discomfort in the future, he would be prudentially rational to undergo it. That is, in Parfit's terms[157]:

(2) In the one-sided case, Brainy's relation to Lefty contains what prudentially matters.

Parfit contrasts this case with a second scenario, which he calls the "two-sided case." In the two-sided case, *both* transplants are successful. *Each* of the two resulting individuals—we'll call them Lefty and Righty—has all of Brainy's memories and

no "whole" which the partial-body can nonetheless sustain.) In any case, this reformulation does not help Parfit in countering closest continuer or multiple-occupancy type objections. (Indeed, it may even weaken his case against the latter, in that it seems to lend support to the view that in any one person, there are two subpersons waiting to get out.)

One further remark: note that if one were to "ask" the "body" (rather than the "mind") about the brain-transplant case, it would surely prefer bodily (i.e. genetic) to mental survival, and that even if the person who results from the brain transplant is identical with the person whose brain was transplanted, the children to whom he or she will give birth will inherit genetic information from the recipient body. Darwin wins, not Lamarck: if you put A's brain in B's body, the resulting person will give birth to B's children.

> *As evidence for the widespread acceptance of this view, consider the following "Incredible Brain Fact" from the back cover of the box enclosing "The Incredible Growing Brain" (a children's toy made in Taiwan): "We would not be ourselves if our brain was transplanted [*sic*]." (I thank Teresa Robertson for the citation.)

[157] Parfit uses the expressions "what matters" "what prudentially matters" and "what matters for prudential concern" interchangeably, preceded variously by the verbs "contains" "has" and "is."

psychological states and a body almost indistinguishable from the one that Brainy had before the accident. Parfit points out that:

> (3) Brainy's relation to Lefty is intrinsically the same in the one- and two-sided cases.

And, he insists, were Brainy to know (before the accident) that there were some discomfort he could undergo in order to save Lefty greater discomfort in the future, he would be prudentially rational to undergo it[158]. That is, Parfit maintains that:

> (4) In the two-sided case, Brainy's relation to Lefty contains what prudentially matters.

But Lefty and Righty are not the same person. After all, the two occupy distinct spatial locations, undergo different experiences, and have no particular causal effect on one another. And if Lefty and Righty are different people, and Brainy is a single person[159], then Lefty and Righty cannot both be identical to Brainy. Thus[160], Parfit holds:

[158] This particular test comes from Unger; he calls it "The Avoidance of Future Great Pain Test" (see Unger 1990, 27-34).

[159] For a denial of this assumption, see Lewis 1976 and Lewis 1983b. See also Noonan 1989, 164-168, 197-198 and Mills 1993.

[160] Strictly speaking, of course, additional assumptions are required for this step. The simplest argument would make a straightforward appeal to some principle of sufficient reason as follows: Since by stipulation there is no relevant difference between Brainy's left and right lobes, there is no reason that Brainy would be identical with Lefty as opposed to Righty, or with Righty as opposed to Lefty; and since he cannot be identical with both, then he must be identical with neither. More elaborately, one might proceed by describing *two* versions of the one-sided case, the first with the right lobe, the second with the left; from this and the fact that Brainy is not identical to both Lefty and Righty, it follows that Brainy in the two-sided case bears an intrinsic relation like that borne in cases of identity towards at least one person with whom he is not identical.

(5) In the two-sided case, Lefty *is not* Brainy.

Since in the two-sided case, Brainy and Lefty are not identical, and since their relation contains what prudentially matters, Parfit concludes that:

(6) In the two-sided case, identity is not what prudentially matters.

And if we have what prudentially matters in the two-sided case, and whatever that is is not identity, then how can it be identity that prudentially matters in the one-sided case? So:

(7) In the one-sided case, identity is not what prudentially matters.

But the only relevant difference between Brainy's survival as Lefty in the one-sided case and Brainy's survival in ordinary cases is the mutilation that Brainy has undergone as a result of the accident and subsequent medical treatment. And we have already conceded that in the one-sided case, Lefty *is* Brainy. The mere fact that he has been through a terrible accident should not affect the reason *why* Brainy cares for himself tomorrow. So, Parfit concludes:

(8) Even in ordinary cases, identity is not what prudentially matters.

4.2.3 The Crucial Premise

The key moves in Parfit's argument are the moves from (3) to (4), and from (4) to (6), (7) and (8). At each step, Parfit invokes what I will call the crucial premise:

> *The crucial premise*: Whether a relation contains what prudentially matters depends only on the relation's intrinsic features.

If the crucial premise is correct, then (4) follows from (2) and (3), since the relation contains what matters in the one-sided case, the intrinsic features of the relation are the same in the two-sided case, and, according to the crucial premise, these features are the only ones that are relevant. And if (4) is true, then given the crucial premise, (6)-(8) follow immediately. Since the relation in the two-sided case contains what prudentially matters, and the relation in the two-sided case is not identity, what prudentially matters in the two-sided case cannot be identity. And since whether a relation contains what prudentially matters depends only on its intrinsic features, the same is true in the one-sided case, and in ordinary cases.

4.2.4: Crucial Terms in the Crucial Premise: Intrinsic and Extrinsic Relations

Given that the crucial premise does so much work in the argument, and given that it speaks of intrinsic features, our first task is to get clear about what such features are. Intuitively speaking, the *intrinsic* features of a relation are facts about the relata themselves (and the relations between them), whereas the *extrinsic* features of a relation are facts about objects distinct from the relata (and the relations between them). On the basis of this distinction, we might say that a relation between X and Y is *intrinsic* if whether it holds depends only on intrinsic features[161], and *extrinsic* if

[161] Cf. the following formulations of the principle for cases of identity, which Wiggins and Noonan endorse, and Nozick and Parfit deny. Wiggins's *only-a-and-b principle*: "for a relation R to be constitutive of the identity of *a* and *b*, *a*'s having R to *b* must be such that objects distinct from *a* and *b* are irrelevant to whether *a* has R to *b*" (Wiggins 1980, 96). Noonan's *Only x and y principle*, which he defines (for personal identity) as: "whether a certain later person P2 is identical with a certain earlier person P1 can depend only on facts about P1 and P2 and the intrinsic relationships between them; no facts about individuals other than P1 and P2 can be relevant to whether P1 is the same person as P2" (Noonan 1989, 16). Nozick: "If x at time t_1 is the same individual as y at a later time t_2 , this can depend only upon facts about x, y , and the relationships between them. no fact about any other existing thing is relevant to (deciding) whether x at t_1 is (part of the same continuing individual as)

Continues on next page…

whether it holds depends on other things as well. *Being taller than* is an intrinsic relation: whether Goliath is taller than David depends only on the height of Goliath, the height of David, and their comparative magnitudes; nothing about any other object could affect the relation in question. Other relations that are intrinsic in this sense include “is the same shape as,” and “is exactly five years older than.”

By contrast, one way in which a relation might be extrinsic is if ascertaining whether the relation holds between X and Y requires ascertaining whether such a relation also holds between Y and someone else. Whether Isaac is the only son of Abraham depends not only on the relation between Abraham and Isaac, but also on the relation between Abraham and Ishmael. Isaac is Abraham’s only son only if he is Abraham’s son *and* if, in addition, no other male bears to Abraham the same intrinsic relation that Isaac does—that is, only if no one else is Abraham’s son. Other extrinsic relations of this sort are “being the oldest sibling of” and “being the richest passenger on.”

This provides a way of articulating an asymmetry between identity on the one hand, and what matters for prudential concern on the other. The former is extrinsic in the way just described; whether Brainy is identical with Lefty depends not only on facts about Brainy and Lefty, but also on facts about Brainy and *Righty*. But if the crucial premise is true, what matters for prudential concern is *not* extrinsically

y at t_2 “ (Nozick 1981, 31). Parfit: “Intrinsicness of Personal Identity: If some future person will be me, that fact must depend only on the intrinsic features of the relation between me now and that future person. It cannot depend on whether the same relation holds between me now and some other future person.” (Parfit 1994, 31).

determined; whether Brainy's relation to Lefty contains what matters for prudential concern depends *only* on facts about Brainy and Lefty.

But isn't this tantamount to saying that the crucial premise does *all* the work? If what prudentially matters must be an intrinsically determined relation, and identity is not an intrinsically determined relation, then identity is not what prudentially matters; end of argument. And if this one-sentence *modus tollens* is really the argument, then what is going on in (1)-(8) above?

Note first that the argument presented in (1)-(8) is essentially the following:

(a) [(1), (2)] In the one-sided case, Lefty is Brainy and their relation contains what prudentially matters.

(b) [(3)] In the two-sided case, Lefty's relation to Brainy is intrinsically the same as it is in the one-sided case.

(c) [(4), (5)] In the two-sided case, Lefty is not Brainy and their relation contains what prudentially matters

(d) [(6), (7), (8)] The relation which contains what prudentially matters need not be identity.

The first question to be asked is: what justifies (c)? That is, what makes it the case that the relation between Brainy and Lefty in the two-sided case contains what prudentially matters? There are only two possibilities: either (c) follows from (a) on the basis of (b), or (c) can be justified without appeal to (b)[162]. ((b) itself is just an

[162] One way of characterizing the difference between my view and Parfit's is to point out that on Parfit's view, (4) must either be justified on the basis of (2), (3) and the crucial premise, or it must be a bare truth. On my view, (4) may be taken as a derivative truth—something that holds as a consequence of a certain background pattern of other truths also holding.

application of the definition of ‘intrinsically the same’ to this particular case.) Suppose the former. That is, suppose that what justifies (c) is the assumption that the relation between Brainy and Lefty in the one-sided case contains what prudentially matters, along with the fact that the relation between Brainy and Lefty is the same in the two-sided and one-sided cases. If so, then Parfit is relying on the assumption that whether a relation contains what prudentially matters depends only on intrinsic features of the relation; clearly this gets us from (a) and (b) to (c), and it is unclear what *else* could get us from (a) and (b) to (c). But if this is correct, then it is undeniable that the argument makes appeal to the crucial premise. And if so, then so much the worse for the rest of the argument if from this, it follows that if identity is extrinsically determined, then identity cannot be what prudentially matters. Surely this is not the fault of the interpretation.

Suppose instead that (c) is independently justified; that is, suppose that we have grounds for thinking that in the two-sided case, the relation between Lefty and Brainy contains what prudentially matters, but that this is established without appeal to the crucial premise. If so, the argumentative work must be being done in the move from (c) to (d). On this reading, (c) is simply an existence claim; it presents us with an instance of a case where what prudentially matters is present and identity is not, and this is true on grounds independent of the crucial premise. So we need to ask: what follows?

On one reading of (d), all that follows is an existential quantification: there exists a case where we have what prudentially matters even though we do not have identity[163]. Clearly, such a conclusion follows directly from (c), without added assumptions; but equally clearly, this is not all that the argument seeks to show. On a second reading of (d), the conclusion is an unrestricted generalization: what prudentially matters is, in general, not identity[164]. But this stronger conclusion follows from (c) only if we assume that we are searching for necessary conditions for prudential concern, such that these conditions will be the same in all cases[165]. But what could justify such an assumption?

In the rhetorically effective version of the argument presented in (1)-(8), the move from (6) to (8) is justified on the grounds that the relation between Brainy and Lefty is intrinsically the same in the one-sided and in the two-sided case ((6) to (7)), and that the relation is intrinsically the same in the one-sided case and in normal cases ((7) to (8)), hence the relation is intrinsically the same in the two-sided case and in normal cases ((6) to (8)). And my contention is: this is what justifies the move from (c) to (d) in the reconstruction I have just presented. The alternative is to say that the move is justified by the bare assumption that whatever makes prudential concern justified in the two-sided case is what makes prudential concern justified in all cases where it is justified.

[163] This is (6) above.

[164] This is (8) above.

[165] In the sense of being a necessary condition for (rational) prudential concern, see section **4.2.5**.

So it turns out that even if we grant that Parfit can get from (a) to (c) without appeal to the crucial premise, we must acknowledge that he uses the premise to get from (c) to (d). But if so, then once again the objection that the reconstruction errs by placing the entire burden on the crucial premise fails. For once again, it turns out that the original argument does the same thing[166].

4.2.5 More Crucial Terms in the Crucial Premise: "What Matters"

The second crucial element of the crucial premise is *containing what prudentially matters*. What I will try to show is that there are actually two possible readings of "what matters" which have been standardly conflated. On the first reading, Parfit's argument is sound, but it shows much less than he intends; on the second reading, the argument—if successful—would indeed show what Parfit intends, but on this reading, I will suggest, the crucial premise is false and the argument unsound.

What, then, are the two readings? On first reading, the argument concerns the *necessary conditions* for rational prudential concern; that is, what features *must be present* in any situation where rational prudential concern obtains. We might call these the N-features. On this reading, the argument's conclusion is that identity is not among the N-features: that is, identity is not among the features that are necessarily

[166] Parfit (personal communication) suggests another option: that the crucial premise is neither redundant, nor so compelling that all the rest follows immediately from it. Rather, he suggests, what his argument establishes is that a certain collection of premises hang together and form a coherent world picture.

present in all cases where rational prudential concern obtains[167]. On the second reading, what is at issue is what *explains* rational prudential concern; that is, what features make it the case that rational prudential concern is *justified*. We might call these the E-features[168]. On this reading, the argument's conclusion is that identity is not among the E-features: that is, identity is not among the features that explain the obtaining of rational prudential concern[169].

How do we go about ascertaining whether Brainy's relation to Lefty contains what prudentially matters in one of these senses? Answer: we enumerate the features of the sub-relations that hold between them, and we check to see whether among these features are either the N-features or the E-features. What I will argue is that in the two-sided case, Brainy's relation to Lefty contains—in the relevant sense—the N-features but not the E-features. That is, I will argue that if Parfit's argument is taken

[167] One might wonder why I am not suggesting a sufficient condition reading; after all, it seems that the argument seeks to establish that whatever relation holds between Lefty and Brainy is *sufficient* for rational prudential concern, not that it is *necessary*. But this is a confusion. The argument from (1) to (8) is essentially a *negative* argument; it seeks to establish that identity is *not necessary* for rational prudential concern. It so happens that in so doing, the argument *also* establishes that something less than identity is sufficient for rational prudential concern. But this corollary does not need to be built into my characterization. Clearly, if x is not necessary for Y, and there is at least a partial ordering of characteristics, then it follows that something less than x is sufficient for Y.

[168] Because what these features explain is what makes such concern *justified*, some readers have suggested that I refer to them as the J-features. Although I see there point, I think to follow this suggestion would be misleading. Even on the necessary condition reading, we are concerned with the presence or absence of a relation which makes prudential concern *rational.* So there is a sense in which justification is at issue even when we discuss the N-features.

[169] I thank Scott MacDonald for pressing me to clarify these claims.

in the *necessary condition* sense, then (4) is true, but if it is taken in the *explanatory* sense, then (4) is false.

We might think of these two readings as expressing two versions of Parfit's argument. On the first reading, the "necessary condition" reading, we attach the words "in the necessary condition sense" to each of the sentences where "what prudentially matters" appears; and on the second reading, the "explanatory" reading, we attach the words "in the explanatory sense" to each of these sentences. The resulting arguments would read as follows:

"Necessary condition" version:

> (1') In the one-sided case, Lefty *is* Brainy.
>
> (2') In the one-sided case, Brainy's relation to Lefty contains what prudentially matters *in the necessary condition sense* (i.e. the N-features are present).
>
> (3') Brainy's relation to Lefty is intrinsically the same in the one- and two-sided cases.
>
> (4') In the two-sided case, Brainy's relation to Lefty contains what prudentially matters *in the necessary condition sense* (i.e. the N-features are present).
>
> (5') In the two-sided case, Lefty *is not* Brainy.
>
> (6') In the two-sided case, identity is not what prudentially matters *in the necessary condition sense* (i.e. identity is not among the N-features).
>
> (7') In the one-sided case, identity is not what prudentially matters *in the necessary condition sense* (i.e. identity is not among the N-features).
>
> (8') In ordinary cases, identity is not what prudentially matters *in the necessary condition sense* (i.e. identity is not among the N-features).
>
> *Crucial Premise (necessary condition version*): Whether a relation contains what prudentially matters *in the necessary condition sense* depends only on the relation's intrinsic features (i.e. the N-features are intrinsic features).

"Explanatory" version:

> (1") In the one-sided case, Lefty *is* Brainy
>
> (2") In the one-sided case, Brainy's relation to Lefty contains what prudentially matters *in the explanatory sense* (i.e. the E-features are present).
>
> (3") Brainy's relation to Lefty is intrinsically the same in the one- and two-sided cases.
>
> (4") In the two-sided case, Brainy's relation to Lefty contains what prudentially matters *in the explanatory sense* (i.e. the E-features are present).
>
> (5") In the two-sided case, Lefty *is not* Brainy.
>
> (6") In the two-sided case, identity is not what prudentially matters *in the explanatory sense* (i.e. identity is not among the E-features).
>
> (7") In the one-sided case, identity is not what prudentially matters *in the explanatory sense* (i.e. identity is not among the E-features).
>
> (8") In ordinary cases, identity is not what prudentially matters *in the explanatory sense* (i.e. identity is not among the E-features).
>
> *Crucial Premise (explanatory version*): Whether a relation contains what prudentially matters *in the explanatory sense* depends only on the relation's intrinsic features (i.e. the E-features are intrinsic features).

So, for instance, what (2') means is that whatever Brainy's relation to Lefty is in the one-sided case, that relation contains, in the relevant way, the N-features; what (4") means is that whatever Brainy's relation to Lefty is in the two-sided case, that relation contains, in the relevant way, the E-features. (6')-(8') claim that identity is not among the N-features; and (6")-(8") claim that identity is not among the E-features. And on the first reading the crucial premise says that to ascertain whether the relation between Lefty and Brainy contains the N-features, we need only look at the intrinsic sub-relations that hold between them; on the second reading, it makes a corresponding claim about the E-features.

4.2.6 Three Questions

On the readings I propose, the move from (6) to (7) and (8) is trivial; whatever the N-features and E-features are, they are the same whether we are talking about the two-sided case, the one-sided case, or ordinary cases. So if these readings are correct, three questions arise:

- Why does Parfit seem to think that one needs an argument to get from (6) to (8)?
- How, if I am denying the crucial premise in the double-prime case, do I get from (6") to (8")?
- Doesn't this show that it was not the crucial premise to which Parfit implicitly appeals in his move from (6) to (8)?

I answer each in turn.

I think the appearance of needing an argument to get from (6) to (8) (in either the single-prime (necessary condition) or double-prime (explanatory) sense) is a result of the following line of reasoning: Although (4) implies (6) even without the crucial premise, (4) alone does not imply (7) or (8). However, if we add the crucial premise, then (7) and (8) follow. So there is a way to get from (4) to (6) that is not available as a way to get from (4) to (7) or (8). It is for this reason, I suggest, that it seems to Parfit that he needs to make an argument to get to (8) from (6)[170].

[170] Cf. Parfit: "I went on to claim [A] 'since identity would not be what matters [in the two-sided case], it is never what matters.' Johnston rejects [A]. Though he concedes that, in my example, identity may not be what matters, he believes that we

Continues on next page…

But how do I get from (6") to (8"), if I am denying the crucial premise in the double-prime case? The answer is that I take the expression “what prudentially matters in the explanatory sense” to refer to the same thing, regardless of which case we are talking about. That is, I take it that “what prudentially matters in the explanatory sense” (that is, the E-features) is the same in the two-sided case and in the ordinary case, *not* because the E-features are something that depend only on the intrinsic features of the relation between the individuals in question, but rather because the E-features explain prudential concern *in general*, and not only in some particular subset of cases. Likewise, I can justify the move from (6') to (8') not on the grounds that the N-features are intrinsically determined (though I think they are) but on the grounds that whatever the N-features are, they are the features that necessarily obtain in *all* cases where rational prudential concern is warranted, and not something that differ case-by-case.

Why, then, do I think it is the crucial premise that does the work when Parfit argues from (6) to (8)? The reason is simple: clearly, he does not think the move from (6) to (8) is trivial (or he would not argue for it); and, as I pointed out in answering the first question, one difference between (6) and (8) is that the move from (4) to (6) can be made without the crucial premise, but the move from (4) to (8) cannot. But

can *quarantine* this solution. We can claim that, in ordinary cases, identity is still what matters. [A], I agree, needs more defense” (Parfit 1994, 36). As should be clear to the reader by now, I think the apparent need for a defense of A results from a failure to distinguish between the necessary-condition reading and the explanatory reading. If we hold “what matters” constant, A requires no defense. So it is only by shifting between meanings of the term (or by assuming that what matters in one case

Continues on next page…

since Parfit already requires the crucial premise to get from (2) to (4), he is free to appeal to it again in justifying the move from (4) to (6), and thence to (7) and (8).

4.2.7 A Further Question

But a further question arises. Why, on my view, do we respond to the case in the way Parfit expects? That is, why does it occur to us to say that identity is not what matters for prudential concern in the necessary condition sense if not because we accept the stronger claim that identity is not what matters for prudential concern in the explanatory sense[171]? The reason, I think, is that there is a claim which is easily confused with (4"), which I will call the intermediate claim. The intermediate claim says:

> *Intermediate claim*: In the two-sided case, Brainy’s prudential concern for Lefty would be rational.

I think the reason one might be tempted to endorse (4")—that in the two-sided case, Brainy’s relation to Lefty *contains what matters* for prudential concern in the explanatory sense—is because one has confused it with the intermediate claim—that Brainy’s relation of prudential concern for Lefty would be rational in such a case. But there is a substantial difference between the two claims. Whereas (4") says that the E-features are present in the two-sided case, the intermediate claim says only that Brainy’s prudential concern for Lefty would be rational in such a case, *even if the*

could be different than what matters in another), that the need for such a defense arises.

[171] Another way to ask this is: why do we accept (4') if not because we accept (4")?

Continues on next page…

relation Brainy bears to Lefty does not "contain what matters" in the explanatory sense[172]. So unlike (4"), the intermediate claim is not strong enough to get us (6"); from the intermediate claim, it does not follow that identity is not what matters in the *explanatory* sense, even in the two-sided case.

What I will try to show in the remainder of the chapter is how the apparent success of Parfit's argument results from a failure to make the distinctions I have just drawn. Because the necessary-condition and explanatory readings are conflated, showing that (8') is true is taken as sufficient for showing that (8") is. And because the intermediate claim is taken to be equivalent to (4"), accepting it seems tantamount to endorsing the crucial premise in the explanatory sense, and with it (6"), (7") and (8"). So it seems that accepting that Brainy would be rational to bear prudential concern for Lefty in the two-sided case gets us all the way to the conclusion that what makes prudential concern rational in ordinary cases is not identity. But, as I have been suggesting, this reasoning rests on several serious mistakes.

4.2.8 What Does Parfit Need to Show?

I have conceded that the crucial premise is true on the necessary condition reading. Whatever it is that *has to* be present in all cases where prudential concern is rational is indeed something which depends only on intrinsic features. But Parfit needs much more than this to establish what he hopes to establish. If we respond to the case in the way he expects, he thinks we *should* change our views about what underpins our prudential concern for our future selves. But all the necessary condition reading shows is that identity need not be present in every case where prudential concern is rational. This is something weaker than what Parfit means when he says "identity is not what matters." Parfit needs to show that identity is not what matters *in the explanatory sense*; he needs to show that identity is not what *justifies* prudential concern. It is precisely this which I deny. Even though identity is not a necessary condition for rational prudential concern, it is still what explains it.

4.3 What Makes Personal Identity an Extrinsic Relation?

Assuming, as most moderns do, that psychological characteristics supervene (at least weakly) on physical microstructure, it follows that persons are, at base, complex physical objects. It is a general fact about complex physical objects that their identity conditions seem to allow for multiple-candidacy; this is a direct consequence of the relation that underpins their identity allowing for preservation of identity under conditions of growth, change, or mutilation. As a result of this unavoidable leniency, there are at least two sorts of cases in which a process that is ordinarily identity-preserving may instead be entity-creating[173]. The first sort—Ship of Theseus cases—

[172] That is, the intermediate claim allows that the rationality of Brainy's prudential concern for Lefty might depend on another relation for its "borrowed luster."

[173] There are two other sorts of cases which might be candidates for this description, or at least for the description: "a process that is often identity-preserving which may instead be entity-creating." One would be cases involving creation of an exact qualitative duplicate; the other would be cases involving instantaneous loss of a bulk of matter. Both sorts of processes are sometimes considered person-preserving in circumstances where there is no equally good or better competitor: the first in teletransportation, the second in survival-as-brain.

Continues on next page…

play off the fact that two distinct sorts of procedures are, in ordinary cases, sufficient for the continued existence of an entity; the second sort—fission cases—play off the fact that an entity might continue to exist despite the sudden loss of half or more of its matter. In this chapter, I will consider only cases of the second kind.

The series of assumptions that underlie the need to consider fission cases can be put in terms of these statements, which might be characterized as "we would consider the following sort of being to be relevantly like us" statements.

> We would consider the following sort of being to be relevantly like us:
>
> (9) an organism whose psychological activity—knowledge, belief, memories, intentions, tacit motivations, etc.—supervenes on the physical structure of its brain. [Note: this need not be its brain; it could be anything (compact and physical), in which case replace "brain" with that throughout]
>
> (10) an organism meeting condition (9) from which it is technically possible to remove the brain surgically in such a way that, suitably resituated, the brain would support the same psychological activity before and after the surgery.

I think cases of this second sort can be assimilated either to cases like Theseus cases, which I have discussed in chapter 3, or to cases like fission cases, where the same criterion is equally well-met twice over by some bit of organized substance. So I don't think these cases raise genuinely new issues.

Cases of the first sort (teletransportation cases) *do* present a genuinely new kind of case, but here, I think, we have overstepped the bounds of identity. That qualitative features alone (or qualitative features constrained by some suitably replicable causal process) could serve as the criterion for *numerical* identity of entities of a certain kind would make sense only in a context where there is some sort of nomological near-guarantee of uniqueness (which is exactly what the supposition of teletransportation denies), and even here, it seems to me highly suspect (for reasons discussed by, among others, Wiggins 1980, 208; Williams 1956; and Williams 1960).

> (11) an organism meeting condition (10) such that when the brain is surgically implanted in a body other than the original, its relation to the new body with regard to sensori-motor control is qualitatively like its relation to its original body.
>
> (12) an organism meeting condition (11) whose brain is such that all features structurally relevant to its psychological functioning are realized in duplicate.
>
> (13) an organism meeting condition (12) where the duplication would be such as to allow the brain to be surgically divided into two distinct sub-brains, each meeting the removal and transplant conditions as described in (10) and (11).)

Defenders of fission contend that the depictions in (9)-(11) describe ways that persons might actually (turn out to) be, so they pose no conceptual problem for the example. We *would* (and *should*), they think, consider those sorts of beings to be relevantly like us. As far as defenders of fission are concerned, the only potential problems lie with (13), where the depiction might be problematic for two distinct reasons[174]. The first, "deep," reason would be that a single consciousness could never be divided in such a way as to produce two distinct streams; the second, "shallow," reason would be that the lower brain of human beings, as it is actually constituted, could not be divided without rendering it non-functional. Both of these problems are treated as undamaging by defenders of fission. Because of empirical data suggesting actual instances exist of divided upper hemispheres resulting in divided consciousness[175], the first reason has been widely rejected[176]. And though it has been

[174] Cf. Parfit 1984, 255.

[175] Cf. Nagel 1971.

a subject of great controversy in the literature, the second reason is standardly dismissed on the grounds that this fact about human beings plays no role in our central beliefs about the nature of persons[177].

Contra the defenders of fission, I think the problems begin before we reach (13). It is important to notice that there is a great deal of arbitrariness in treating certain presumably contingent features of human beings as fixed, and others as malleable. Certainly it seems to be the same *sort* of biological fact that the basis for our psychological activity is localized in the brain (9), and that the basis for our linguistic ability is localized in one of the hemispheres (the denial of (12)). It is far from trivial to suppose that all psychological characteristics might be realized in the brain twice over (12), and far from trivial to suppose that the brain stem and spinal cord would enjoy the same sort of duplicatory structure (13). It would be at least as plausible to suppose that the source of psychological activity might be totally non-localized (the denial of (10)), so that the brain transplant with which the fission case begins would produce nothing but confusion (the denial of (11)). If one were to argue that a non-localized center of psychological activity would run counter to evolutionary laws, which suggest that the center of intelligence should be located in a

[176] I remain unconvinced, both by the evidence Parfit himself adduces, and by the other articles I have read on the subject, that these cases describe a genuine division of consciousness of the sort necessary to support his conclusion. Cf. Robinson 1988, esp. 325; Putnam 1981, 89-92. But since I think Parfit's argument is problematic for other reasons, I will concede this point for the sake of argument.

[177] Both Robinson 1988 and Wilkes 1988 disagree, as does Wiggins 1980; but see Johnston 1989, 376-377; Garrett 1990, 178-180 for replies.

reasonably compact, well-protected area of the body, one would face equal difficulties explaining how duplication of *all* psychological features (or even of a significant proportion of high-level psychological features[178]) could be countenanced on evolutionary grounds[179]. Such mechanisms are extraordinarily expensive, far beyond the purchasing power of a satisficing spender like natural selection. It seems that (12) is already counter-nomological, and that (13) is even more so.

But for the sake of argument, let us grant to the defender of fission that we might be creatures of the kind described in (13). What I will argue below is that *even if it should turn out that we are such creatures*, we would not consider the creature described in (13) to be relevantly like us in the sense that (8") requires. That is, even if it turns out that we are creatures for whom it is not just metaphysically possible—but also medically possible—that we should divide, we should not conclude that identity is not what matters for prudential concern. What matters for prudential concern might be connected to certain contingent facts about the way things (merely) happen to be.

4.4 Two Unsuccessful Strategies

Parfit contends that there is an asymmetry between identity on the one hand, and what matters for prudential concern on the other. The former is extrinsic in the way described above; and if Parfit is right that the crucial premise is true, the latter is

[178] I thank Hilary Putnam for pressing me on this point.

not. We have conceded that the crucial premise is true in the necessary condition sense; identity is not what among the conditions necessary for rational prudential concern. So the question is whether prudential concern *in the explanatory sense* is an extrinsic relation. How might one go about answering this question?

4.4.1 An Unsuccessful Attack on the Crucial Premise

A first strategy would be to challenge Parfit's argument by denying the crucial premise outright, on the grounds that what prudentially matters is, like identity, just one of the many relations that is extrinsic. In "Reasons and Reductionism," Mark Johnston considers this strategy, though he ultimately rejects it. Johnston writes:

> Whence the plausibility of the crucial principle that whether one has reason to be specially and directly concerned about some future person depends only on the intrinsic aspects of the relation between oneself and that future person? Certainly not from the plausibility of the general claim that extrinsic features do not matter. We often take extrinsic features to be highly relevant to how we evaluate some fact or relation (Johnston 1992, 609).

Johnston goes on to cite the examples of "exclusive ownership, winning, unique achievement, and intimacy" (cf. Sosa 1990, 319-320). With the possible exception of intimacy, all of these are clearly examples of the sort of extrinsic relation Parfit takes to be paradigmatic. Exclusive ownership obtains when I bear an intrinsic relation of ownership to some object, and no one else bears the same relation; winning obtains when I reach the finish line and no one else has yet done so; and unique achievement obtains when I alone succeed in doing thus-and-such.

[179] This is not the same thing as suggesting (as has empirically been demonstrated)

Continues on next page…

But it is hard to see how these analogies could convince anyone who does not already accept that rational prudential concern is a relation that can only hold one-one. Indeed, it is hard to see how they could be relevant, unless one were to *stipulate* that whether my relation to some future person gives me reason for prudential concern depends on the relation holding exclusively. Nothing about the concept of prudential concern seems to tell one way or the other[180]. And nothing about the analogous examples seems to provide *reasons* for thinking that prudential concern is a relation with an explicit exclusiveness clause. So while appeals to these sorts of analogies may remind us that there are a range of relations which are extrinsically determined, as a means to establishing the crucial premise, they are question-begging at best.

4.4.2 An Unsuccessful Defense of the Crucial Premise

But the simplest argument for the view that what prudentially matters is an *intrinsic* relation is equally question-begging. One reason one might think that fission preserves what matters for rational prudential concern is because being one of the two survivors of a fission transplant would feel, from the inside roughly like being someone who wakes up from a complicated operation. But whatever other reasons there are for accepting Parfit's analysis,

that certain memory traces are stored at multiple locations in the brain.

[180] Cf. Sosa: “We can see the sorts of value that would be threatened by having too many spouses; but I for one have no inkling of what important values would be endangered by [division]...it is not easy to see what values would be endangered in

Continues on next page…

the fact that a situation would seem, from the inside, qualitatively identical to ordinary survival does not give a reason to think that it is just as valuable.

This can be vividly illustrated by realizing that as far as perception "from the inside" goes, being the product of fission is no better or worse than being hooked up to what Nozick calls an *experience machine* (see Nozick 1974, 42-45; see also sections **1.3.1** and **1.3.2** above), that is, a machine which would stimulate the brain and produce whatever set of experiences one would have in an ordinary life. What Nozick's example plausibly shows is that something that appears from the inside to be worthwhile—such as seeming to undergo a certain sort of experience—might turn out upon reflection not to be nearly so valuable. So to the extent that intuitions converge in the fission case *because* readers imagine the case from the inside, these intuitions are not informative. Imagining the experience machine case from the inside would produce the same convergence of intuitions. If we would not trust them in the latter case, we should not trust them in the former.

A first retort to this objection might be the following. That a non-standard form of survival would feel from the inside like ordinary survival is at most a necessary condition of our valuing it. Against a background of veridicality, however, feeling the same from the inside becomes a much stronger indication of reasonable valuation. That is, if it feels the same from the inside, *and* it gives us accurate information about the world, then our

such cases *except* only for the true survival of the mainstream protagonist" (Sosa

Continues on next page…

desire to have this as a result of normal processes may be only derivative. Parfit seems to have something like this in mind when, in contending that it is irrational to prefer that continuation have its normal cause, he appeals to the following analogy:

> Consider artificial eyes which would restore sight to those who have gone blind. Suppose that these eyes would give these people visual sensations just like those involved in normal sight, and *that these sensations would provide true beliefs about what can be seen*. This would surely be as good as normal sight. It would not be plausible to reject these eyes because they were not the normal cause of human sight (Parfit 1987, 285, italics added).

That is, Parfit argues that we value our eyes only derivatively—as a source of visual sensation; that sight have its normal cause is not what matters to us. What does matter to us is that "these sensations would provide true beliefs about what can be seen." Similarly, what we care about in the case of continuation is not only that there will be someone whose inner life will feel the same to her as my inner life does to me, but also that her experiences will be, for the most part, veridical.

However, this response is not sufficient to draw a sharp line between the fission case and the experience machine case. Assuming that what we want when we require veridicality of experience includes memories of past actions as well as sensations of present ones, both forms of continuation have a non-veridical component. Just as every post-hookup experience of the person connected to the machine is false, so too, on Parfit's view, is every I-

1990, 311).

involving memory of Lefty in the two-sided case[181]. So Parfit's analogy is inadequate. In the case of artificial eyes, both function and veridicality are preserved; in the case of fission, we have only the former. Against a background assumption of accuracy, phenomenological identity might justify equal valuation. But no such background conditions hold in the fission case. So the judgment that bearing a relation of prudential concern is rational for intrinsic reasons will be reliable only if it rests on more than imagining the case from the inside. It will have to account for the fact that veridicality of memory, normally valued for its own sake, does not matter in this case.

Of course, Parfit would retort that this is begging the question: the appropriate consideration, he would counter, is that Lefty's *quasi*-memories be veridical, that is, that they accurately reflect *someone's* actual experiences. Indeed, they don't simply reflect *someone's* experiences, they reflect the experiences of someone to whom he is related as a continuer, namely Brainy. Surely, Parfit would say, this is close enough to what we wanted when we sought veridicality[182]. But at this point we have reached a stalemate: quasi-

[181] This is so despite the fact that Lefty's memories have their normal causes. Because Brainy has two continuers, *neither* of them is identical with Brainy (barring a multiple-occupancy view). So every (quasi-)memory that Lefty has that involves strict I-thoughts is, strictly speaking, false. One might try to get around this by replacing I-thoughts with I*-thoughts (see Rovane 1990), but this would be no help in terms of settling the question at hand. Precisely what is at issue is whether I*-thoughts are "just as good" as I-thoughts. (I thank Terry Irwin for showing me the need to include such a footnote.)

[182] I thank Nick Sturgeon for pressing me on this point.

memory “counts” only if identity doesn’t. And nothing about imagining the case from the inside can resolve that question one way or the other.

4.5 Deeper Reasons

4.5.1 Why is the Fission Argument So Compelling?

In the remainder of the chapter, I want to try to get at what I think are the deeper reasons that the fission argument seems so compelling. Doing so will allow me to bring out why our responses to such cases are so tricky to interpret. There is a complicated interplay between trying to make sense of the scenario on its own terms, and trying to make sense of the scenario in a way that illuminates our beliefs about ourselves. In the case of fission, these aspects are extremely hard to disentangle.

What I will suggest is that there are two ostensibly undeniable principles which play a tacit role in making the fission argument seem as forceful as it does. I will try to show that while both of these principles may be legitimately invoked when what is at issue are the necessary conditions for prudential concern, neither can be appealed to when what is at issue is what *explains* it. But the fact that it is so hard to see this when we contemplate the case gives us reason to be wary of the methodology.

4.5.2 The First General Principle

The first general principle—which in decision theory is called the *independence of irrelevant alternatives*—is that my ranking of A and B should not change simply because C is introduced as an additional alternative[183]. This principle is so fundamental that to deny would seem like a joke.

Indeed, there is a joke about the denial of this principle that runs as follows. The waiter comes to the table and asks the customer: “Would you like tuna salad or egg salad?” The customer answers: “I’d like egg salad.” A few minutes later, the waiter returns to the table. He says, “I forget to mention that we also have grilled cheese.” “Oh,” says the customer, “in that case I’ll take tuna salad.”

The reason this joke is funny is that it seems completely irrational to change one’s rankings of A and B just because C came into the picture. There would be no joke if the customer said: “Oh, in that case I’ll have grilled cheese.” One’s rational preferences can certainly change in light of novel information: to change one’s order to grilled cheese when one is newly aware of its availability is perfectly sensible; what seems crazy is to change one’s order from egg salad to *tuna salad* because grilled cheese is also available.

The connection of this to the fission example is the following. Let’s call the choice of ordinary survival *A,* and the choice of survival-as-Lefty—in a way whose details are to be specified—*B.* Now take one version of B, namely the one-sided case. In this case, Brainy’s continuation as Lefty clearly contains what matters for prudential concern, just as Brainy’s ordinary survival does; in both cases, the person towards whom Brainy bears prudential concern is *himself*. As far as prudential concern goes, Brainy ranks A and B on a par; that is he ranks continuation as Lefty in

[183] Cf. Luce and Raiffa 1957, 127.

the one-sided case and survival in the ordinary sense equally highly. But according to the first general principle, Brainy's ranking of these two options should be unaffected by whether a third option, survival as Righty, is also available. His ranking of A and B should not change simply because C is introduced as an additional possibility[184].

Parfit's opponent seems to be asking Brainy to violate this principle. The opponent accepts (2") while denying (4"). That is, she accepts that what matters for prudential concern is present in the one-sided case while denying that it is present in the two-sided case. But this seems tantamount to suggesting that it would be rational for Brainy to change his rankings of A and B just because C is now in the picture. To say that it would be rational for Brainy to bear a relation of prudential concern

[184] But isn't the following an immediately obvious relevant difference between the cases? In the one-sided case, Brainy's survival-as-Lefty takes the form survival-as-himself, whereas in the two-sided case (that is, the case where C is introduced as an additional option), Brainy's survival-as-Lefty takes the form survival-as-someone-else. To suggest that C is not relevant in this case would be like suggesting there is no difference among the following cases:

Case 1: A=[gain $5000]; B=[gain $1M (in a way to be specified)].

Case 2: A=[gain $5000]; B=[gain $1M]; C=[gain the $1M by winning the lottery].

Case 3: A=[gain $5000]; B=[gain $1M]; C=[gain the $1M as a pay-off on a loved-one's life-insurance policy].

Surely one might rationally prefer B to A in case 2, but prefer A to B in case 3. The reason, of course, is that it makes a great deal of difference how B is, so to speak, cashed out.

This suggests that the mistake made by Parfit and his allies is in fact a failure to see that the specification of *how* Brainy survives as Lefty may be relevant to the question of whether Brainy's relation to Lefty contains what matters for prudential concern.

This may be so. But if it is so, it is a consequence of a prior commitment to the crucial premise. Once the crucial premise has been endorsed, nothing about the details of what survival-as-Lefty consists in can be relevant to its valuation (except

Continues on next page…

towards Lefty *unless* Righty happens also to exist seems to violate the general principle we have identified. If so, the opponent seems to be defending an indefensible position.

But consider the following apparent counter-example to this seemingly undeniable rule. Time has come to order dessert, and the waiter asks me: "Would you like Ben and Jerry's ice cream or homemade apple pie?" I answer: "I'd like ice cream." A few minutes later, the waiter comes back and says: "I forgot to mention that we also have homemade pumpkin-hazelnut torte with a ginger lime cream sauce." "Oh," I reply, "in that case I'll have the apple pie." The waiter giggles. "No," I insist, "it's not funny. You need to know the following three things about me: I don't much like pie, I am allergic to hazelnuts, and I am something of a gourmet. Faced with the initial choice, I selected ice cream because I don't generally care for restaurant pie. But upon hearing about the hazelnut torte, I gained the information that this chef is something special. Hence I decided that it would make more sense, given my desires, to order the apple pie[185]."

On closer inspection, of course, this is not a counterexample to the first general principle; it is not a case where my *criteria for ranking* change in light of the new information. Rather, it is a case where I come to realize that I have misranked the

insofar as these details concern the intrinsic relation between Brainy and Lefty). (I thank Nick Sturgeon for this objection, and for the analogy.)

[185] A similar example, involving steak, salmon, snails and frog-legs, can be found in Luce and Raiffa 1957, 288. They point out that the example "illustrates the important assumption implicit in [the principle], namely, that adding new [options] *does not*

Continues on next page…

available options *given* my criteria. When I ordered the ice cream I was *making a mistake*. I thought I knew enough about A and B to judge between them. But upon hearing that hazelnut torte was also an option, I gained new information *about B*, namely: chances are that, like its fancier cousin, this apple pie is a *special dessert*.

So the example does not show that information about some local fact can rationally bring me to reorganize my preferences. But it does show that information about some local fact can make me *aware* of some global feature of a situation that is relevant to my evaluation of the situation as a whole, including my evaluation of the relative merits of A and B. And while the basic principle remains unchallenged, we now have a sense of where its vulnerability lies. Ascertaining whether an alternative is indeed irrelevant is far more difficult than it might initially have seemed. Apparently extrinsic features may play a role in the ranking of preferences without the first general principle being thereby violated. So if the appeal of the crucial premise can be traced to the fact that its denial seems to violate this principle, we are back to a state of truce. It may be that rejection of the premise that prudential concern is an intrinsic relation only *seem*s to run counter to the principle we have been discussing.

4.5.3 Another Way of Being Extrinsic

Thus far, we have only considered relations that are extrinsic because they hold exclusively, such as winning, unique achievement, and being the only son of.

alter one's a priori information as to which is the true state of nature" (Luce and

Continues on next page…

These are relations whose extrinsic determination stems from the fact that they hold of X only in the absence of competitors. But there are also relations which are extrinsic not in this local sense, but in a more global one: relations where whether the relation holds between X and Y depends both on the intrinsic sub-relation that holds between them, *and* on the existence of some contextual factor that makes that particular intrinsic relation "count" as the relation it does. Here is an example of the sort of case I am thinking of. If Ali says to Ayesha three times: "In the name of Allah, I divorce you," Ali is divorced from Ayesha *...if* the two live in a country governed by Muslim law. But if they have recently emigrated to the United States, his utterance of these words will not be sufficient to make his relation to Ayesha that of ex-husband. Nonetheless, the *intrinsic* relation between Ali and Ayesha is exactly the same in the two cases. When Ali says these words to Ayesha in their apartment in Cleveland, both of them *believe* that the words are sufficient to bring about the dissolution of their marriage. *From the inside*, there is no way to tell whether or not the act has been sufficient; it is only by "zooming out" to the circumstances surrounding our main characters that the true impact of Ali's utterance can be ascertained[186].

What I want to argue is that what prudentially matters is an extrinsic non-exclusive relation of this sort—but with a twist. The contextual factor that makes the intrinsic relation "count" as the relation it does—namely as a relation of *rational* prudential concern—is the background truth that the facts of life guarantee that I will

Raiffa 1957, 288; italics in original).

[186] This is true of all cases of play-acting as well. The phenomenon I am describing is just a special case of the more general problem of context-dependency.

Continues on next page…

have at most one continuer[187]. Rational prudential concern, that is, has its exclusiveness built in at the *beginning*, not at the *end*[188]. Let me recapitulate where we stand so far, and then make a positive argument for this claim.

[187] Note that there is an asymmetry between the cases. In the divorce case, if the words are uttered in a context where Islamic law is not in force, they are not sufficient for (state-recognized) divorce. This is true even if most of the time when these words are uttered, they *are* sufficient for divorce (because they are ordinarily uttered in countries governed by Muslim law). By contrast, what I will say about prudential concern is that if, most of the time, when X bears rational prudential concern for Y the E-features are present, then even when the E-features are *not* present, prudential concern may be rational. (I thank Terry Irwin for pointing out this disanalogy.)

The disanalogy does not undermine the use to which I am putting the example (indeed, a perfectly parallel case would be less illuminating), which is to bring out that there are certain relations such that whether that relation holds between X and Y depends on:

(a) intrinsic features of the (sub)relation between X and Y, and
(b) certain contextual factors that affect whether that intrinsic (sub)relation "counts" as the relation in question

In the Islamic divorce case, (a_I) is the thrice-fold utterance of the words "In the name of Allah I divorce you" and (b_I) is the presence of the parties in a country governed by Muslim law. In the personal identity case, (a_{PI}) is that the relation between X and Y is intrinsically the same as a relation of identity; (b_{PI}) is the fact that the world guarantees that I will have at most one continuer. What happens when (b) is not satisfied? In the Islamic divorce case, there is zero-tolerance; if (b_I) does not obtain, the divorce does not go through. In the personal identity case, there is slightly more flexibility; in certain very special (purely imaginary) circumstances, even if (b_{PI}) does not obtain, rational prudential concern might still be justified. So the disanalogy between the two cases concerns the way in which (b) is specified, and not the structure of the relations under consideration.

Alternatively, one might think of (b_{PI}) in the personal identity case as saying: (b_{PI}') the world guarantees *with almost complete certainty* that I will have at most one continuer. And here the analogy can be pressed without difficulty. As I will argue below, if (b_{PI}') is false (that is, if the world does not guarantee with almost complete certainty that I will have at most one continuer), then (a_{PI}) will not be enough to give us rational prudential concern.

[188] Cf. Johnston: Fission "violates a presupposition of our future directed self-concern by providing more than one future person to continue an earlier person's mental and

Continues on next page…

4.5.4 Recapitulation

We are trying to figure out why one might believe the crucial premise that what matters for prudential concern in the explanatory sense is an intrinsic relation. In section **4.4**, I discussed two arguments, one in favor of the premise and one against it; both were unsuccessful. Simple argument by analogy to show that rational prudential concern is possible only in the absence of a competitor involves the question-begging assumption that identity *is* what matters. Nothing about the *concept* of rational prudential concern seems to have built into it this sort of from-the-outside exclusiveness constraint. But I then showed that the simplest argument for the contrary view—that "what it feels like from the inside" is sufficient to make prudential concern rational—can be established only by assuming equally question-beggingly that identity is *not* what matters in the explanatory sense.

I then suggested that the appeal of the crucial premise might be traced to the fact that its denial seems to violate the principle that one's rankings of A and B should not change in light of information about C. But I went on to point out that extrinsic features can be relevant to the evaluation of a fact or relation in a global as

bodily life" (Johnston 1992b, 603). Or again: "The very case of fission itself undermines essential unity, violates the presupposition the one will have at most one continuer, threatens the ordinary idea that only intrinsic features matter to identity, and so undermines the basis for the principle that only intrinsic features can matter" (Johnston 1992b, 609). See also Gale 1991.

well as a local sense, and I suggested that what prudentially matters might have such a structure. Like Islamic divorce, it might depend on the presence of some contextual factor that makes an otherwise sufficient intrinsic relation "count" as the relation it does. What I will try to show next is that this diagnosis can be used to help defend the apparently implausible claim that even in circumstances where prudential concern is rational *without* identity, identity is still what matters for prudential concern.

4.5.5 The Second General Principle

On its surface, this suggestion seems to violate a principle of rationality at least as fundamental as the principle of irrelevant alternatives. This principle, which Mill called the *method of agreement*, says that: "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon" (Mill I:451)[189]. To see how counterintuitive it is to deny this principle, consider the following scenario: Whenever I strike a match against the side of a matchbox and say "let there be light," the match bursts into flame; whenever I strike a match against the side of a matchbox and say nothing, the match bursts into flame; and whenever I simply hold the match in the air and say "let there be light," the match remains unlit. It would be insane for me to conclude from this that in the first case—when I strike the match and utter the phrase—that it is what I have *said* that has caused the match to burst into flame. If, whenever we have Y alone Z occurs, but whenever we have X alone Z does not occur, then we seem rationally compelled to conclude that when we have X and Y together and Z occurs, it is Y rather than X that does the explanatory grounding.

This is true not only in cases where what we are trying to explain are physical phenomena, but also when we are trying to explain our attitudes. Suppose you ask me: “What is it that you like so much about the Star Spangled Banner: the tune, or the text?” And I say: “Hmmm... when I hear the words sung to a different tune, I always like that song exactly as much as I like the Star Spangled Banner, and when I hear the tune of the Star Spangled Banner with any other text, I don’t like it at all. So I guess I’d have to say that what I like about the Star Spangled Banner itself is the tune.” As before, such a conclusion would seem insane[190].

[189] Strictly speaking, the principle which concerns me is a cousin to this principle; I am concerned with justification rather than cause and effect.

[190] But isn’t this a counterexample? I love to eat bread-and-butter sandwiches. And I love to eat bread on its own. But the idea of consuming a plain pat of butter repulses me. Now suppose you ask me: “What do you like so much about bread-and-butter sandwiches?” And I reply: “I’d have to say what I like so much is the butter.” Surely this is *not* a crazy answer. Why not? Because I have, presumably, spoken non-literally. What I like so much about bread-and-butter sandwiches is the butter *when it is on the bread*. That is: what I like is the bread-and-butter *combination.*

But if I spoke truly when I said that I like plain bread just as much as I like bread-and-butter, aren’t I committed to saying that what I like about bread-and-butter sandwiches is the *bread*? The answer is *no;* it is possible for me to like bread alone exactly as much as I like bread-and-butter, *but* for the reasons to differ in the two cases. Perhaps the reason I like plain bread is because it allows me to concentrate on the subtle texture of the loaf, whereas the reason I like bread-and-butter is that the butter brings out nuances of the bread’s flavor. But if we allow this, don’t we have to allow the possibility that what makes the match burn when I am silent is the fact that I strike it against the edge of the box, but that what makes the match burn when I strike it while saying “let there be light” is the *combination* of my words and my deed?

If the answer is *no*, then there must be some relevant difference between the bread-and-butter case and the match-and-incantation case. The difference is this: In the bread-and-butter case, even though I value the bread-and-butter the same amount

Continues on next page…

How does this connect to our discussion of fission? For the time being, let us grant Parfit that what personal identity consists in is non-branching psychological continuity with the right sort of cause, which Parfit calls Relation R. That is, Y is identical to X iff Y bears relation R to X, and no one else bears relation R to X. We have already conceded that whenever both relation R and identity obtain, rational prudential concern is warranted. And we have also conceded that in the two-sided case—where relation R obtains *without* identity—rational prudential concern is warranted. And given the analysis of identity we have granted to Parfit, we have also conceded that rational prudential concern is never warranted in the absence of R. So if we apply Mill's method of agreement to this case, it seems clear that even when both identity and relation R obtain, it is relation R that matters for prudential concern. How, then, can I possibly claim that when both identity and relation R obtain, it is *identity* that matters[191]?

as I value the bread, I don't value it in the same way. *Qua* source of texture, I prefer bread alone to bread-and-butter; *qua* source of taste, I prefer bread-and-butter to bread alone. So there is no mystery: if the question concerns what I value about bread-and-butter when what is at issue is the taste, then the right answer is "the combination." But of course, in such a circumstance I also prefer X-and-Y together (that is, bread with butter) to either X or Y alone. If, by contrast, the question concerns what I value about bread-and-butter when what is at issue is the texture, then the correct answer is, "the bread." So again there is no violation of the principle, for here I prefer X alone to X-and-Y together, and X-and-Y together to Y alone. (Or, to change the case a little, if *qua* source of texture I value X and X-and-Y equally, then again it would be correct to say that I value X-and-Y *because* of X; and again, the principle would not be violated.)

[191] Cf. Parfit: "In all ordinary cases, personal identity and [R] coincide. When they diverge, [R] is what matters. That strongly suggests that, in all cases, [R] is what matters...If, when two facts come apart, one of them is what matters, why think the *other* is what matters when they coincide?" (Parfit 1994, 38).

Continues on next page…

4.5.6 How Absent Features can be Explanatory

Note first that there is something slightly misleading about putting the case the way I just have. Identity isn't something different from relation R, in the way that saying "let there be light" is something different from striking the match against the side of the box. The fact that a relation of personal identity holds just *is* the fact that relation R holds in a non-branching fashion. So I propose that we frame the situation as follows[192]. There are three features at play: a basic feature, in this case *being R-related*; an enhanced feature, in this case *being identical*[193]; and a defeating feature, in this case *there being multiple R-related candidates*. The basic feature is determinable in different ways[194]. If the defeater obtains, we might speak of the basic feature obtaining "merely" or "barely." So if there are multiple R-related candidates, then Brainy and Lefty would be *merely* R-related. On the other hand, if the enhanced feature obtains (which is the same thing as saying that the defeating condition does not obtain), then we might speak of the basic feature obtaining "specially." So if there

[192] Cf. (although her point is somewhat different) Susan Hurley: "Perhaps the failure of preferences [to meet the desired constraint] suggests...that we have not yet drawn the right distinctions among the various objects of preferences...Perhaps the distinctions between different values, or objects of preference...are bound up with the applicability of an independence condition. If the relationship between the values in one set seems not to be characterized by Independence, then you haven't individuated the values correctly; go back and try to find some other way of interpreting the objects of preference so that the condition does hold" (Hurley 1989, 74).

[193] That is, being uniquely R-related.

[194] I thank Jennifer Whiting for suggesting that the point be put this way.

Continues on next page…

are *not* multiple R-related candidates, then Brainy and Lefty would not be merely R-related; they would be *identical*.

What Parfit contends is that since prudential concern can be rational even if R holds on its own, what matters for prudential concern is the mere obtaining of the basic feature. I will argue that this move can be blocked. In certain cases where it is rational to bear a certain attitude in the face of the mere obtaining of the basic feature, what makes that attitude rational are facts about the *enhanced* feature[195/196].

[195] Cf. McDowell: “There is no reason to assume that what is left when [a] requirement is dropped will stand on its own as an adequate explication of a concept, even though the result is necessary and sufficient conditions for its application. It takes more than an arithmetic of subtracting necessary conditions to guarantee us an autonomously intelligible concept” (McDowell forthcoming, 17). Cf. also his discussion of the “highest common factor conception” in McDowell 1994.

[196] Robert Nozick raises the following objection. We might characterize the case I have described as follows: the basic feature is P, the defeater is Q, and the enhanced feature is merely P-and-not-Q. But then the enhanced feature is “enhanced” only in the most attenuated sense of the word; all I have described is a case where we have a particular property (P), and a defeater (Q) which may or may not be present. I seem to be arguing that the basic content be treated (by courtesy) in the same way as the enhanced content, despite the presence of the defeater. But in the case I have described, the defeated-enhanced content is just the same thing (that is, P) as the basic content.

I have two replies. First, it is not true that basic-P and defeated-enhanced-P are the same, even though both turn out to have the same characteristics. Basic-P is indeterminate between P-and-Q and P-and-not-Q; defeated-enhanced-P is definitively P-and-Q. So there *is* a difference between the basic content and the defeated-enhanced content. Second, the *explanation* for the difference between the two cases is that the properties of basic-P are determined locally (by looking at P and its intrinsic relations), while the properties of enhanced-P and defeated-enhanced-P are determined non-locally (by looking at P and its *extrinsic* relations). The *reason* basic-P is indeterminate between P-and-Q and P-and-not-Q is because in order to determine whether Q obtains, we need to look at something other than P. So to say that we might treat the defeated-enhanced content in the same way as the enhanced-content is *not* to say that the defeater is irrelevant; rather, it is to say that once we are looking at

Continues on next page…

Here is an example. The basic feature is: *being a human body*; the enhanced feature is *being a living human body*. To the extent that respect is an attitude constrained by rationality, it is surely rational to treat a living human body with respect. Even in cases where our actions will not cause pain to another, we bear certain obligations towards living human bodies that preclude our treating them with wanton disregard. It is also rational, to the extent that respect is an attitude constrained by rationality, to treat a *non-living* human body with respect. Again, we bear certain obligations towards non-living human bodies that preclude our treating them with wanton disregard. *But*, I want to claim, what makes it rational to treat a non-living human body with respect is that it is rational to treat a *living* human body with respect, and not, as the standard analysis would seem to suggest, the other way around[197].

extrinsic features, issues concerning general patterns of relations may come into play in the proper evaluation of the situation.

[197] One might object to this example on the grounds that what deserves respect is not a *non-living* human body, but rather a *dead* body—one that once embodied the life of some particular person. And having-been-the-embodiment-of-a-particular-person is not a feature that is *absent* from the non-living body; rather, it is a property that the body has in virtue of its causal history. (I thank Terry Irwin for this objection.) I reply as follows. It is true that in all actual cases, non-living human bodies are dead bodies, and it is true that dead bodies have a certain causal history which in part explains the respect we have for them. But *because* this is so, a body could not be *human* and *non-living* without being *dead**. What this means is that this apparent objection does not tell against my main point. What I need to show is that some feature that is not present in the non-living body case—namely, being alive—explains the rationality of our attitudes towards that body. That *not being alive* coincides with *being dead* for human bodies means that in addition to being describable as a body from which life is absent, a non-alive body can also be described as a dead body, that is, a body that *has been alive*. The difference is that having-been-alive is a feature that the dead body has, whereas being-alive is a feature that it does not have. And if it is having-been-

Continues on next page…

So I have shown that there are cases where a feature that is not present does explanatory work. With this in place, I am much of the way to my conclusion. All that remains to be demonstrated is that rational prudential concern is such a relation.

4.6 Conclusion

4.6.1 Exceptional Cases and Conceptual Change

Since it is easy to forget when discussing issues of personal identity, let me begin this section by reminding you that—as a matter of fact—human beings come into existence only through the sort of process I described in the first paragraph. That is, in all *actual* cases relation R holds only in conjunction with identity. What I will argue in this final section is that it is *because of this* that we are right to accept the intermediate claims: that in the two-sided case, Brainy's relation of prudential concern to Lefty would be rational. That is, it is because, I am assuming Parfit's

alive that does the explanatory work, then it seems that I have not described a case where it is the enhanced rather than the basic feature that matters; it seems that all I have done is to describe a case where some present feature (having-been-alive) matters. Indeed, one might even argue that what the case shows is that the relevant feature is having-ever-been-a-living-human-body, a feature which is present *both* in the live body case and in the dead body case. Given the point I am trying to make, however, this is not a problem. That having-ever-been-a-living-human-body applies to *all* human bodies means that it is *being living* that, at base, *explains* the rationality of our valuing bodies as such. So the feature that does the explanatory work is, indeed, a feature that is not present in the basic case.

*It is perhaps conceptually possible that something could be a non-living human body that is not dead, were it to be *not-yet-living*, but my sense is that what we mean by "human" precludes this possibility. Even if not, this does not affect my reply to the objection.

fission example to be an *exceptional* case, that we are able to extend our concepts to it—as a courtesy[198].

Consider the following analogous case. The basic feature is*: being a recording that sounds exactly like a recording of some great cellist playing Bach*[199]. And the enhanced feature is: *being a recording of some great cellist playing Bach*[200]. Suppose that, as with R-relatedness, there are no actual instances of the basic feature alone: nothing sounds exactly like a recording of a great cellist playing Bach that is not, in fact, a recording of a great cellist playing Bach. What I want to argue is that it would be rational to value the sound-alike just as much as we value the actual recording, but that the *reason* it would be rational to value the sound-alike is *because* it is rational to value the actual recording as we do. That is, as in the dead body case,

[198] Cf. Johnston: "[T]he most that follows [from Parfit's arguments] is that self-concern might be sensibly extended to certain bizarre cases, were the cases ever in fact to arise" (Johnston 1992, 600).

[199] One might suppose that the recording came into being as the result of a bunch of computer sound technicians in a CD studio playing around with the equipment, each of them controlling some range of frequencies.

[200] The case I am imagining differs somewhat from the one described by Nelson Goodman in "The Perfect Fake" (Goodman 1976, 99-102). Goodman supposes that "we have before us, on the left, Rembrandt's original painting *Lucretia* and, on the right, a superlative imitation of it. We know from a fully documented history that the painting on the left is the original; and we know from X-ray photographs and microscope examination and chemical analysis that the painting on the right is a recent fake" (Goodman 1976, 99-100). Goodman is concerned with the question of whether there could be "any aesthetic difference between the two pictures for *x* at *t*, where *t* is a suitable period of time, if *x* cannot tell them apart merely by looking at them at *t*" (Goodman 1976, 102). In my example, there is—by stipulation—no *qualitative* difference between the authentic recording and the sound-alike; the only difference between them is their causal history. Nonetheless, aspects of Goodman's discussion are relevant to my point. See especially Goodman 1976, 105-106.

Continues on next page…

what matters in an explanatory sense for our valuing the sound-alike recording is a feature—having thus-and-such a causal history—which is not present.

First: why would it be rational to value the sound-alike recording even if it is not actual[201]? The answer, presumably, is that it would provide us with precisely the same sort of aesthetic pleasure that the actual recording does; experientially, hearing the sound-alike would be exactly the same as hearing an actual recording. By stipulation, there is no qualitative difference between the two; no matter how perfect one's musical sense, there would be no discernible discrepancy between them.

But why isn't this tantamount to saying that, in the basic case, what matters is what the recording sounds like? Indeed, why isn't this tantamount to saying that in the *enhanced* case, what matters is what the recording sounds like? Why am I not saying that what matters for our valuing the *actual* recording is that we value the sound-alike? The reason is the following.

[201] A puzzle: Nelson Goodman distinguishes between works of art that are *autographic* and works of art that are *allographic*. Autographic works are works for which "even the most exact duplication of it does not thereby count as genuine;" allographic works are works which are not autographic. Goodman contends that painting is autographic (even a perfect replica of a Rembrandt is not a Rembrandt), whereas music is allographic (even a beginning pianist's performance of a Chopin Etude is a performance of Chopin). Goodman concedes that "there may indeed be forgeries of performances...that purport to be by a certain musician etc.; but these, if in accordance with the score, are nevertheless genuine instances of the work" (Goodman 1976, 113; cf. 118). In my example, however, even though the sound-alike is "in accordance with the score," I think it is not a "genuine instance of the work." Nor is it properly speaking a *forgery*. "A forgery of a work of art is an object falsely purporting to have the history of production required for the (or an) original of the work" (122). But there is no sense in which the sound-alike that I describe *purports* to have one history or another. Perhaps the difficulty we have in classifying this case

Continues on next page…

In general, we believe that we can tell the difference between something that is produced by normal means and something that is produced by abnormal means: a hill that was produced by glacial forces and a hill that was produced by mining; a painting that was produced by Rembrandt and a painting that was produced by an assistant. We think, in general, that how an entity came into being affects its value: a forgery isn't as good as an original. But why not? What if, no matter how hard we tried, we couldn't tell the difference? What if they are genuinely qualitatively identical? Is it then irrational to care about the process by which the entity came into being?

One reason it might not be is if we think it is valuable that someone perform the actions that bring the entity into being via the normal route; that is, we might care about how it came into being for *valuational* reasons, because we care about the process itself. On these grounds, we might prefer homemade brownies to brownies from a bakery, or a union-made shirt to one produced in a non-unionized factory. A second reason we might care about process is that we might think it is valuable that we respond to the world in a way that tracks what is actually going on. So we might care about how an entity came into being for *epistemic* reasons, because we want to know that we can take certain things as *evidence*. On these grounds, we might be disturbed by apparent photographs produced by means of Digitized Image Technology, or by particularly effective forgeries.

can be traced to the fact that it violates our ordinary assumptions about the line between the artificial and the natural.

In the music case, both sorts of reasons seem to be at play. In all *actual* cases, music of this kind is produced by great musicians playing pieces composed by great composers on instruments created by great instrument-makers. Why might we value such recordings differently if this were not so? The most obvious reason is that in such circumstances, people would be less likely to practice the cello or compose music or craft high-quality instruments, and we might think these are valuable endeavors in themselves. But it is a mistake to take this reason as the only one. Additional explanatory work is done by the second reason: that under such circumstances, we would no longer be justified in taking a great recording as evidence for the fact that a great musician had produced it. In the situation I have described, the world would be such that we could be categorically fooled. But, as I will argue in the next section, such a situation would be intolerable. Our values and concepts are too closely tied to the way things are in the world for such a global mismatch to be tenable.

4.6.2 The Final Step

If apparent recordings were regularly produced by means that we now consider unnatural, there would be general de-coupling of things that sound like recordings of great cellists playing Bach from things that *are* recordings of great cellists playing Bach. If you accept that under such circumstances we might value music less—or at least differently—then I have a foothold for my final argument.

Here is the argument. If we say that our valuing the basic feature of what the recording sounds like is what explains our valuing the enhanced feature of the recording actually being of such-and-such, then only way we can account for such a change in value would be on the grounds that there would likely be less music-producing activity. But this doesn't seem to capture the extent of the transformation. Surely it is not *only* because there would be less cello-practicing that our attitudes would change. On the other hand, if we say that the enhanced feature—being an actual recording—is what explains our valuation of the basic feature, then we can account for why it is that in the face of such global disruptions, we would be inclined to rethink our valuational scheme. We can account for the fact that in a world where *all* music was produced in that way, it might be rational to value music less than we do, even if in the case where such a process is exceptional, we would not discount a recording simply because it had an abnormal cause.

There is reason to think this phenomenon of local adaptation is quite general. Think about what has happened to the concept of motherhood in the face of recent technological advances. As it has become possible to implant the egg of one woman into the uterus of another, a previously unnoticed distinction has been drawn between *genetic mother* on the one hand, and *birth mother* on the other[202]. Since the cases are

[202] Consider the following parody of a new-age greeting card, taken from *The New York Times* Sunday Magazine:

> YOU'RE SOMEONE SPECIAL
> Mom...
> Gee, it feels funny to call you that.
> But after all, you are the woman
>   who brought me to term.
> And even though it was "just a job,"
>   I feel as though we have a lasting bond.

Continues on next page…

exceptional, there has been an effort to “save” the concept; and since it is the egg-donor whose genetic information is carried on to the child, it is the birth mother who has been given the status of “surrogate.” But were the practice to become widespread, the concept of motherhood would break down entirely. We would no longer have the idea of filial concern for one’s mother, because there would be no unitary concept of “mother” that lay behind it. We might well have two similar concepts—filial concern for one’s birth mother and filial concern for one’s genetic mother; and we might well think these concepts were more similar to one another than either one would be to the concept of concern for one’s child, or one’s spouse, or one’s sibling. But I think that in such circumstances, the concept of maternal-filial concern *simpliciter* would have no application.

And this is the final piece to the puzzle. In a world in which fission was the norm, there would not be a concept of prudential concern; there might be a somewhat similar concept, such as the concept of concern-for-one’s-R-related-continuer. But there wouldn’t be the same concept as the one we have, the concept that describes the relation we bear to our future selves[203]. In such a world, Brainy’s relation to Lefty in

> I know it can’t have been easy
>   carrying around someone else’s baby,
>   especially a big eater like me!
> So I just want to say,
>   thanks for being my birth mother!
> The time we spent together
>   will always mean something special to me.

(Rubiner 1996, 60.)

[203] But why is this not like the story of the Oxford misogynist, who, at a debate over whether All Souls College should admit women, contended that All Souls College

Continues on next page…

both the one-sided and two-sided cases would indeed contain what matters for concern-for-one's-R-related-continuer. But the relations would not contain what matters for prudential concern, because there wouldn't *be* prudential concern[204].

Now, if we say, as I have been arguing, that *identity* is what matters in the explanatory sense, then we have some way to account for the fact that our concepts would change in the face of such a global disruption[205]. But the same option is not

could not admit women, for if it were to do so, it would cease to be All Souls College? (I thank Derek Parfit for the story, and for the objection it raises.)

The answer, I think, is that the role played by All Souls College before and after the admission of women as fellows would be largely unchanged, whereas the role played by R-continuer concern in such a world would be *too different* from the role played by prudential concern in our world for the two to be considered the same concept. To argue fully for this would require another chapter (at least), but the line of thought is as follows. The concept of prudential concern is tied up with concepts of fairness, responsibility, justice, and rationality. Our views about the sorts of rational and moral obligations we have to ourselves and others considered as beings who exist through time rest on the assumption that each of us will have at most one continuer, and that that continuer is someone with whom we will be identical. Disruption of this background assumption would result in disruption of the entire framework by which we make sense of this wide range of concepts. And to the extent that prudential concern is interconnected with them, it too would be disrupted.

For discussion of some of these issues, see Korsgaard 1989; MacIntyre; Rovane 1990; Whiting 1986; and Wolf 1986. (I thank Dick Boyd for helpful conversation on these topics.)

[204] Something like this seems to be what McDowell is suggesting when he writes: "According to the view I recommend, a context of facts about the objective continuation of lives helps to make intelligible a face-value construal of what Locke actually says, that continuous 'consciousness' presents an identity through time" (McDowell (forthcoming), 6).

[205] Cf. Shoemaker: "[W]hile I think that there are possible cases in which identity and the proper object of special concern come apart, e.g., cases of 'fission,' I think that there is nevertheless a close conceptual link between these. I think it is a constraint on the concept of a person that the truth conditions for judgments of personal identity should, so far as possible, make it true that persons are identical with the future

Continues on next page…

available if we say that what matters is R-relatedness. So if follow Parfit in accepting that what matters for prudential concern in the explanatory sense is not identity, then we have no way to account for the fact that the concept would not exist under the conditions I described. If we say instead that what matters for prudential concern *is* identity, we rightfully acknowledge that the concept depends crucially on the way things happen to be[206].

persons for whom they rationally have a special sort of concern" (Shoemaker 1995, 36).

[206] But doesn't the following argument suggest that I have only succeeded in showing something rather weaker, namely that what matters is R-relatedness plus uniqueness, and not identity as such? Imagine a world in which everyone undergoes fission at the age of 18, but where one of the resulting survivors always dies within a few days. It seems plausible to say of such a world: (a) that there would be prudential concern, and (b) that as a rule, this prudential concern would not have to be for someone with whom one was identical. This seems to show that it is uniqueness rather than identity that matters. (I thank Sydney Shoemaker for this objection.)

I reply as follows. Case one: If the example supposes that the quick death of one of the fission products is guaranteed by the biological laws that govern the imagined world, I think the pre-fission persons are identical to their post-fission survivors. Their way of surviving over time involves an odd lapse in what we would consider normal, but I think this is no more disruptive than imagining a world in which at the age of 18, everyone spent four days in a state of deep unconsciousness from which they completely recovered. So if the world imagined is such that the process Shoemaker imagines is nomologically grounded , then I think there is room to deny (b). Case two: If, on the other hand, it is mere coincidence that in all instances thus far the death of one of the fission products has been an early one, then I would reply in one of two ways. First, it is not clear to me that we really know how to make sense of such massive coincidence; my suspicion is that this case would either collapse into the first case (where what is going on really is law-governed), or that it would be a case where we are imagining a world whose relation of coincidence to law is sufficiently foreign as to be incomprehensible to us. But to the extent that the second case is comprehensible, I would be inclined to deny (a); if in such a world there is the always-present but never-realized live possibility of multiple-continuation, then I think the concept of prudential concern would begin to break

Continues on next page…

4.6.3 Summary

I began by suggesting that there is a danger to philosophical inquiry that ignores what I have been calling the facts of life. That human beings come into existence only through the predictable sequence of events that I described at the beginning of the chapter is one of the background truths against which we organize our concepts. At the same time, it seems possible that there could be circumstances—fission is one—where a process that is ordinarily identity-preserving would turn out to be entity-creating. That is, it seems possible that there could be a process with the following character: if it happened in one way (what we have been calling the "one-sided case") it would result in the continued existence of some entity over time; but if it happened in another way (what we have been calling the "two-sided case") it would result in the creation of two new entities. But if the entities in question are self-conscious, as human beings are, then this possibility raises the following puzzle. To the extent that the process itself would—intrinsically—be same in both cases, how could the rationality of one's attitude towards one's continuer depend on whether the process ended up being identity-preserving, or whether it ended up producing two new human beings? Presumably one's attitude towards one's continuer would—rationally—be the same in both the one-sided and the two-sided case. And with this much, I said I agree.

down. In short, the more compelled I am to accept (a), the less compelled I am to accept (b), and *vice versa*.

The question that has concerned me in this chapter has been the question of what lessons can be drawn from this fact. Parfit contends that from it, we can conclude that what makes my prudential concern for myself tomorrow rational is not the fact that Tamar-tomorrow will be identical to Tamar-today, but only that she will be connected to me by the right sort of causal process that will result in the right sort of relation of psychological continuity and connectedness. I have tried to show that this conclusion can be blocked. I pointed out that Parfit's reasoning rests on what I have called the crucial premise: that whether a relation contains what prudentially matters depends only on the relation's intrinsic features. And I have tried to show that the crucial premise can be denied—not on the grounds that some local fact might affect the rationality of one's attitude towards one's continuer, but on the grounds that some global fact—such as a general uncoupling of continuation from identity—might. What this means is that Parfit's argument shows much less than he takes it to show. It shows only that there are conceivable circumstances where it might be rational to bear a relation of prudential concern towards a continuer with whom one was not identical. But it does not show that identity is not what—in the explanatory sense—matters.

Although most of my argument has focused on one example, I should make it clear that I take my discussion to have general implications. In the case I described, we are asked to consider a scenario in which a pair of features that coincide in all actual situations are imaginatively separated, and to make a judgment about which of the two features has conceptual primacy. I have argued that the proper interpretation of the case is precisely the opposite of what it has generally been taken to be. And I think the reason its implications have been so misunderstood is this: certain patterns of features which coincide only fortuitously may nonetheless play a central role in the organization of our concepts. To the extent that imaginary scenarios involve disruptions of these patterns, our first-order judgments about them are often distorted or even inverted. It is this that I have tried to demonstrate during the pages above.

5. Conclusion

At the beginning of this dissertation, I raised the question of how thinking about imaginary cases can help us to learn new things about the world. I went on to suggest that the answer has something to do with the role played by exceptional cases. I then identified a structure which I suggested is common to thought experiments in science and philosophy. And I closed the introductory chapter by submitting that the question has at least three different facets, one concerning factive, one concerning conceptual, and one concerning valuational thought experiments.

In the next three chapters, I presented detailed case studies of three thought experiments, each representing one of the three sorts of case. And in each of the case studies, I tried to establish a number of specific claims. In the first case study, I discussed a famous thought experiment of Galileo's, and tried to show that the guided contemplation of an imaginary scenario can provide us with new scientific knowledge in a way that argument alone cannot. In the second case study, I tried to show that standard interpretations of the puzzle of the Ship of Theseus founder because they ignore the importance of the background norms against which we can make sense of local instances of extrinsically-determined identity. And in the third case study, I tried to show that certain thought experiments in the personal identity literature are inconclusive because they disregard the explanatory role played by contingent facts about the ways human beings come into existence.

The aim of this conclusion is to tie these strands together. I will do so by discussing each of the thought experiments in turn. What I try to show is that exceptionality plays a role in their evaluation at three distinct levels: (a) in the patterns of classification of certain states of affairs as normal or exceptional; (b) in the ways in which the particular exceptional (that is, unusual) situation described by the scenario should be accounted for; and (c) in the ways in which this exceptionality is accounted for at the level of explanation. It is the interplay among these three aspects of exceptionality that explains the respective successes and failures of the particular cases I have discussed.

5.1 Factive Thought Experiments: Galileo

Scientific thought experiments are typically factive[207]; they are attempts to elicit physical intuitions about what would happen under certain conditions. Such thought experiments are puzzling because they seem to describe cases where we learn something new about the physical world, even though we have no new *empirical* information about the world[208].

Three sorts of responses have been offered to this puzzle. In chapter 2, I discussed the *empiricist* response, as exemplified by the writings of John Norton.

[207] The reader will recall that *factive* thought experiments are cases where an imaginary scenario is described, and the reader is asked to think about *what would happen* in such a case.

[208] Thomas Kuhn poses the problem concisely as follows: "How…relying exclusively upon familiar data, can a thought experiment lead to new knowledge or to a new

Continues on next page…

Norton contends that scientific thought experiments have demonstrative force only insofar as they are arguments in disguise. Otherwise, he contends, we would not be able to account for the fact that they provide us with new information that is not new *empirical* information. In chapter 2, I argued that such a view cannot account for Galileo's famous thought experiment, where as a consequence of thinking about a particular imaginary case, the Aristotelian is brought to reject certain framework-defining assumptions about the sort of property natural speed is. Prior to thinking about the case that Galileo describes, the Aristotelian is committed to the view that natural speed is a function of weight. After thinking about the case in the way that Galileo encourages, he recognizes that such a conception is untenable. He comes to understand that other things he knows about the ways that middle-sized physical objects behave belie his previous representation of natural speed.

This realization is not a sudden flash of Platonic inspiration, which illuminates previously invisible *a priori* physical truths. Such a view can be found in the writings of James Robert Brown, against whom Norton directs his argument. Brown contends that in some cases, "thought experiments give us (fallible) *a priori* beliefs of how the physical world works. With the mind's eye, we can see the laws of nature" (Brown 1991, 155)[209]. But where is no need to posit such mysterious entities, they ought to be

understanding of nature?" (Kuhn 1977, 241). (See also the quotes from Norton in section **2.1.3** above.)

[209] Brown considers the thought experiment of Galileo's which we have been discussing to be one of the two "best examples of….a *platonic thought experiment*"* (Brown 1991, 77). For his discussion of the case, see Brown 1991, 1-3 and 77-79.

*The other is the 1935 Einstein, Podolsky and Rosen thought experiment (generally referred to as "EPR").

Continues on next page…

avoided. And, as I have argued above and will argue below, there is a satisfactory explanation of the phenomenon which does not make appeal to such enigmatic abstracta.

Between immoderate empiricism and immoderate platonism lies a third sort of explanation. Eschewing the assumption that there is sharp line between theory on the one hand, and raw data on the other, the position introduces an element of *constructivism*. It suggests that the new knowledge in the Galileo case comes neither from argument nor from insight, but from a reconfiguration of conceptual commitments on the part of the Aristotelian which enables him to see old phenomena in a new way[210].

The thought experiment directs the Aristotelian's attention to inadequacies in his conceptual scheme, inadequacies which he recognizes to be inadequacies when he performs a particular experiment in thought. The experiment in thought involves asking himself what sorts of objects and motions there are in the world. Through this,

[210] The three proposals I have been considering are *explanatory* rather than *genetic*. Each would be compatible with the proposal made by Roger Shepard in his William James lectures. Shepard suggests that "every thought experiment is preceded by an enormous number of actual experiments," namely, the selectively-relevant experiences had by our ancestors in the long course of evolution of the species. (See Shepard 1994). So, he suggests, the knowledge gained as a result of thought experiment is knowledge of naïve physics implanted in us by evolution.

I am not sure Shepard's answer gives the whole story. But it is surely true that some knowledge of naïve physics is hard-wired. Both folk psychology and cognitive science tell us so; and there are additional arguments (some due to Shepard himself) which seem to refute convincingly the proposal that each of us is born a *tabula rasa*. But as with the discussion of Mach above (see section **2.3.3**), I have little more to add here; this question lies outside the scope of the dissertation.

he realizes that entification is not fixed by the world, whereas natural speed is. Such an explanation appeals to exceptionality at three levels, which I now discuss.

As I noted in the text, the Aristotelian needs to explain as aberrant what the Galilean takes as typical: that the rate of fall of bodies of radically different weights is sometimes (nearly) identical (see section **2.3.1**). And as I went on to note, he does so by appealing to additional principles which show how the world's complexity has prevented the true principles from being manifest; he suggests, for instance, that from great enough heights, the relevant differences would be apparent. Part of the power of Galileo's thought experiment comes in helping the Aristotelian to see that the cases he has been taking as anomalous are in fact the norm, that what cries out for explanation is precisely the opposite of what he initially thought; the cases which require explanation are those where heavier objects fall faster than lighter ones. So the first way in which exceptionality comes into the explanation is that the Galilean thought experiment brings the Aristotelian to see that the cases he has been taking to be exceptional are in fact typical, and *vice versa.* We'll call this *exceptionality in the first sense*.

Secondly, the specific case which Galileo asks the Aristotelian to contemplate is exceptional in the following sense. In ordinary cases involving falling bodies, the objects which fall are objects for which criteria of individuation are clear. We ask: what would happen if we were to drop thus-and-such body from this-and-that height? Or: what would happen if we took these two things up to the top of that tower and let them go? But the case which Galileo asks the Aristotelian to think about is one where the criteria of entification are not so obvious. Is that strapped-together thing one body or two? The world doesn't really tell us.

But despite the fact that the object in the example is odd in this way, the conclusions which can be drawn from thinking about what would happen to it are general conclusions. Descriptive physical science aims at generality, and if the theory must be modified to account for a case which is (in this second sense) exceptional, then the theory must be modified in general. That it is here appropriate to allow exceptions to drive norms is confirmed by the second test (see section **1.1.3**). Even if all objects were like this strapped-together object, it would still make sense to have a theory about the natural speed of falling bodies[211]. So the second way in which exceptionality comes into play is in bringing out why the evaluation of the imaginary scenario is rightly taken to reveal general truths about cases beyond the scenario (see section 1.3.2).

The third and most interesting way that exceptionality comes in is at the level of the explanation itself. In chapter 4, I argued that absent features play explanatory roles by providing the background assumptions against which explanations make sense; that is, when we are concerned with explanations, we are concerned with norm-driven exceptions. What this means in the Galileo case is that what I above called *exceptionality in the first sense* (that the inversion of exceptional and normal in the order of explanation) is sufficient to bring with it new knowledge. Let me explain.

[211] Indeed, in certain presentations of modern physics, all bodies *are* conceived of as strapped-together in this sense—that is, as particles held together by adhesive forces. (I thank Simon Saunders for pointing this out to me.)

Continues on next page…

Galileo’s thought experiment brings the Aristotelian to see that the cases he had been taking as remarkable are in fact ordinary, that what demands explanation is precisely the inverse of what he previously supposed. In one sense, this transposition is trivial; all that has happened is that two classes of cases have been differently diagnosed. But if we take seriously what we have learned about exceptionality in the third sense, then it becomes apparent that this inversion has far-reaching consequences.

As I argued in chapter 4, exceptions in explanation are norm-driven; we can make sense of anomalous cases only against a background of regularity, and we recognize these as cases of the relevant sort only, so to speak, out of courtesy. The Galilean thought experiment helps the Aristotelian to see that explanation is needed for cases where two bodies of differing weights fall with *different* natural speeds (and not cases where two bodies of differing weights fall with the same natural speed). And what this means is that it helps him to see that the cases where bodies fall with different natural speeds need to be understood *through the lens* of cases where bodies fall with the same natural speed. They need to be seen *as exceptional*; more specifically, they need to be seen as exceptional with respect to a norm that says: as a rule, bodies fall at the same speed regardless of their weight.

And this is the deepest sense in which the contemplation of an imaginary scenario forces the contemplator to make sense of an exceptional case, and thereby brings him to new knowledge. Because of the relation of exceptions and norms in the case of explanation itself (third sense), because of the relation of exceptions to norms in the case of physical theories (second sense), and because of the inversion of exceptions and norms in what demands explanation in this particular case (first sense), it can truly be said of this thought experiment that it brings the Aristotelian to see new things in the world; it does so by bringing him to see the world in a new way, against a newly configured background of norm and exception.

5.2 Conceptual Thought Experiments: The Ship of Theseus

Conceptual thought experiments are cases where an imaginary scenario is presented, and the reader is asked to think about *how such a situation should be described*. In chapter 3, I discussed a particular conceptual thought experiment, that of the Ship of Theseus; I argued that the implications of that thought experiment have been standardly misunderstood.

We might express the problem using the vocabulary of section **1.3.2**. Clearly, the scenario described is imaginable; it is an example of what I called in section **1.2** a *hypothetical* case, which means that for all we know, such a case might actually have occurred. So there is no reason to level an *unimaginability* criticism against the case. Moreover, standard analyses of the scenario, to the extent that they recognize that the case involves pitting two principles of identity against one another, are largely reasonable. For the most part, there is no reason to level *unsound argument* criticisms against interpretations of the case[212]. Instead, the problem with discussions of the Ship of Theseus is in the *application* of the lessons of this particular case to cases of identity in general. The appropriate level at which to object to the solutions presented in sections **3.5** and **3.6** above is on *inapplicability* grounds.

Returning to the vocabulary of this conclusion, we might say that the Ship of Theseus is an exceptional case in the second, flatfooted sense; that is to say, it is an unusual case. What should we make of this? In the Galileo example, the unusual case of the falling strapped-together body served to establish that the speed at which such bodies fall is independent of their weight. And the proper conclusion to draw on the basis of the case was that the previous classification of normal and exceptional cases was inverted. Heavy and light bodies falling at the same speed are normal; heavy and light bodies falling at different speeds are anomalous. Because the proper way to account for the exceptional case is by changing the norms to meet the exception, the proper way to understand the patterns of exceptionality is to reverse the previous assumptions.

[212] Dissolutions like those offered by Parfit and van Inwagen suggest that apparent depth of the puzzle is a result of false assumptions about the relation of our concepts to the world. Such positions err not in their reasoning about the particular case of the Ship of Theseus, but in the conclusions they draw about such cases in general. Solutions like those offered by Hirsch and Nozick give what I take to be the right answer to the question; where they err is in the generalizations that they make on its basis.

We might classify the four views as follows:

	Solution	**Dissolution**
Closest continuer	Nozick	Parfit
Not closest continuer	Hirsch (traditional)	van Inwagen

In the story of the Ship of Theseus, the unusual case presents us with an instance where a process which is ordinarily entity-preserving is instead entity-creating. But here it is not true that the proper conclusion to draw is that such processes are in general entity-creating (or that they provide dubious mechanisms for identity-preservation). And the *reason* that this is the wrong conclusion to draw is again brought out by the generalization test (see section **1.1.3**). For if cases like Theseus were the norm rather than the exception, it would not make sense even to speak of identity-candidacy for ships. Ships would be like amoebae or cloud formations or World Wide Web sites—messy sorts of entities with obscure criteria of individuation and persistence; in general, we would not know what to say about their identity conditions over time. So the case of the Ship of Theseus is one where the proper interpretation of the case is as a norm-driven exception. We can make sense of this anomaly, but only against a background of normal cases.

What, then, of the third level at which exceptionality enters the picture, the level of explanation? In the Galileo case, it was this level that solved our epistemological puzzle. It helped us to see why the perspective shift afforded by the accommodation of this exceptional case was sufficiently powerful to make pre-shift perceptions so radically different from post-shift perceptions that we might without exaggeration speak of *new* knowledge. In the case of the Ship of Theseus, where no such shift occurs, the consequences are more modest. I suggested that what the case reveals is that processes may be extrinsically determined in ways that we might not have realized. Explanation goes via normal cases; they provide the background against which we make judgments about the atypical. But the sort of radical shift in perspective that ensued as a result of the scientific thought experiment is not possible here. The fulcrum around which we seek to reorganize our conceptual commitments is not sufficiently independent of the commitments themselves to permit such a radical reconfiguration.

Thinking about the case in this way sheds new light on the errors of those who misgeneralize on its basis. The sorts of radical shifts in perspective which they propose are precisely the sorts of alterations we would expect if it were indeed reasonable to treat this as a case of exception-driven norms. So once again our schema has helped us to understand the import of an imaginary exceptional case.

5.3 Valuational Thought Experiments: Personal Identity

Valuational thought experiments are thought experiments where were are asked to make judgments about how we would *evaluate* what would happen in a particular imaginary scenario. In chapter 4, I discussed a particular valuational thought experiment, the case of fission, and I argued that the case does not succeed in showing what Parfit takes it to show, namely, that personal identity is not what matters in justifying rational prudential concern.

In contrast to the other two examples, the case of fission is what I called in section **1.2** a *counterfactual* case; even without looking at the world, we know the case to be imaginary[213]. This means that the case must be approached with a certain

[213] Baruch Brody considers this sufficient to make the case irrelevant. He writes “Parfit [and others]…see the very possibility of such a case as posing a conceptual crisis about identity. I say that only the actual occurrence of such cases would pose a

Continues on next page…

degree of caution. Not only must we be careful not to misunderstand the implications of the exceptional case we are asked to consider, but we must be careful *in the interpretation of the case itself* not to import assumptions which would be inapplicable in that imaginary context. This suggests that the proper sort of criticism in the case of fission is at the level which in section **1.3.2** I referred to as *unsound argument*. In making the judgment about the case which he does, Parfit offers an *evaluation* of the scenario which is fundamentally misguided. His conclusion rests on certain assumptions about how we would value certain states of affairs, assumptions which are illegitimate in light of the very scenario which they are meant to illuminate.

Let us examine this in light of the categories of the conclusion. Clearly, the case of fission is exceptional in the flat-footed second sense; division of the sort Parfit describes is meant to be extraordinary, not commonplace. And, as I argued extensively in the body of chapter 4, the appropriate way to account for this case is to recognize that for it to make sense to us, it *must* be exceptional. Consider again the sound-alike recording. A recording that sounds just like a great cellist playing Bach is as valuable (as far as the sensory experience it provides) as an actual recording of a

crisis, and that it would be a scientific crisis about how to describe and explain what has occurred…[S]ince such cases do not occur, and it is only their occurrence that would cause a problem on my account, it follows that there is, at the moment, no splitting problem if my account is true" (Brody 1980, 59). For the reasons I discussed in the introduction, I do not believe we can be rid of imaginary cases so easily as Brody seems to think. But Brody is surely right to point out that there are *practical* questions raised by actual cases which are not raised by cases which are purely imaginary, and that changes in the ways that the world is *do* pose crises for concepts that we otherwise use unproblematically. (I touch briefly on these issues in section **1.2** (see especially footnote 23) and section **4.6.2** (see especially footnote 160).)

great cellist playing Bach. But the reason this is so is because, in all actual cases, such recordings *are* recordings of great cellists playing Bach. The value that attaches to the sound-alike is derivative; such disks shine like the moon, not like the sun. If the patterns whereby such recordings come into existence were disrupted, sound-alikes would lose their value[214]. Our ability to judge their worth is parasitic on our ability to judge the worth of actual recordings.

What this means it that when we perform experiments in thought concerning the scenarios described in valuational thought experiments which involve counterfactual cases, our very judgments are filtered through the background patterns of fortuitous coincidence which play central roles in the organization of our conceptual scheme. Imaginary cases allow us to separate out apparently contingent concomitants. But some of our conceptual commitments may presuppose the coincidence of such concomitants, even though we can make sense of their coming apart.

When we make judgments about such cases, we do so standing on the shoulders of the normal cases from which they deviate. But we cannot, as it were, throw away this ladder. We can see as far as we do standing where we do, but we should not mistake this for standing on the ground. Questions about how we would extend our practices are not ways of finding out what underpins our practices.

In the situations Parfit describes, we can make sense of the cases he adduces. But we cannot make sense of them *as* cases about persons. They tell us what sorts of

[214] It might be that we would take them to have as much value as ordinary recordings

Continues on next page…

beliefs and commitments we might expect to find among a group of individuals for whom different possibilities were live, but they do not tell us about our *own* beliefs and commitments. What seems to make these cases so puzzling is that the sorts of circumstances we imagine underpinning these live possibilities are circumstances which strike us as being at most technologically impossible[215]. But, as I have argued throughout, merely contingent constraints on the way things happen to be may play central roles in our categories of evaluation.

This brings us full circle to the Galileo case. There, we saw that the thought experiment is able to bring new knowledge to the Aristotelian by allowing him to see all cases involving falling bodies *through the lens* of cases where bodies fall with the same natural speed. This is what explains the thought experiment's power in the Galileo case. And in the fission case, precisely the same feature explains the thought experiment's limit. We cannot suddenly step outside the framework within which we

now do. But if so, it would be for different reasons.

[215] Consider the following analogy: Suppose there were a community in which it were possible to arrange for a total-matter-replacer to keep one's book collection from decaying. Each day, a quick vacuuming of one's bookshelves would result in all of the matter in the books being replaced by new matter, configured in an identical way, such that all of the observable features of the books—the color of the pages, the words on them, dents on the covers, penciled notes in the margins, post-it notes in the bibliographies—would remain the same. People faced with this option might well be rational to accept it, and it would seem at best unkind to say of them that they represented a culture that did not care about books. On the other hand, this does not show, I think, that all *we* care about when we care about some book is the qualitative features that the book manifests. Nor do I think it shows that all we care about when we care about some book is that it have the qualitative features it has brought about in the ordinary way because that's the only way we can get those features.

ordinarily make sense of what we value and believe. The way the world is and the way we see the world are too deeply intertwined for this to be possible.

Appendix A: Three Examples of Conceptual Change

In this appendix, I try to give a clearer sense to what I mean when I speak of coming to make novel conceptual distinctions. In the three subsections which follow, I present a trio of examples of conceptual change that involve differentiation[216]. In each of the examples, some collection of phenomena previously described or understood by means of a single expression or concept comes to be described or understood by means of a pair of expressions or concepts. In the first example, the newly differentiated concepts are *instantaneous velocity* and *average velocity*; in the second they are *heat* and *temperature*; and in the third they are *promoting democracy abroad* and *protecting the interests of the United States*.

A.1 Instantaneous and Average Velocity

In "A Function for Thought Experiments" (Kuhn 1964), Kuhn discusses a series of experiments conducted by Piaget concerning children's conceptions of motion and speed (see Piaget 1964/1970). In the experiments, the children were

[216] Conceptual change may occur by means of other mechanisms as well. For instance, Wiser and Carey identify at least two others: *coalescence* and *shift from property to relation.* (Examples of coalescence would include the way in which Newtonian gravitational mass and inertial mass are collapsed in Einsteinian mechanics, or the way in which Aristotelian natural and violent motion are collapsed in Galilean mechanics. Examples of shifts would include the way in which in pre-Newtonian theories, weight is treated as a property, whereas in post-Newtonian theories it is treated as a relation. See Wiser and Carey 1983, 268-269.) In all three sorts of cases (differentiation, coalescence, and shift from property to relation), there is a recognition that certain patterns of apparent saliency were just that: features that

Continues on next page…

instructed to observe a pair of toy cars which were moved uniformly along a straight line towards some pre-specified goal. One of the cars was blue, the other red, and at the completion of each trial, the children were asked which of the cars had moved faster.

The children confronted three sorts of trials. In the first sort, both cars moved the same distance but one took longer to do so. So, for instance, the blue car and the red car might leave the starting line together, but the red car would move at a steady pace of *p* along the entire course, while the blue car moved at a steady pace of *p+1*. In the second sort of trial, both cars moved for the same amount of time but one covered a greater distance. So, for instance, the blue car and the red car might leave the starting line together, the blue car moving at a pace of *p+1* and the red car at a pace of *p*, and both would stop after some specified period of time. In the third sort of trial, neither the times nor the distances were the same. So, for instance, the blue and red cars might set off together, with speeds as before, but the blue car might stop after a certain amount of time, while the red car continued on a bit longer. After each trial, Piaget asked the child which car had gone faster, and why[217].

The children, aged about five, tended to describe as faster the car which reached its goal first; this was so even when they recognized that the "slower" car had

initially seemed irrelevant turn out to have been relevant and *vice versa*; predicates that seemed to be one-place turn out to have been two-place and *vice versa*.

[217] The experiments are described in Kuhn 1977, 243-256; I make use throughout my discussion of Kuhn's description of the cases and (with the exception of the example used to illustrate the *goal-reaching* criterion), I make use of his translation. For the complete text, see Piaget 1964/1970; the examples in question can be found in chapters 7 and 9.

traveled a greater distance during the same period of time. Kuhn suggests that the child's response here manifests what he calls the *goal-reaching* criterion for the application of *faster* (Kuhn 1977, 244). That is, the child uses the term "faster" to describe the object which reaches its goal first, regardless of differences in the length of path covered, or the relative rates of motions of the vehicles. So, for instance, in the following example, the red car and the blue car both travel for the same amount of time, but the red car starts from a point well behind the blue car, and catches up with it.

> P(iaget): Same speed or not?
> C(hild): Both the same.
> P: Do they do the same distance?
> C: No, one starts here and the other there.
> P: The same length?
> C: No, that blue one is shorter.
> P: Were they both going just as hard?
> C: Yes.
> P: One of them wasn't faster than the other?
> C: No.
> P: Which one went a long way?
> C: Red.
> P: And what if I said to you that one was going harder than the other?
> C: Well, I think they went at the same speed. (Piaget 1964/1970. 158).

The child seems to be assuming that because the cars reached their goal simultaneously, both traveled at the same speed, even though the red car had a much greater distance to traverse[218].

[218] Or again, in an experiment involving two dolls moving along concentric circular tracks, one of which has a radius approximately half the length of the other:

> C: They both go at the same speed.
> P: Why?
> C: Because they finished at the same time.

Continues on next page…

But in addition to the “goal-reaching” criterion, children of this age also apply a second criterion to determine which of two objects is moving more quickly. Kuhn refers to this criterion as “perceptual blurriness:” One object is “perceptually blurrier” than another if its more rapid motion can be directly observed. So, for instance, when the case just described was modified so that the red car began moving *much* later than the blue car, and correspondingly moved much more rapidly, the same child responded as follows:

> P: Did one go more quickly than the other?
> C: The red.
> P: How did you find that out?
> C: I WATCHED IT[219].

These two criteria—goal-reaching and perceptual blurriness—coexist unproblematically in most cases. In most of the trials presented by Piaget, and in most

> P: But which one has a longer road to go?
> C: The one on the little circle.
> P: Then which one goes quicker?
> C: They’re the same.
> P: Why?
> C: Because they started off at the same time?
> P: But did both hurry the same or was one hurrying more?
> C: Both the same (Piaget 1964/1970, 142).

[219] Or again, in a race where the blue car moves at a speed four times that of the red:

> P: Blue was faster?
> C: Yes
> P: Why?
> C: Because I can see it.
> P: If you shut your eyes during its journey and you saw only the ends of the journeys, could you tell?
> C: No
> [We do so]
> P: Why was the blue faster than the red?
> C: I just can’t see why. (Piaget 1964/1970, 160).

Continues on next page…

daily experience, faster-moving objects reach their goals earlier than slower-moving objects[220]. However, as Kuhn notes:

> Not very often (or the children could not have preserved the concept for so long) but occasionally nature will present a situation in which a body whose directly preserved speed is lower nonetheless reaches the goal first. In this case the two clues conflict; the child may be led to say that both bodies are "faster" or both "slower" or that the same body is both "faster " and "slower." That experience of paradox is the one generated by Piaget in the laboratory (Kuhn 1977, 244-245).

That is, in most cases, the object which reaches its goal first is the object which moves most quickly along the designated path. Occasionally, however, the faster-moving object will begin its journey much later than the slower-moving object, or will trace a much longer path. Under such circumstances, the child's undifferentiated concept is inadequate to make sense of these phenomena[221], phenomena for which it is expected to account. "As a result," reports Kuhn, some of the children "changed their concept of 'faster,' perhaps by bifurcating it. The original concept was split into

[220] This suggests that "The Tortoise and the Hare" teaches cognitive as well as moral lessons.

[221] Cf. also this example of a child's failure to differentiate two concepts which adults take to be clearly distinct is reported in this conversation between Susan Carey (S) and her daughter Eliza (E), aged three-and-a-half:

> E: Isn't it funny—statues aren't alive but you can still see them?
> S: What's funny about that?
> E: Grandpa's dead and you can't see him.
> S: Oh, I see. Well, you know, people and animals can be alive and dead—first they are alive and then when they die, they're dead. But other things, like chairs—they aren't ever alive, so they can't die.
> [E]: That's right. Tables and chairs are not alive and they're not dead and you can still see them. Isn't that funny, they're not alive, but you can still see them.

(Carey 1988, 178)

Continues on next page…

something like the adult’s notion of ‘faster’ and a separate concept of ‘reaching-goal-first’” (Kuhn 1977, 245).

Kuhn uses Piaget’s studies as a way of illuminating “a historical, but otherwise similar, case of concept revision, this one...promoted by the close analysis of an imagined situation” (Kuhn 1977, 246). The historical case involves Aristotelian concepts of speed, which like those of the children, give paradoxical answers in certain cases. On one strand of the Aristotelian picture, “two things are of the same velocity if they occupy an equal time in accomplishing a certain equal amount of motion” (*Physics* 249a7-8) or again, “there is equal velocity where *the same* change is accomplished in an equal time” (*Physics* 249b4-5). This notion might be called “average velocity;” it is a quantity equal to the ratio between the difference covered and the time spent covering that distance. Kuhn suggests that this notion corresponds to the child’s notion of goal-reaching; what it measures is something about the *overall* motion of the object under examination.

In addition to this notion of speed, however, the Aristotelian framework seems to apply the same concept to a second sort of quantity, one which corresponds to what might be called “instantaneous velocity.” So, for instance, at 230b23-25, Aristotle writes: “whereas the velocity of that which comes to a standstill seems always to increase, the velocity of that which is carried violently seems always to decrease” (*Physics* 230b23-25). Kuhn suggests that “Aristotle [here] is grasping directly...an aspect of motion which we should call 'instantaneous velocity' and which has properties quite different from average velocity" (Kuhn 1977, 247).

By the mid-fourteenth century, these notions had been partially disambiguated; fourteenth century natural philosophers distinguished between *total velocity* and *intensity of velocity*. The latter was, in many ways, close to the modern notion of instantaneous velocity; the former, with important modifications from Galileo, might be profitably compared to the modern notion of average velocity[222]. But the medieval distinction applied only to motions which took place over the same distance or which required the same amount of time; it could not be used to compare motions where both distance and time differed[223]. The distinction used in modern science was made precise only by Galileo, in a discussion presented in his *Dialogue concerning the Two Chief World Systems.*

In the passage in question, Galileo's spokesman Salviati[224] begins by asking his listener to imagine a pair of planes: a vertical plane, CB, and an inclined plane, CA. Although CA is longer than CB, the two cover the same vertical distance. Salviati illustrates the situation as follows:

[222] For discussion of these matters, see Claggett 1959, Part Two.

[223] See Kuhn 1977, 248 for a concise discussion of these issues; for more detailed discussions, see Claggett 1959; Clavelin 1974; Dijksterhuis 1961.

[224] The characters in this dialogue are the same as those described in chapter 2.

He then asks his listener to imagine two bodies sliding from C without friction, one towards A and the other towards B, and expects him to concede that: “the impetus of that which descends by the plane CA, upon arriving at point A, will be equal to the impetus acquired by the other at point B after falling along the perpendicular CB.” That is, at the conclusion of the fall, the velocity of the two bodies will be equal. His companion agrees, conceding that “two equal movable bodies, descending by different lines and without any impediment, will have acquired equal impetus whenever their approaches to the center are equal” (All quotes from Galileo 1632/1967, 23).

At the same time, however, the interlocutor admits that it seems “necessary that the motion by the perpendicular CB should be faster than by the inclined plane CA[225]” (Galileo 1632/1967, 23). After all, the ball which moves only from C to B will reach its goal before the ball which moves from C to A, and that same ball will seem, to revert to the terminology used to describe Piaget’s children, “perceptually blurrier.” But then the ball that moves along CB must move both faster than, and at the same velocity as, the ball that moves along CA. How could this be so?

What Galileo suggests is that two sorts of velocity have not been distinguished. Initially, he puts the point paradoxically: “I...say categorically that the speeds of the bodies falling by the perpendicular and by the incline are equal....[T]his proposition is quite true, just as it is also true that the body moves more swiftly along the perpendicular than along the incline” (Galileo 1632/1967, 24). On one notion of

[225] Sagredo does not actually assent to this, but he does speak of its intuitive appeal.

speed, what we would call average velocity, “velocities [are] equal when the spaces passed over are in the same proportion as the times in which they are passed.” On this notion, “the body along the perpendicular moves more swiftly than that along the incline” (Galileo 1632/1967, 24). On the other notion of speed, what we would call instantaneous velocity, “the proposition ‘motion along the perpendicular is faster than along the incline’ is true not universally, but only for motions which begin from the initial point, that is, the point of rest.” That is, “motion along the incline is in some places faster and in some slower than motion along the perpendicular” (Galileo 1632/1967, 25).

So what Galileo suggests is that two concepts have been conflated. What the Aristotelian has failed to recognize is that the object moving along CB moves *faster* than the object moving along CA in the sense that it reaches its goal first, and *at the same speed* as the object moving along CA in the sense that their terminal velocities are identical. In a parallel fashion, Piaget’s children have failed to recognize that one car might be *faster* than another in the sense that it reaches its goal first, while being *slower* in the sense that it is perceptually “less blurry.”

The novel conceptual distinctions afforded by such disambiguations do two things: they capture what is intuitively appealing in the previous undifferentiated concepts, and they avoid the paradoxes to which the undifferentiated concepts lead. So, for instance, both goal-reaching and perceptual blurriness capture something about what we mean—intuitively—by “faster,” as do average and instantaneous velocity. At the same time, a concept of “faster” which fails to distinguish these elements is inadequate to the world it seeks to describe[226]. So the disambiguation is forced by exceptional cases which require the user of the concept to isolate within it two *ways* in which it captures the truth about the world.

A.2 Heat and Temperature

Our second example of conceptual disambiguation is taken from seventeenth- and eighteenth-century discussions of heat and temperature. My discussion here is guided by Marianne Wiser and Susan Carey's discussion of the episode in an article entitled "When Heat and Temperature Were One." In that article, Wiser and Carey examine the role played by the concept of *heat* in the work of the seventeenth-century Accademia del Cimento (Academy of Experiments) in Florence, where the first systematic studies of thermal phenomena were undertaken. Their central thesis is that

[226] That this is so is a function of the fact that not all motions occur with uniform speed. Kuhn notes that "in both these cases the concepts are contradictory only in the sense that the individual who employs them *runs the risk* of self-contradiction. He may, that is, find himself in a situation where he can be forced to give incompatible answers to one and the same question." But that this is so, contends Kuhn, does not mean that the concepts in question are "self-contradictory;" at most, he suggests, they are "confused" or "inadequate." But even this may be too strong: "Ought we demand of our concepts, as we do not and could not of our laws and theories, that they be applicable to any and every situation that might conceivably arise in any possible world?" In a world where all speeds were uniform, "the Aristotelian concept of speed could never be jeopardized by an actual physical situation, for the instantaneous and average speed of any motion would always be the same...if we found a scientist in this imaginary world consistently employing the Aristotelian concept of speed, [we would not say] that he was confused...Instead, given our own broader experience and correspondingly richer conceptual apparatus, we would like say that, consciously or unconsciously, he had embodied in his concept of speed his expectation that only uniform motions would occur in his world" (All quotes from Kuhn 1977, 254). (However, it is an interesting question to ask whether in, such a world, there would really *be* a concept of speed.)

the self-titled Experimenters "did not distinguish between heat and temperature" (Wiser and Carey 1983, 269). Let us try to make sense of what this might mean.

A century after the period studied by Wiser and Carey, the mid-eighteenth century Scottish physicist Joseph Black conducted a series of experiments through which he ascertained the basic principles of thermal science still accepted today. According to modern (post-Blackian) theories, heat and temperature are two completely different *kinds* of physical magnitudes: *heat*, which is the measure of the total kinetic energy of all the molecules in the body, is an extensive quantity; *temperature*, which is a measure of the average kinetic energy of individual molecules in the body, is an intensive quantity[227]. But in the work of the Experimenters, a single variable was used to refer to both of these features[228].

Wiser and Carey argue that this single variable was "neither modern heat nor modern temperature, but a mixture of both," and offer two reasons for this diagnosis: "First, [the Experimenters] lacked critical components of both modern concepts, components necessary to distinguish the two. Second, in those cases where they had knowledge of the properties that do now distinguish the two, these properties were then aspects of the same thermal concept" (Wiser and Carey 1983, 290-291).

[227] That is, heat, like all extensive quantities, is additive: the amount of heat in two cups of water is the amount of heat in the first plus the amount of heat in the second. Temperature, on the other hand, like all intensive quantities, is mediative: the temperature of a cup of water at 60 degrees F added to a cup of water at 80 degrees F will be 70 degrees F. (Cf. Carey 1988, 171.)

[228] Note that the quick and simplistic characterization that I offer in this section in no way does justice to the subtle analysis provided by Wiser and Carey.

So, for instance, where modern theories of temperature make use of a scale with two fixed points (such as the freezing and boiling points of water) and some sort of calibration (such as “degrees”) corresponding to portions of the interval between them, the Experimenters made use of three sorts of thermometers: the 50-degree, the 100-degree, and the 300-degree. These numbers corresponded to the number of degree-markings on the neck of the instrument. The thermometers were not calibrated in reference to fixed points; rather, the calibrations were designed to provide maximum use of the length of the neck. For instance, all 100-degree thermometers were designed to measure 20 when placed in snow, and 80 when placed in the sun. But a 50-degree thermometer would be calibrated otherwise; for instance (according to the reports of the experimenters themselves) “at the greatest cold of...winter [the 100-degree thermometer] subsid[ed] to 17 or 16 degrees [, while the 50-degree thermometer subsided] usually to 12, or 11.” (My discussion here relies on Wiser and Carey 1983, 286-287.)

What this means is that the single variable used by the Experimenters served neither the role of modern heat, nor of modern temperature. Unlike modern heat, the Experimenters’ variable could not be measured by “amount”; unlike modern temperature, it did not employ any consistent metric, corresponding to modern degrees. Moreover, the variable had both intensive and extensive aspects. The property measured was taken to be intensive in that its effects were held to be identical at all points in a homogenous body; but it was taken to be extensive in that increasing a source’s volume was held to increase the quantity of the property in question[229].

So the property that concerned the experimenters was in some ways like the property of heat, and in others like the property of temperature. In some ways, it measured the *total* kinetic energy of the molecules of a given body (heat), in others, the *average* kinetic energy of the molecules of that body (temperature). What forced the ultimate differentiation of the concept into two concepts corresponding to modern heat and modern temperature was the inability of the Experimenters’ theory to account for phenomena that they themselves acknowledged to be within its theoretical purview. So, for instance,

> [i]n one of their artificial freezing studies, they were surprised to find that adding more ice and more salt to the source did not influence the course of freezing of the water in the vessel. This would not have been puzzling...had they distinguished the temperature of the ice-salt mixture from the total heat capacity of the mixture (Wiser and Carey 1983, 292).

In the extant theoretical framework, where heat and cold were taken to be separate entities passed from a source body to a recipient body, *more cold* (produced by adding more ice and more salt) should have made the water freeze faster; that it did not was deeply perplexing to the Experimenters. In the modern framework, the puzzle is easily resolved: *less heat* does not mean *lower temperature*; the total may decrease while the average remains constant.

[229] For other differences between modern notions of heat and temperature and the concepts employed by the Experimenters, see Wiser and Carey 1983, 290-294.

Hence from our perspective, it seems that the Experimenters were guilty both of making distinctions without difference (in identifying one body as the source and the other as the recipient) and failing to mark differences with distinctions (in conflating heat with temperature). The shift in theoretical framework that followed the differentiation of the one and the consequent conflating of the other offers a second example of what I am calling *making a novel conceptual distinction.*

A.3 Promoting Democracy and Protecting Interests

Our third example of conceptual disambiguation is taken from a book widely used in university-level history courses, David W. Levy's *The Debate Over Vietnam*. One of Levy's central contentions is that the Second World War had produced in America an unprecedented consensus about the legitimacy of the country's foreign policy. He suggests that because the country had gained interests abroad—both commercial and territorial—a case could be made that self-interest required America to intervene in Europe to help reset the balance of power. At the same time, the forces they would find themselves opposing were unquestionably immoral, brutal and ruthless, deeply opposed to the ideals of democracy, freedom, fairness and peace. Typical expression of this view can be found in FDR's Quarantine Speech of October 5, 1937, where he argued that:

> the present reign of terror and international lawlessness...has now reached a stage where the very foundations of civilization are seriously threatened...Innocent peoples, innocent nations, are being cruelly sacrificed to a greed for power and supremacy...If those things come to pass in other parts of the world, let no one expect that America may expect mercy, that this Western Hemisphere will not be attacked and that it will continue tranquilly

> and peacefully to carry on the ethics and the arts of civilization (cited in Levy, 11-12).

Levy analyzes the speech, along with other letters and speeches of the period, contending that two separate arguments were at play: the first involving claims about the legitimate self-interest of the US in protecting itself and its allies, the second involving claims about the country's need and obligation to defend the ideals of western civilization against a brutal and barbaric enemy. He suggests that in the minds of many, these two arguments were "scrambled...together until the two propositions were almost inextricably joined, combined so intimately, so unconsciously, that it was hard to see that the blend was, in fact, composed of two ingredients. Indeed, only by means of the historian's trick of retrospective analysis are we able to see that *two* presumably separate arguments had been woven together" (Levy, 11).

Because the two justifications led to an agreement about practical questions, Levy contends, there was a fairly deep consensus across a wide political spectrum for nearly three decades—from the late 1930s until the mid-1960s—concerning the legitimacy and expediency of American foreign policy. But, he continues, this consensus was illusory. At base—perhaps consciously, perhaps unconsciously—people subscribed to the undifferentiated amalgam for different sorts of reasons: some were primarily concerned with the nation's interests abroad; others were primarily concerned with promoting democracy or freedom (or with opposing communism or tyranny); and some, presumably, were concerned with both goals—whether or not they happened to coincide. Looking back, Levy contends, this conflation of reasons seems striking; but at the time, there was little need to make the conceptual distinction. It was only when a differentiating case forced the question that the two justifications came conceptually unraveled. He writes:

> In retrospect, it is, perhaps, a little surprising that nobody paused to point out that Americans were subscribing to this foreign policy for different reasons, that they had come to their position of general support out of quite different concerns. Some were led by their deep attachment to certain ideals, beliefs, and principles; others, by the profound commitment to the necessity to keep their country safe. And in retrospect, it is, perhaps, a little surprising that nobody wondered very much about what would happen to the general consensus if these two articles of faith ever appeared to separate, to conflict, to lead in different directions (Levy, 20).

That is, because no *practical* differences separated those who supported the policy for one reason and those who supported it for another, these retrospectively distinct reasons remained conceptually undifferentiated.

There are, of course, cases where political alliances are created between groups who disagree fundamentally, without any confusion about the fact that these are strange bedfellows: the American alliance with the Soviet Union during World War II, collaborations between feminists and religious groups to restrict pornography, or the left-right alliance to legalize drugs. But what is interesting about Levy's argument—whether or not it is an accurate analysis of American popular opinion at mid-century—is that it suggests that in politics, as in child development and in the history of science, encounters with actual or hypothetical cases can lead to a disentangling of previously undifferentiated concepts.

Appendix B: Recent Characterizations of *Thought Experiment*

Nicholas Rescher: “Thought Experimentation in Pre-Socratic Philosophy”

> A “thought experiment” is an attempt to draw instruction from a process of hypothetical reasoning that proceeds by eliciting the consequences of an hypothesis which, for aught that one actually knows to the contrary, may well be false. It consists in reasoning from a supposition that is not accepted as true—perhaps is even known to be false—but is assumed provisionally in the interests of making a point or resolving a conclusion.

James G. Lennox: “Darwinian Thought Experiments: A Function for Just-So Stories”

> Thought experiments are:
>
> > [a] tests of a theory’s explanatory potential which
> > [b] posit hypothetical or counterfactual test conditions and
> > [c] invoke particulars which are irrelevant to the generality of the theory, and which
> > [d] are selected to instantiate features of the theory under special consideration.

James Robert Brown: “Thought Experiments: A Platonic Account”

> Thought experiments are performed in the laboratory of the mind. Beyond that bit of metaphor it’s hard to say just what they are. We recognize them when we see them: they are visualizable; they involve mental manipulations; they are not the mere consequence of theory-based calculation; they are often (but not always) impossible to implement as real experiments, either because we lack the relevant technology, or because they are simply impossible in principle.

Andrew D. Irvine: “Thought Experiments in Scientific Reasoning”

> This standard definition of thought experiment is too inclusive: a thought experiment is an instance of reasoning which attempts to draw a conclusion about how the world either is or could be by positing some hypothetical, or perhaps even counterfactual, state of affairs...Thought experiments, in the first instance, are simply arguments concerning particular events or states of affairs of a hypothetical (and often counterfactual) nature which lead to conclusions about the nature of the world around us...However, in addition, like physical experiments, thought experiments must stand in a privileged relationship both to past empirical observations and to some reasonably well-developed background theory.

John Norton: “Thought Experiments in Einstein’s Work”

> Thought experiments are arguments which:
> (i) posit hypothetical or counterfactual states of affairs, and
> (ii) invoke particulars irrelevant to the generality of the conclusion.

D. A. Anapolitanos: “Thought Experiments and Conceivability Conditions in Mathematics”

> A thought experiment is an exploratory ideal process meant to answer a theoretical or metatheoretical question in the general framework of the given discipline and carried out according to rules specified by logic and the particularities of the discipline itself.

J.N. Mohanty: “Method of Imaginative Variation in Phenomenology”

> A genuine thought experiment has to be an imaginative reconstruction of experience, of imaginative “transformation” of realities into fictional possibilities, in order to test hypotheses.

Roy Sorensen: *Thought Experiments*

> An *experiment* is a procedure for answering or raising a question about the relationship between variables by varying one (or more) of them and tracking any response by the other or others.
>
> A *thought experiment* is an experiment that purports to achieve its result without benefit of execution.

Barbara D. Massey: “Frege’s Thought Experiment Reconsidered”

> The notion that we are constructing a situation in imagination and then *observing* it in order to determine “what would be the case” seems basic to the thought experiment mode of inquiry. The thought experiment seems to discover facts about how things work in the laboratory of the mind.

Allen I. Janis: “Can Thought Experiments Fail?”

> First, a thought experiment is a description of an experimental procedure as well as its outcome or possible outcomes...Second, the outcome or possible outcomes must be deduced by reasoning consistent with a given theoretical framework.

Ronald Laymon: “Thought Experiments as Ideal Limits and as Semantic Domains”

A *thought experiment* is an ordered pair $<\Phi,\theta>$ where Φ is a set of persons (audience and/or presenter) and θ is a set of statements $\{T, P_1, P_2,...P_n, Q\}$ where:

(1) *T* is a description that is not in fact true (because it is idealized) of any experiment in this world;

(2) Members of Φ believe that $P_1, P_2,...P_n$ are scientific laws or principles;

(3) Members of Φ believe that $\exists x(Tx)\ \&\ P_1, P_2\ \&...\ P_n \Rightarrow Q$

A *thought experiment* (as defined above) is *successful* with respect to the set of persons Φ if:

(4) If $\Diamond\exists x(Tx)$ or if $\sim\Diamond\exists x(Tx)$ but *T* is asymptotically approachable, then members of Φ believe that it is possible to construct a series of real experiments, $e_1,e_2...e_n$, such that the description of each successive member more closely approximates *T*.

(5) If $\sim\Diamond\ \exists x(Tx)$, then the members of ϕ believe that it is possible to construct a set Ω of real experiments $\{e_1,e_2...e_n\}$ and a set Ψ of theories or analyses of inferring causes such that Ψ can be selectively applied to members of Ω so as to yield a series of residual analyses that converges to *T*.

(6)$\forall\alpha\in\Phi$, if α did not believe *Q* before being presented with θ then α believes *Q* after being presented with θ.

Peter King: “Mediaeval Thought-Experiments: The Metamethodology of Mediaeval Science”

Mediaeval philosophers took thought-experiments to be rather like consistent sets of sentences, enabling them to bypass “mentalistic” aspects and focus on logical aspects. A thought experiment consists of a set of sentences D which are to be understood as a description of a situation, together with the claim that D describes a case in which P, where P describes “what happens” in the situation.

Appendix C: The Ring of Gyges

In this Appendix, by discussing two criticisms of Plato's story of the Ring of Gyges, I try to show why unimaginability claims (see section **1.3.2**) are largely irrelevant when leveled against scenarios that are *idealizations* along dimensions where clear tendencies can already be identified. I will argue that this is so not only for unimaginability claims where charges of *underdescription* are leveled, but even for certain cases where the charges are of *incoherence*. I contend, however, that such unimaginability objections *do* bring out certain preconditions for the ordinary application of our concepts-features which are *coinstantiated* with the features under consideration. If these features are imaginatively separated within the context of a scenario, the ordinary background conditions for making judgments may be altered. Thus first-person projections may be rendered untrustworthy. “What would I do?” or “What would I care about?” may be questions whose answers cannot be reliably ascertained by means of such imaginative projection.

Appendix C thus serves two purposes. It underscores the fruitfulness of the taxonomy described in section **1.3.2** by showing how the taxonomy is able to account for arguments made by two of the most widely-cited contemporary discussants of the technique of thought experiment (Roy Sorensen and Kathleen Wilkes). And it reinforces the main contention of the dissertation by providing yet another example of a case where imaginary projection is rendered unreliable due to a disruption of the ordinary background conditions under which we make judgments of a certain sort.

C.1 The Story

In a famous debate with Socrates concerning the nature of justice, Glaucon is reported to have maintained that “those who practice justice do it unwillingly and because they lack the power to do injustice” (*Republic*, 359b), that is, that the only reason anyone acts justly is our of fear that she will be caught and punished. In good scientific fashion, Glaucon does not merely assert his view; rather, he proposes to establish it by means of a controlled experiment. He suggests two cases to be compared. In the first case, all ordinary constraints are in place, and we are presented with two persons, one of whom (A) acts justly, the other of whom (B) does not. In the second case, the possibility of being caught is eliminated from the equation, and we are to “follow both of them and see where their desires would lead” (*Republic* 359c). The experiment is set up so that whatever differences there are between the actions of A in the first case and the actions of A in the second case can be attributed to the sole difference between the cases, namely, that in the second case of the possibility of being caught and punished for wrongdoing has been eliminated. Since A is presumed to be representative of all just persons, whatever we learn about A from this comparison can be generalized to just persons in general[230].

[230] What role is the *unjust* character (B) playing in the experiment? If the only reason he is considered is to show that the absence of punishment is a sufficient but not a necessary condition for acting unjustly, then his presence seems like unnecessary clutter. That impunity is not a necessary condition for wrongdoing can be established simply by pointing out that injustice and punishment coexist. Rather, I suggest that his presence enables Glaucon's experiment to play a role parallel to that of the *experimentum crucis* in Newton's Opticks (I:I:II:II:6)—though with a conclusion that runs in the opposite direction.

In the passage in question, Newton is trying to establish that “The Light of the Sun consists of Rays differently Refrangible” and that “Lights which differ in Colour, differ also in Degrees or

Continues on next page…

The experiment Glaucon proposes, however, is not to be *actually* conducted. Rather, he suggests, we can see his point "most clearly...if *in our thoughts* we grant to a just and an unjust person[231] the freedom to do whatever they like" (*Republic*, 359bc, italics added). That is, Glaucon is suggesting that we already in some sense *know* how a just person would act under conditions of impunity; we just do not realize it because

Refrangibility" (Opticks, 26, 20). In earlier experiments, he had shown that when a beam of white light is projected through a prism, it is broken up into a rainbow of colors, which can be projected on a wall in a roughly oblong shape. In the experiment in question, he isolates in turn each of the colors which had been produced by projecting the light through the first prism, and then projects that single color through a second prism (see illustration).

The result of this experiment is that the contrasts between the differently-refractable colors are exaggerated; projected through a second prism, red light bends proportionately less than violet light, and the differences between them—already present in the first projection—are magnified still more.

By contrast, Glaucon is taking two individuals who differ already in terms of justice, and testing whether impunity exaggerates, holds constant, or erases the difference between them. That is, impunity might be something with constant effects that would make both just and unjust persons better or worse, or leave them both the same. *Or* it might be something with differential effects such that it increased the differences between just and unjust persons. *Or*—and this is Glaucon's hypothesis—it might be something with differential effects that levels both just and unjust to a common denominator.

Thus the reason the second character is there is to provide information about the kind of thing impunity is. He is supposed to show that impunity has the following effect: nothing about the characters of persons in an impunitous world could tell you *which* ones would be just in a world with punishment; whereas anything about the characters in a world with punishment could tell you which ones would be unjust in an impunitous world—namely *all* of them.

[231] By a "just person," Glaucon means: one who acts in keeping with laws and norms under ordinary circumstances. On this picture, a just person differs from an unjust person in terms of (actual) *behavior*, but not necessarily in terms of *character* (that is, dispositional behavior). In my discussion of the Ring of Gyges case, I will use "just person" in this Glauconian sense.

we have not organized the information in a way that renders its implications apparent to us[232]. The proposed experiment, then, is to be a thought experiment. The particulars are as follows:

> The freedom I mentioned would be most easily realized if both people had the power they say the ancestor of Gyges...possessed...[He found a ring which] if he turned the setting inward, he became invisible; if he turned it outward, he became visible again[233]...Let's suppose, then, that there were two such rings, one worn by a just and the other by an unjust person. Now no one, it seems, would be so incorruptible that the would stay on the path of justice or stay away from other people's property, when he could take whatever he wanted from the marketplace with impunity...Rather, his actions would be in no way different from those of an unjust person, and both would follow the same path...[W]henever either person thinks he can do injustice with impunity, he does it (*Republic*, 359c-360c).

Glaucon’s thought experiment fits cleanly into the tripartite structure proposed in section **1.3.2**. He begins by presenting an imaginary scenario: A ring is described which, if worn facing inwards, renders its wearer invisible, and it is supposed that there are two such rings, one issued to a just, the other to an unjust person. The next segment of the text involves reasoning within the context of this imaginary scenario: Glaucon contends that, armed with the freedom to take whatever he wants with impunity, the just person acts exactly as the unjust person does.

[232] Thus the way in which this experiment in thought functions is like an ordinary experiment: it takes information already (in some sense) available to us through ordinary experience, and allows us to organize it in such a way that its implications become apparent.

[233] The ancestor of Gyges used the ring in the following way: “He at once arranged to become one of the messengers sent to report to the king. And when he arrived there, he seduced the king's wife, attacked the king with her help, killed him, and took the kingdom” (*Republic*, 360ab). Presumably, this is not sufficient to make Glaucon’s point since we do not know whether Gyges ancestor was a just man *before* he found the ring.

Finally, he applies the results of his reasoning to the actual world: “whenever either person thinks he can do injustice with impunity, he does it.”

C.2 Criticisms of the Case

What I want to do in the next few pages is to look at two recent critiques that have been offered of the Gyges example, both of which are presented as criticisms of the case's imaginability. The first, from Kathleen Wilkes, suggests that the case is *(resolvably* or *unresolvably) underdescribed*; the second, from Roy Sorensen, suggests that the case is *(tacitly) incoherent*. Wilkes presents her objection in her own voice; Sorensen presents his as coming from an imaginary interlocutor, and then offers replies to the objection. But both share the view that, for the thought experiment to work, the objections need to be answered. I will contend that this is not so (at least not in the way that Wilkes and Sorensen propose to answer them), and that the *reason* the objections need not be answered (in that way) is that the scenario involves idealization along dimensions where clear tendencies are already evident. At the same time, I will argue that each of these objections can be *reframed* as an objection that shows the conceptual centrality of the *actual* concomitance of some feature with all instances of human agency.

What underlies each of the objections is this: the thought experiment will work (that is, provide informative data about whether fear of punishment is the only thing that leads to just behavior) *only if* we already (in some sense) *know* how someone would act under such circumstances—if we can work it out from information we already have by recombining it in a novel but determinate way. If what we already believe *fixes* what we would say in this case—that is, if there is only one answer that can be offered that can consistently and completely account for our commitments—then the scenario described can serve the role of a controlled situation around which information can be perspicuously reconfigured.

For instance, if we already know what happens in A-type situations, and B-type situations can be shown to be like A-type situations in all relevant ways, then (assuming that we are correct about A-type situations) we might learn something about what happens in B-type situations. Our learning comes from working out an implication: A-type situations have feature x, B-type situations are relevantly like A-type situations, therefore B-type situations have x. The work of the thought experiment in such a case is likely to be in helping us to see the relevant similarities: consideration of a scenario in which the salient features of A and B are highlighted can help us to realize that A-type and B-type situations are relevantly similar; having seen this, we can then conclude that B-type situations also have x[234].

By contrast, thought experiments that do not bring us to see that an unfamiliar case is relevantly similar to a familiar case must work by alternate means. Typically, they operate as follows: instead of positing the property in question explicitly, they identify a *surrogate* for the property. (So, for instance, in presenting the Ring of Gyges case, Glaucon does not posit explicitly that the characters act without risk of discovery or punishment for misdeed; rather, he identifies a *surrogate* for impunity, namely invisibility, and describes a situation in which invisibility plays a particular

[234] For an argument that such relevant similarities rarely obtain, see Dancy 1993.

role.) The reason for this is that the property in question is—*ex hypothesi*—not one about which we are able to make a judgment (for if it were, we would be back to a case of the first sort).

So a global criticism can be made that inevitably, the following dilemma will hold: either the surrogate will be something about which we can make determined judgments, but it will not test the novel property; or the surrogate will test the novel property, but it will not be something about which can make determined judgments. In the first case, we have not succeeded in testing what we wanted to measure, so we must find a new way to get the information desired. But in the second case we are not yet at a loss. What we *can* do is to perform an experiment in thought—namely—imagine ourselves to be in such a situation, and figure out what we would do[235]. But if we are going to take *this* as reliable, then we have to assume that what I imagine I would do is a good guide to what I would, in fact, do. This pattern is clearly illustrated by the examples from Wilkes and Sorensen.

C.2.1 Underdescription: Wilkes

As part of a general critique of thought experiments, Kathleen Wilkes offers a paragraph-long objection to the Gyges case. She writes:

> Before we can make sense of [the Gyges ring] thought experiment, several points press to be answered...For instance, is the owner of the ring to be intangible as well as invisible? That makes a substantial difference to *the issue*

[235] Why can't we just imagine *someone* being in that situation, and figure out what *she* would do? Perhaps because we don't have enough of a full-fledged sense of normative rationality to say what someone *ought* to do in that (sort of) situation.

> *at issue*: if he is not intangible, he might by mistake bump up against, and get arrested by, a policeman, or get his hand slammed in a till drawer. Thus a potential criminal may yet have self-interested reasons for staying within the bounds of morality...[I]f you are both invisible and intangible...could *you* hold a gun, or a caseful of banknotes?...[W]ould others know that one owned such a ring? If so, then there might be extra reason for *remaining* moral: viz., that unsolved crimes might otherwise be ascribed to you...The point is that...[t]he background is inadequately described, and the results therefore inconclusive (Wilkes 1988, 11).

In this paragraph, Wilkes offers two criticisms of the Gyges thought experiment; I will discuss each in turn.

C.2.1.1 Wilkes's First Objection

Her first criticism is that, as it stands, the case is *underdescribed*: we do not have sufficient grounds for predicting the behavior of the ring-wearer, since certain crucial features of the situation remain unspecified. For example, our predictions about what an invisible but tangible agent would do differ from our predictions about what an invisible and intangible agent would do, and since the original version of the story says nothing about tangibility, we are unable to make a judgment about this case.

An obvious response to this first criticism is that the underdescription Wilkes has pointed out is at worst resolvable and at best irrelevant. Following Wilkes's lead, we might consider the case as involving two sub-cases about which we are obliged to make judgments, one involving an agent who is invisible and tangible, and the other an agent who is invisible and intangible. If our judgments about these two cases were to diverge, then some disambiguation would indeed be necessary. In that case, we would end up with two different thought experiments—one about invisible-intangible agents, the other about agents who were invisible yet tangible—and Wilkes would have successfully shown that the case suffered from resolvable underdescription where resolution led to a bifurcation of the case. As a matter of fact, however, this disambiguation is not necessary. Both of the sub-cases are cases about which we would make the same *sort* of judgment, differing only in degree. Assuming that the characters act on the basis of self-interest, we would expect an invisible but tangible agent to violate *some* of the ordinary moral constraints, and (for the reasons that Wilkes herself gives) we would expect the invisible-intangible agent to violate still more[236]. So the failure to specify whether the ring gives intangibility as well as invisibility seems irrelevant to the force of the example. Underdescription is a fatal flaw only if it is unresolvable, and if something turns on it. Wilkes's first criticism fails to establish either one.

Wilkes might reply: “But that wasn’t what I meant when I suggested that we need information beyond that specified in the original case. What I meant was: if you don't *realize* the ambiguity, you might predict that the invisible character would act justly in certain cases (for instance, where she runs a risk of getting caught) without realizing that you are making this prediction because her self-interest is still playing a role in your calculation. You would then mistakenly conclude that the just and unjust person would act differently under conditions of impunity, where all you had *really*

[236] If we assume that the characters do *not* act on the basis of self-interest, we would expect intangibility (or even invisibility) to make no difference at all to their behavior.

considered were conditions of invisibility." This suggests the Wilkes *real* first objection is not about what has or has not been specified in the case; her *real* first objection is that invisibility is not a perfect surrogate for impunity. If we are trying to determine whether the possibility of being detected plays a role in our decision to act justly, imaging ourselves capable of invisibility alone will not allow us to rule out all possibility of being detected.

C.2.1.2 Wilkes's Second Objection

Wilkes second objection is that we don't quite know what the world would be like for someone who is invisible and intangible. For instance, she asks: "[C]ould prison walls hold you? And if they could not, could *you* hold a gun, or a caseful of banknotes?" (Wilkes 1988, 11). Again, it seems that this underdescription is either resolvable (we can go through each of the cases and ascertain what we would say[237]) or irrelevant (if we were to say the same thing in each case). But again, we can reinterpret this objection in a way that seems to cut much deeper.

On the reinterpretation, Wilkes argument is this: The thought experiment is supposed to help us decide whether the prospect of being caught is what leads us to act justly. In order to test this, it invokes a supposedly reliable surrogate for impunity, namely invisibility. But as the first objection shows, invisibility is not strong enough

[237] That is, the case where being intangible means that you can hold a gun and the case where it means that you cannot, etc.

to serve as such a surrogate; it becomes strong enough only strengthened to invisibility-plus-intangibility[238]. And once we strengthen it that much, we become rather unclear on what it would be like to be that way. That our actions have consequences in the world, and that those consequences can be, at least in some cases, attributed to us as agents, is so deeply a part of our conceptual scheme that we don't really know what to say about actors so divorced from (certain of) their actions that such connections could never be drawn by others. Would one act justly under such conditions? Wilkes would say: we simply don't know. And our not knowing is not coincidental. The possibility of being caught is so central to our conception of being an agent that we cannot fully abstract away from it and still be talking about *agency*. So the reason we cannot judge reliably how people would act under conditions of impunity is that part of our concept of *action* is tied up with our concept of it being possible for others to attribute that action to its doer. I will return to this argument after presenting the second case.

C.2.2 Incoherence: Sorensen

In his discussion of the Ring of Gyges, Roy Sorensen presents and then dismisses an objection that seems similarly irrelevant.

> Plato’s Ring of Gyges...has serious run-ins with contemporary opthamology. To see, the eye lens must bend light to form an image on the retina. But a transparent lens has the same index of refraction as the air (and so cannot bend light) and a transparent retina cannot absorb light. Hence, the invisible

[238] And even then it may not be strong enough. For instance (as Wilkes notes) the risk of circumstantial evidence must also be blocked.

> man could not see! But an ethicist who tried to rebut Plato by stressing the disadvantages of blindness would be laughed down...There are two rationales for the amusement. The first...concedes that the Ring of Gyges case is flawed but stresses the triviality of the flaw. Plato's main point is easily salvaged by substituting another scenario...[and] we are so confident of the possibility of a backup that we do not bother to actually construct it...The [second] rationale admits no error at all...say[ing] that we should relativize to the overt beliefs of Plato's community, not to the actual world. If so, then the bearer of the ring can see even though he is invisible. (Sorensen 1991, 287-288)

The objection Sorensen puts in the mouth of his interlocutor certainly seems, on the face of it, to have missed the entire point of the case. To remark on a minor biological mistake in the construction of the story seems to be evidence that one has misunderstood the issue at hand.

But there is an alternative way of understanding the objection Sorensen raises and dismisses as parallel to the argument I have reconstructed for Wilkes. Her objection, as I have reconstructed it, is that the case in uninformative because we do not really know what it would be like to *act* unobserved. Sorensen's reconstructed argument concerns one particular aspect of this problem: it says that we do not know what it would be like to *observe* unobserved. In both cases, the argument rests on the suggestion that something that might initially appear to be a coincidental feature of our being agents is actually central and indispensable.

On the reconstruction I am proposing, Sorensen's objector might begin by saying this: it is not coincidental that invisibility ends up interfering with the possibility of seeing, since seeing is something we do as agents in the world, and being agents in the world means we are, at least in principle, observable. To this the defender could reply: but one can observe unseen—from a mountaintop, or with a telescope, or through a hidden video camera, or while someone is asleep—and these possibilities seem to cause no conceptual difficulties. Here Sorensen’s interlocutor could retort: But in each of those situations, when we are on the other side of things (when we are in a valley, or near an open window, or taking money from an ATM, or asleep) we know that there is a *chance* that we are being observed, even if we are unable to ascertain that we are being observed *right then*.

He could continue: Now if invisibility were possible, there are two options: either it would fall in exactly that category, so that the only way in which things would be different is that one could *never* be sure that one was unobserved; there would always be a sneaking suspicion that someone might be watching. (The defender might say: but that's the case right now—on any reasonably fallibilist picture of things, we might be wrong at any time about whether we are being observed or not. Reply: This is certainly true in *some* sense, but the scenario in which invisibility is a *live* possibility is a scenario in which an entire *dimension* of being wrong that one is unobserved has been introduced.) Or it would alter things so radically that the idea of *being unobserved* would lose all sense. (Think of devoutly religious people who act always with the awareness that God is watching.) In the first case, we can observe unseen, but only in a local sense. In the second, we can observe unseen in the global sense, but only because “observe” has taken on a different role in the world. Moreover, if knowledge obtained unseen is to make any difference in our lives, we open ourselves to the possibility of discovery on circumstantial grounds. That is, either we risk being “found out” (by acting on knowledge we have obtained through such surreptitious means), or the knowledge must remain compartmentalized in a way that does not affect our actions. That there is a *physical* reason Plato's story won’t work is simply a function of the way in which, metaphorically speaking, the agency lump reappears at that point under the rug.

C.3 Evaluation of the Arguments

Both the Sorensen and the Wilkes objections assume the following. (1) That we don’t know in advance what role impunity plays in our decision to be moral. (2) That we are using this thought experiment as a means of finding out. And (3) that in order for the thought experiment to serve this purpose, it must describe a surrogate for impunity with the following properties: (a) the surrogate (almost) perfectly tracks impunity and (b) the surrogate is something we can imagine in sufficient and precise enough detail to answer (most) “what it would be like” questions consistently and unambiguously.

What (3a) is supposed to guarantee is that we are testing what we are interested in, given the constraints imposed by (1); what (3b) is supposed to guarantee is that the test we are using is a reliable guide to our *actual* intuitions—our intuitions about what we (tacitly) think is relevant to our decisions to act morally. We are interested in what our impunitous counterparts would do because we are interested in why we ourselves do what we do.

What the objections we are considering suggest is that (3a) and (3b) cannot be met simultaneously. And they suggest that this is for a single simple reason. The apparently irrelevant *physical* fact that how we act and how we observe ineliminably involve certain mechanical processes is not, they suggest, a coincidental aspect of what it is that we mean when we say “act” and “observe.” Rather, “act” and "observe" refer to physical processes, *and* physical processes are in principle discoverable by us. So, the objection goes, since acts are physical and physical processes are in principle discoverable, then acts must be in principle discoverable. It doesn't make sense to ask whether persons act justly because they face a risk of being caught, since part of what it is to *act* is to face a risk—however slight—of being caught. It is not just *as a matter of fact* that it is part of what it is to act; it is part of what it is to act *as part of the concept of acting*.

Extreme as this argument may seem, I am actually convinced that it is correct. But I am equally convinced that as a criticism of the Ring of Gyges case at the level of imaginability, it is almost entirely irrelevant. And the *reason* that it is irrelevant seems to me quite a general one. At the same time, I think the criticism helps bring out one of the several reasons that Glaucon's predictions about the imaginary scenario he presents are wrong. And this, too, is on rather general grounds.

Since I think the reconstructed Wilkes and Sorensen are right that (3a) and (3b) cannot be met simultaneously, and since I will assume that they are right about (1) and (2),[239] then if their criticism is unsuccessful, it must be because they are wrong about (3). That is, they must be wrong that in order for the thought experiment to help us learn about what role impunity plays in our decision to act justly, it must describe a surrogate that: (a) (almost) perfectly tracks impunity and that: (b) we can

[239] (1) That we don't in advance know what role impunity plays in our decision to be moral. (2) That we are using this thought experiment as a means of finding out.

imagine in precise enough detail to answer (almost all) “what it would be like” questions consistently and unambiguously.

I agree with the reconstructed Wilkes-Sorensen that both (3a) and (3b) are important conditions. Where I think they are wrong is in requiring that they be met simultaneously. Rather, for a situation to be *relevantly* imaginable, the surrogate employed need only meet the conditions one at a time. In the case of Gyges, the scenario meets (b) but not (a) in that there are a spectrum of *fully imaginable* cases where one is less or more invisible (that is, less or more likely to be seen), even though the last case on the spectrum doesn't come *close* to approximating impunity. And the scenario meets (a) but not (b) in that invisibility (precisely because it carries with it all of the problems Wilkes and Sorensen point out) almost perfectly tracks impunity for a wide range of cases.

All this is simply to say that we can distinguish between cases where one is likely to be caught (because one is seen), and cases where one is extremely unlikely to be caught (because one is unseen), and compare the proportionate tendency to act unjustly. The parking lot attendant who would not steal a car in broad daylight as the owner looked on might be part of a ring that removes stereo systems late at night. And the housesitter, curious to know things about her friends that they would never tell her directly, may snoop through their files, carefully replacing each to its original location in the drawer so as to hide what she has done[240]. These cases are fully

[240]Like a computer hacker, one might even take pride in successfully snooping unseen—so much pride that (ironically) one leaves a message to tell the snooped-upon that one has secretly intruded upon her

Continues on next page…

imaginable, and *not being seen* is a reasonably good surrogate in them for *not being caught*, even though it is no guarantee. (For instance, a passerby might overhear the conspirators as they pry a sound-system from the dashboard of a vehicle, or the hapless housesitter might later make public reference to heretofore secret information about her friend's divorce.)

Where Glaucon’s analysis of the case founders is in his inappropriate use of analogic extrapolation. Although more closely observed behavior generally leads to unjust action being less frequent, and less closely observed behavior generally leads to unjust action being more frequent, it surely does not follow that *unobserved* behavior leads to unjust action in *all* cases[241]. (Nor that constant scrutiny leads to unjust action in *no* cases.) And one of the reasons this analysis founders is because the ends of the spectrum are, as the Wilkes-Sorensen objection brings out so clearly, ill-defined. That is, one reason that when we (somewhat incoherently) imagine ourselves invisible, we do not imagine ourselves murdering (even though we may imagine ourselves spying or stealing or taking vengeance on an enemy) is because the spectrum along which we are extrapolating becomes fuzzy that far down[242].

purportedly impenetrable files. (See “Hackers Taking a Byte Out of Computer Crime” *Technology Review* (MIT alumni monthly) April 1995.)

[241] There is an old joke about this. The pilot announces over the loudspeaker that the flight will be landing 20 minutes later than expected because one of the engines has fallen off. A few minutes later, the pilot comes on again to say that the plane will be landing an hour late, because a second engine has fallen off. Twenty minutes later, the co-pilot announces apologetically that the plane will be landing four hours late, because a third engine has fallen off. At this point, our hero turns to the passenger sitting next to him and says: “Well, I hope the fourth engine doesn’t fall off, or we’ll be up here all night.”

[242] Additionally, it may be that our reasons for not committing certain sorts of crime differ in *kind* from our reasons for not committing certain others; or the spectrum may not be fully sequential.

None of this is sufficient to undermine the Gyges story as a way of learning about whether we think people act justly out of fear of being caught, or as a way of learning about whether we think they do so for reasons independent of the possible gaze of others. A situation may be relevantly imaginable without being fully imaginable; to contend otherwise is to misunderstand the role of imagination in the acquisition of new knowledge.

C.4 Conclusion

The purpose of this Appendix has been twofold: to bring out the fruitfulness of the taxonomy described in section **1.3.2**, and to reinforce the main contention of the dissertation by providing an additional example where background conditions play a role in enabling certain sorts of judgments. In section **C.2**, I presented two criticisms of the story of the Ring of Gyges: Wilkes's critique of the story as *unimaginable* because it is *underdescribed*, and Sorensen's critique of the story as *unimaginable* because it is *incoherent*. I suggested that as they stand, neither objection is convincing.

But I went on to argue that each can be reformulated as a contention that a certain feature of human finitude (that we can neither act nor observe without the possibility of our actions being detected by others) plays a crucial role in our conception of human agency. In this light, I suggested, both Sorensen and Wilkes are right to call the case *unimaginable*. At the same time, I contended that imaginability is not the relevant criterion in evaluating the thought experiment's informativeness.

Such objections are largely irrelevant when leveled against scenarios that are idealizations of the sort described in this case.

Bibliography

Ackrill, J. L., editor (1987). *A New Aristotle Reader*. Princeton: Princeton University Press.

Anapolitanos, D. A. (1991). “Thought Experiments and Conceivability Conditions in Mathematics.” In Horowitz and Massey (1991).

Armstrong, D.M. (1983). *What is a Law of Nature?* Cambridge: Cambridge University Press.

Baillie, James (1993). “Recent Work on Personal Identity.” *Philosophical Books* Vol. XXXIV, No. 4 (October, 1993): 193-206.

Bealer, George (1996). “The Autonomy of Philosophy.” Unpublished manuscript. Presented at “Rethinking Intuition: The Psychology of Intuitions and their Role in Philosophical Inquiry.” University of Notre Dame, April 1996.

Brewer, Scott (1994). “Semantics, Pragmatics, and the Rational Integrity of Legal Analogy.” Unpublished manuscript.

Brody, Baruch (1980). *Identity and Essence*. Princeton: Princeton University Press.

Brooks, D.M.H. (1994). “The Method of Thought Experiment. ” *Metaphilosophy* Vol. 25, No. 1 (January 1994): 71-83.

Brown, James Robert (1991a). *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. New York and London: Routledge.

Brown, James Robert (1991b). “Thought Experiments: A Platonic Account.” In Horowitz and Massey (1991): 119-128.

Brown, James Robert (1993). “Why Empiricism Won’t Work.” *Proceedings of the Philosophy of Science Association* 2, 271-279.

Brueckner, Anthony (1993). “Parfit on What Matters in Survival. ” *Philosophical Studies* 70, 1-22.

Burke, Michael B. (1994a). “Dion and Theon: An Essentialist Solution to an Ancient Puzzle.” *Journal of Philosophy*, Vol. 91, No. 3 (March 1994): 129-139.

Burke, Michael B. (1994b). “Preserving the Principle of One Object to a Place: A Novel Account of the Relations Among Objects, Sorts, Sortals, and

Persistence Conditions. *Philosophy and Phenomenological Research* Vol. LIV, No. 3 (September 1994): 591-624.

Butler, Joseph (1736/1975). “Of Personal Identity.” First Appendix to *The Analogy of Religion*. Reprinted with new pagination in Perry 1975a.

Butts, Robert E. and Joseph C. Pitt, eds. (1978). *New Perspectives on Galileo*. Dordrecht/ Boston: Reidel.

Carey, Susan (1988). “Conceptual Differences Between Children and Adults.” *Mind and Language* Vol. 3, No. 3 (Autumn 1988): 167-181.

Cargile, James (1987). “Definitions and Counterexamples.” *Philosophy* Vol. 62: 179-193.

Carrier, Martin (1993). Review of Horowitz and Massey 1991. *Erkenntnis* 39: 413-419.

Carrithers, Michael, Steven Collins and Steven Lukes, editors (1985). *The Category of the Person: Anthropology, Philosophy, History*. Cambridge: Cambridge University Press.

Chisholm, Roderick M (1971). “Problems of Identity.” In Munitz (1971): 3-30.

Claggett, Marshall (1959). *The Science of Mechanics in the Middle Ages*. Madison: University of Wisconsin Press.

Clavelin, Maurice (1974). *The Natural Philosophy of Galileo: Essays on the Origins and Formation of Classical Mechanics*. Trans. A. J. Pomerans. Cambridge: MIT Press.

Clement, John (1983). “A Conceptual Model Discussed by Galileo and Used Intuitively by Physics Students.” In Gentner and Stevens (1983): 325-340.

Cockburn, David, ed. (1991). *Human Beings*. Cambridge: Cambridge University Press.

Cooper, Lane (1935). *Aristotle, Galileo, and the Tower of Pisa*. Ithaca: Cornell University Press.

Cummins, Robert (1996). “Reflections on Reflective Equilibrium.” Unpublished manuscript. Presented at “Rethinking Intuition: The Psychology of Intuitions and their Role in Philosophical Inquiry.” University of Notre Dame, April 1996.

Dancy, Jonathan (1985a). *An Introduction to Contemporary Epistemology*. Oxford: Basil Blackwell.

Dancy, Jonathan (1985b). “The Role of Imaginary Cases in Ethics.” *Pacific Philosophical Quarterly* 66: 141-153.

Dancy, Jonathan (1993). *Moral Reasons*. Oxford: Basil Blackwell.

Dancy, Jonathan (forthcoming). *Reading Parfit*. Oxford: Basil Blackwell.

Damerow, Peter et al. (1992). *Exploring the Limits of Pre-classical Mechanics: A Study of Conceptual Development in Early Modern Science: Free Fall and Compounded Motion in the Work of Descartes, Galileo and Beeckman*. New York: Springer Verlag.

Dauer, Francis W. (1972). “How Not to Reidentify the Parthenon.” *Analysis* 33:2 (December 1972): 63-64.

Davis, Lawrence H. (1973). “Smart on Conditions of Identity.” *Analysis* 33:3 (January 1973): 109-110.

Davis, Philip J. and David Park, editors (1987). *No Way: The Nature of The Impossible*. New York: W.H. Freeman and Company.

Dennett, Daniel (1984). *Elbow Room*. Cambridge: MIT Press.

Dennett, Daniel (1991). *Consciousness Explained*. Boston: Little Brown.

Dijksterhuis, E. J. (1961, repr. 1986). *The Mechanization of the World Picture: Pythagoras to Newton*. Trans. C. Dikshoorn. Princeton: Princeton University Press.

Drake Stillman and I. E. Drabkin (1969). *Mechanics in Sixteenth Century Italy: Selections from Tartaglia, Benedetti, Guido Ubaldo,Galileo*. Madison: University of Wisconsin Press.

Drake, Stillman (1989). *History of Free Fall: Aristotle to Galileo*. Toronto: Wall & Thompson.

Drake, Stillman (1990). *Galileo: Pioneer Scientist*. Toronto: University of Toronto Press.

Ethics 96 (July 1986).

Fodor, Jerry A. (1971). “On Knowing What We Would Say.” In Jay Rosenberg and Charles Travis, *Readings in the Philosophy of Language*. Englewood Cliffs, NJ: Prentice Hall: 198-212.

Foot, Philippa (1978). “The Problem of Abortion and the Doctrine of Double Effect.” In *Virtues and Vices*. Oxford: Blackwell.

Frege, Gottlob (1879/1953 repr. 1980). *The Foundations of Arithmetic*, trans. J.L. Austin. Evanston, IL: Northwestern University Press.

Gale, Richard (1991). “On Some Pernicious Thought Experiments.” In Horowitz and Massey (1991): 297-304.

Galilei, Galileo (1632/1967). *Dialogue Concerning the Two Chief World Systems*. Trans. Stillman Drake. Berkeley: University of California Press.

Galilei, Galileo (1638/1914, repr. 1954). *Dialogues Concerning Two New Sciences*. Trans. Henry Crew and Alfonso de Salvio. New York: Dover.

Galilei, Galileo (1638/1974, rev. 1989). *Two New Sciences, Including Centers of Gravity and Force of Gravity and Force* of Percussion. Trans. Stillman Drake. Toronto: Wall & Thompson.

Garrett, Brian (1990). “Personal Identity and Extrinsicness.” *Philosophical Studies* 59: 177-194.

Garrett, Brian (1991). “Personal Identity and Reductionism.” *Philosophy and Phenomenological Research* LI:2 (June 1991): 361-373.

Gentner, Dedre and Albert L. Stevens, eds. (1983). *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goffman, Erving (1959). *The Presentation of Self in Everyday Life*. Garden City, NY: Doubleday Anchor.

Goldman, Alvin (1989). “Psychology and Philosophical Analysis.” *Proceedings of the Aristotelian Society* 39: 195-209.

Goldman Alvin and Joel Pust (1996). “Philosophical Theory and Intuitional Evidence.” Unpublished manuscript. Presented at “Rethinking Intuition: The Psychology of Intuitions and their Role in Philosophical Inquiry.” University of Notre Dame, April 1996.

Gooding, David C (1993). “What is *Experimental* about Thought Experiments?” *Proceedings of the Philosophy) of Science Association* Vol. 2: 280-290.

Gopnik, Alison (1996). “Whose Concepts are They, Anyway?” Unpublished manuscript. Presented at “Rethinking Intuition: The Psychology of Intuitions and their Role in Philosophical Inquiry.” University of Notre Dame, April 1996.

Grice, H. P. (1941/1975). “Personal Identity.” *Mind* Vol. 50 (October 1941). Reprinted with new pagination in Perry (1975a).

Hacking, Ian (1993). “Do Thought Experiments Have a Life of Their Own?” *Proceedings of the Philosophy) of Science Association* Vol. 2: 302-308.

Haksar, Vinit (1991). *Indivisible Selves and Moral Practices*. Edinburgh: Edinburgh University Press.

Hanson, Norwood Russell (1958). *Patterns of Discovery*. Cambridge: Cambridge University Press.

Harman, Gilbert (1986). “Moral Explanations of Natural Facts: Can Moral Claims Be Tested Against Reality?” *Southern Journal of Philosophy* 24 supplement.

Hart, Ivor B**.** (1923). *Makers of Science: Mathematics, Physics, Astronomy*. London: Oxford University Press.

Helm, Paul (1975). “Are ‘Cambridge’ Changes Non-Events?” *Analysis* 35:4 (March 1975): 140-144.

Hertzberg, Lars (1991). “Imagination and the Sense of Identity.” In Cockburn (1991): 143-156.

Hesse, Mary (1966). *Models and Analogies in Science*. Notre Dame: University of Notre Dame Press.

Hirsch, Eli (1982). *The Concept of Identity*. New York: Oxford University Press.

Hobson, R. Peter (1990). “On the Origins of Self and the Case of Autism.” *Development and Psychopathology* 2, 163-181.

Horgan, Terence (1993). “On What There Isn’t.” *Philosophy and Phenomenological Research,* Vol. LIII, No. 3 (September 1993): 693-700.

Horowitz, Tamara and Gerald Massey, eds. (1991). *Thought Experiments in Science and Philosophy*. Savage, MD: Rowman and Littlefield Publishers.

Horowitz, Tamara (1996). “Philosophical Intuitions and Psychological Theories.” Unpublished manuscript. Presented at “Rethinking Intuition: The Psychology of Intuitions and their Role in Philosophical Inquiry.” University of Notre Dame, April 1996.

Humphreys, Paul (1993). “Seven Theses on Thought Experiments.” In John Earman *et al* editors, *Philosophical Problems of the Intemal and External Worlds: Essays on the Philosophy of Adolf Grünbaum*. Pittsburgh: University of Pittsburgh Press: 205-227.

Hurley, Susan (1989). *Natural Reasons*. New York: Oxford.

Irvine, Andrew D. (1991). “On the Nature of Thought Experiments in Scientific Reasoning” in Horowitz and Massey (1991): 149-165.

Janis, Allen I. (1991). “Can Thought Experiments Fail?” In Horowitz and Massey (1991): 113-118.

Johnston, Mark (1987a). “Human Beings.” *Journal of Philosophy* Vol. LXXXIV, No. 2 (February 1987): 59-83.

Johnston, Mark (1987b). Review of Shoemaker and Swinburne (1984). *Philosophical Review*, XCVI, No. 1 (January 1987): 123-128.

Johnston, Mark (1989a). "Fission and the Facts.” In James E. Tomberlin, ed. *Philosophical Perspectives, 3* Atascadero, CA: Ridgeview Publishing Co.: 369-397.

Johnston, Mark (1989b). “Relativism and the Self.” In Michael Krausz, ed. *Relativism: Interpretation and Confontation*. Notre Dame: University of Notre Dame Press: 441-472.

Johnston, Mark (1992a). “Constitution is Not Identity.” *Mind*, Vol. 101(January 1992): 89-105.

Johnston, Mark (1992b). “Reasons and Reductionism.” *Philosophical Review*, Vol. 101, No. 3 (July 1992): 589-618.

Johnston, Mark (1992c). “Human Concerns without Superlative Selves.” Unpublished manuscript. Forthcoming in Dancy (forthcoming).

Kant, Immanuel (1785/1964). *Groundwork of the Metaphysic of Morals*. Trans. H.J. Paton. New York: Harper Torchbooks.

Kant, Immanuel (1787/1929). *Critique of Pure Reason*. Trans. Norman Kemp Smith. London: Macmillan.

Keil, Frank C. (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: Bradford/MIT.

King, Peter (1991). "Mediaeval Thought Experiments: The Metamethodology of Mediaeval Science." In Horowitz and Massey (1991): 43-64.

Kornblith, Hilary (1996). "The Role of Intuition in Philosophical Inquiry: An Account with No Unnatural Ingredients." Unpublished manuscript. Presented at "Rethinking Intuition: The Psychology of Intuitions and their Role in Philosophical Inquiry." University of Notre Dame, April 1996.

Korsgaard, Christine (1989). "Personal Identity and the Unity of Agency: A Kantian Response to Parfit." *Philosophy and Public Affairs* Vol. 18, No. 2 (Spring 1989): 101-132.

Koyré, Alexandre (1968). *Metaphysics and Measurement*. London: Chapman and Hall.

Kuhn, Thomas (1964). "A Function for Thought Experiments." Reprinted with new pagination in Kuhn (1977).

Kuhn, Thomas (1970). *The Structure of Scientific Revolutions* (second edition). Chicago: University of Chicago Press.

Kuhn, Thomas (1977). *The Essential Tension*. Chicago: University of Chicago Press.

Laymon, Ronald (1991). "Thought Experiments of Stevin, Mach and Gouy: Thought Experiments as Ideal Limits and as Semantic Domains." In Horowitz and Massey (1991): 167-192.

Leibniz, G.W. (1765/1981). *New Essays on Human Understanding*. Translated and edited by Peter Remnant and Jonathan Bennett. Cambridge: Cambridge University Press.

Lennox, James G. (1991). "Darwinian Thought Experiments: A Function for Just-So Stories." In Horowitz and Massey (1991): 223-245.

Lewis, David (1976). "Survival and Identity." In Rorty (1976): 17-40. Reprinted with new pagination in Lewis (1983a): 55-72.

Lewis, David (1983a). *Philosophical Papers: Volume I*. New York and Oxford: Oxford University Press.

Lewis, David (1983b). “Postscripts to ‘Survival and Identity.’” In Lewis (1983a): 73-77.

Locke, Don (1961). “Strawson’s Auditory Universe.” *Philosophical Review* Vol. LXX, No. 4 (October 1961): 518-532.

Locke, John (1710/1975). *An Essay Concerning Human Understanding* (Edited with a forward by Peter H. Nidditch). Oxford: Clarendon Press.

Luce, R. Duncan and Howard Raiffa (1957). *Games and Decisions*. New York: Dover.

Mach, Emst (1933/1960). *The Science of Mechanics* (ninth edition). Trans. Thomas McCormack.

Mach, Emst (1926/1976). *Knowledge and Error* (fifth edition). Trans. Thomas McCorrnack and Paul Foulkes.

Mackie, J. L. (1976). *Problems from Locke*. Oxford: Clarendon Press.

Madell, Geoffrey (1991). “Personal Identity and the Idea of a Human Being.” In Cockburn (1991): 127-142.

Markman, Ellen M. (1989). *Categorization and Naming in Children: Problems of Induction*. Cambridge: MIT/Bradford.

Martin, Raymond (1993). “Real Values: Why the Wilkes-Donagan Prohibition is Mistaken.” *Metaphilosophy* Vol. 24, No. 4 (October 1993): 400-406.

Massey, Barbara (1991). “Do All Rational Folk Reason as We Do? Frege’s Thought Experiment Reconsidered.” In Horowitz and Massey (1991): 99-110.

Massey, Gerald (1991). “Backdoor Analyticity.” In Horowitz and Massey (1991): 285-296.

Massey, Gerald (1995). “*Gedankeneperimente*: Where Science and Philosophy Meet.” Unpublished manuscript.

McDowell, John (1994). *Mind and World*. Cambridge: Harvard University Press.

McDowell, John (forthcoming). “Reductionism and the First Person.” Unpublished manuscript, forthcoming in Dancy (forthcoming).

McMullin, Ernan (1978). “The Conception of Science in Galileo's Work.” In Butts and Pitt (1978): 209-257.

Miller, Fred D. and Nicholas Smith, eds. (1989). *Thought Probes: Philosophy Through Science Fiction Literature*. Englewood Cliffs, NJ: Prentice Hall.

Mills, Eugene (1993). “Dividing without Reducing: Bodily Fission and Personal Identity.” *Mind* Vol 102, No. 405 (January 1993): 37-51.

Mohanty, J. N. “The Method of Imaginative Variation in Phenomenology.” In Horowitz and Massey (1991): 261-272.

Molnar, G. (1969). “Kneale’s Argument Revisited.” *Philosophical Review*. Reprinted in T.L. Beauchamp ed. *Philosophical Problems of Causation*. Belmont CA: Dickenson, 1974.

Moran, Richard (1988). “Making Up Your Mind: Self-Interpretation and Self-Constitution.” *Ratio* (New Series), Vol I, 135-151.

Munitz, Milton K., ed. (1971). *Identity and Individuation*. New York: New York University Press.

Myers, C. Mason (1986). “Analytical Thought Experiments.” *Metaphilosophy* Vol. 17, Nos. 2 and 3 (April/June 1986): 109-118.

Nagel, Thomas (1971). “Brain Bisection and the Unity of Consciousness.” *Synthêse* Vol. 22. Reprinted with new pagination in Perry (1975a): 227-245.

Nersessian, Nancy (1984). *Faraday to Einstein: Constructing Meaning in Scientific Theories*.

Nersessian, Nancy (1992). "How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science" In Ronald Giere, editor *Minnesota Studies in the Philosophy of Science: Cognitive Models of Science,* Minneapolis: University of Minnesota Press, Vol. XV: 3-44.

Nersessian, Nancy (1993). "In the Theoretician's Laboratory: Thought Experiment as Mental Modeling.” *Proceedings of the Philosophy of Science Association* Vol. 2: 291-301.

Noonan, Harol (1980). *Objects and Identity*. The Hague: Martinus Nijhoff Publishers.

Noonan, Harold (1989). *Personal Identity*. London and New York: Routledge.

Norton, John (1991). “Thought Experiments in Einstein's Work.” In Horowitz and Massey (1991): 129-148.

Nozick, Robert (1974). *Anarchy, State, and Utopia*. New York: Basic Books.

Nozick, Robert (1981). *Philosophical Explanations*. Cambridge, MA: Belknap Press.

Nozick, Robert (1993). *The Nature of Rationality*. Princeton: Princeton University Press.

Oderberg, David S. (1989). “Johnston on Human Beings.” *Journal of Philosophy* Vol. LXXXVI, No. 3 (1989): 137-141.

Oderberg, David S. (1993). *The Metaphysics of Identity over Time*. New York: St. Martin’s Press.

O’Neill, Onora (1986). “The Power of Example.” *Philosophy* 61: 5-29.

Parfit, Derek (1971). “Personal Identity.” *Philosophical Review* Vol. 80, No. 1 (January, 1971). Reprinted with new pagination in Perry (1975a): 199-223.

Parfit, Derek (1986). “Comments.” *Ethics* 96 (July 1986): 832-872.

Parfit, Derek (1987). *Reasons and Persons*. Oxford: Clarendon Press.

Parfit, Derek (1992). “The Unimportance of Identity.” Unpublished manuscript.

Parfit, Derek (1994). “Personal Identity.” Unpublished manuscript.

Peirce, Charles Sanders (1867-1893/1992). *The Essential Peirce: Selected Philosophical Writings*. Vol. I. Ed. Nathan Houser and Christian Kloesel. Bloomington and Indianapolis: Indiana University Press.

Perry, John (1972). “Can the Self Divide?” *Journal of Philosophy* LXIX:16 (September 7, 1972): 463-488.

Perry, John, ed. (1975a). *Personal Identity*. Berkeley and Los Angeles: University of California Press.

Perry, John (1975b). “The Problem of Personal Identity.” In Perry (1975a): 3-30.

Perry, John (1976). “The Importance of Being Identical.” In Rorty 1976: 67-90.

Persson, Ingmar (1993). “Critical Study of Peter van Inwagen’s *Material Beings*.” *Noûs* 27:4: 512-518.

Piaget, Jean (1946/1970). *The Child's Conception of Movement and Speed.* Trans. G.E.T. Holloway and M.J. Mackenzie. NY: Basic Books.

Plato (c. 380BC/1974). *Republic*. Trans. G.M.A. Grube revised C.D.C. Reeve. Indianapolis: Hackett.

Poincaré, Henri (1905/1952). *Science and Hypothesis*. New York: Dover.

Popper, Karl (1959/1992). *The Logic of Scientific Discovery*. Trans. Karl Popper. New York: Routledge.

Putnam, Hilary (1975a). "The Meaning of Meaning." In Putnam (1975b): 215-271.

Putnam, Hilary (1975b). *Mind, Language and Reality: Philosophical Papers, Volume 2*. Cambridge: Cambridge University Press.

Putnam, Hilary (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.

Putnam, Hilary (1983a). "Analyticity and apriority: beyond Wittgenstein and Quine." In Putnam (1983b): 115-138.

Putnam, Hilary (1983b). *Realism and Reason: Philosophical Papers, Volume 3*. Cambridge: Cambridge University Press.

Putnam, Hilary (1983c). "There is at least one *a priori* truth." In Putnam (1983b): 98-114.

Putnam, Hilary (1990a). "Is Water Necessarily H_2O?" In Putnam (1990b): 54-79.

Putnam, Hilary (1990b). *Realism with a Human Face*. Cambridge: Harvard University Press.

Putnam, Hilary (1994a). "Rethinking Mathematical Necessity." In Putnam (1994b): 245-263.

Putnam, Hilary (1994b). *Words and Life*. Cambridge: Harvard University Press.

Quine, W.V. (1972). Review of *Identity and Individuation*. *Journal of Philosophy* Vol. LXIX, No. 16 (September 7, 1972): 488-497.

Quinton, Anthony (1962/1975). "The Soul." *Journal of Philosophy* Vol. LIX, No. 15 (July 1962). Reprinted with new pagination in Perry (1975a): 53-72.

Rea, Michael C. (1995). “The Problem of Material Constitution.” *Philosophical Review* Vol. CIV, No. 4 (October 1995): 525-552.

Rescher, Nicholas (1991). “Thought Experiment in Pre-Socratic Philosophy.” In Horowitz and Massey (1991): 31-41.

Robinson, John (1988). “Personal Identity and Survival.” *Journal of Philosophy* Vol. LXXXV, No. 6 (June 1988): 319-328.

Rorty, Amélie Oksenberg, ed. (1976). *The Identities of Persons*. Berkeley: University of California Press.

Rosenberg, Jay F. (1993). “Comments on Peter van Inwagen’s *Material Beings*.” *Philosophy and Phenomenological Research*, Vol. LIII, No. 3 (September 1993): 701-708.

Rovane, Carol (1990). “Branching Self-Consciousness.” *Philosophical Review* Vol. XCIX, No. 3 (July 1990): 355-395.

Rovane, Carol (1993). “Self-Reference: The Radicalization of Locke.” *Journal of Philosophy* Vol. XC, No. 2 (February 1993): 73-97.

Rovane, Carol (1994). “Critical Notice of Peter Unger’s *Identity, Consciousness and Value*.” *Canadian Journal of Philosophy* 24:1 (March 1994): 119-133.

Rubiner , Michael (1996). “Endpaper: Greetings.” *New York Times Magazine* (January 14, 1996): 60.

Scaltsas, Theodore (1980). “The Ship of Theseus.” *Analysis* 40:3 (June 1980): 152-157.

Schechtman, Marya (1990). “Personhood and Personal Identity.” *Journal of Philosophy* Vol. LXXVII, No. 2 (February 1990): 71-92.

Schechtman, Marya (1994). “The Same and the Same: Two Views of Psychological Continuity.” *American Philosophical Quarterly* Vol. 31, No. 3 (July 1994): 199-212.

Sedley, David (1982). “The Stoic Criterion of Identity.” *Phronesis* Vol. 22: 255-275.

Shafir, Eldar (1996). “Philosophical Intuitions and Cognitive Mechanisms.” Unpublished manuscript. Presented at “Rethinking Intuition: The Psychology of Intuitions and their Role in Philosophical Inquiry.” University of Notre Dame, April 1996.

Smart, Brian (1972). “How to Reidentify the Ship of Theseus.” *Analysis* Vol. 32: 145-148.

Smart, Brian (1973). “The Ship of Theseus, the Parthenon, and Disassembled Objects.” *Analysis* Vol. 34: 24-27.

Shepard, Roger (1994). “Mind and World.” William James Lectures, Harvard University, Fall 1994. Unpublished manuscript.

Shoemaker, Sydney (1959/1975). “Personal Identity and Memory.” *Journal of Philosophy* Vol. LVI, No. 22 (October 22, 1959). Reprinted with new pagination in Perry (1975a): 119-134.

Shoemaker, Sydney (1963). *Self-Knowledge and Self-Identity*. Ithaca: Cornell University Press.

Shoemaker, Sydney (1970/1984). “Persons and Their Pasts.” *American Philosophical Quarterly* Vol 7, No. 4 (October 1970): 269-285. Reprinted with new pagination in Shoemaker (1984): 19-48.

Shoemaker, Sydney (1971). “Wiggins on Identity.” In Munitz (1971): 103-118.

Shoemaker, Sydney (1984). *Identity, Cause and Mind*. Cambridge: Cambridge University Press.

Shoemaker, Sydney (1992). “Unger’s Psychological Continuity Theory.” *Philosophy and Phenomenological Research* Vol. LII, No. 1 (March 1992): 139-143.

Shoemaker, Sydney (1994). “The First-Person Perspective.” *Proceedings of the APA* Vol. 68, No. 2 (November 1994): 7-22.

Shoemaker, Sydney (1995). “Self and Substance.” Unpublished manuscript.

Shoemaker, Sydney and Richard Swinburne (1984). *Personal Identity: Great Debates in Philosophy*. Oxford: Basil Blackwell.

Slote, Michael (1991). Review of Sorensen (1992). *Noûs*: 328-333.

Smart, Brian (1972). “How to Reidentify the Ship of Theseus.” *Analysis* 32:5 (April 1972): 145-148.

Smart, Brian (1973). “The Ship of Theseus, the Parthenon, and Disassembled Objects.” *Analysis* 34:1 (October 1973): 24-27.

Snowdon, P.F. (1991). “Personal Identity and Brain Transplants.” In Cockburn 1991:109-126.

Sorensen, Roy (1992a). *Thought Experiments*. New York and Oxford: Oxford University Press.

Sorensen, Roy (1992b). “Thought Experiments and the Epistemology of Laws.” *Canadian Journal of Philosophy* Vol. 22, No. 1 (March 1992), 15-44.

Sosa, Ernest (1990). “Surviving Matters.” *Noûs* 24: 305-330.

Sosa, Ernest (1996). “Minimal Intuition.” Unpublished manuscript. Presented at “Rethinking Intuition: The Psychology of Intuitions and their Role in Philosophical Inquiry.” University of Notre Dame, April 1996.

Stevin, Simon (1955). *The Principal Works of Simon Stevin*. Vol 1: General Introduction and Mechanics. Ed. E.J. Dijksterhuis. Amsterdam: C. W. Swets and Zeitlinger.

Strawson, P. F. (1959). *Individuals: An Essay in Descriptive Metaphysics*. London: Methuen.

Strawson, P. F. (1992). “Comments on Some Aspects of Peter Unger’s *Identity Consciousness and Value*.” *Philosophy and Phenomenological Research* Vol. LII, No. 1 (March 1992): 145-148.

Sunstein, Cass (1993). “On Analogical Reasoning.” *Harvard Law Review* Vol. 106: 741-791.

Sunstein, Cass (1994). Tanner Lectures, Harvard University. Unpublished.

Swinburne, Ricahrd (1992). Discussion of Peter Unger’s *Identity Consciousness and Value*.” *Philosophy and Phenomenological Research* Vol. LII, No. 1 (March 1992): 149-152.

Taylor, Charles (1977). “Self-Interpreting Animals.” Reprinted in Taylor 1985a, 45-76.

Taylor, Charles (1981). “The Concept of a Person.” Reprinted in Taylor 1985a, 97-114.

Taylor, Charles (1985a). *Human Agency and Language: Philosophical Papers, volume 1*. Cambridge: Cambridge University Press.

Thomason, Sarah (1991). "Thought Experiments in Linguistics." In Horowitz and Massey (1991): 247-257.

Thomson, Judith Jarvis (1971). "A Defense of Abortion." *Philosophy and Public Affairs* 1:1 (Fall 1971): 47-66.

Thomson, Judith Jarvis (1986). *Rights, Restitution and Risk*. Cambridge: Harvard University Press.

Thomson, Judith Jarvis (1990). *The Realm of Rights*. Cambridge: Harvard University Press.

Tidman, Paul (1994). "Conceivability as a Test for Possibility." *American Philosophical Quarterly* 31:4 (October 1994): 297-309.

Tye, Michael (1992). "Review of Peter van Inwagen's *Material Beings.*" *Philosophical Review* Vol. 101, No. 4 (October 1992): 881-884.

Unger, Peter (1990). *Identity, Consciousness and Value*. New York and Oxford: Oxford University Press

Unger, Peter (1992). "Précis of *Identity Consciousness and Value*" and "Reply to Reviewers." *Philosophy and Phenomenological Research* Vol. LII, No. 1 (March 1992): 133-137, 159-176.

Van Inwagen, Peter (1990). *Material Beings*. Ithaca: Cornell University Press.

Van Inwagen, Peter (1993a). "Critical Study of Peter Unger's *Identity Consciousness and Value.*" *Noûs* 27:2: 373-379

Van Inwagen, Peter (1993b). "Précis of *Material Beings*" and "Reply to Reviewers" *Philosophy and Phenomenological Research,* Vol. LIII, No. 3 (September 1993): 683-691, 709-719.

Walton, Kendall (1990). *Mimesis as Make-Believe*. Cambridge: Harvard University Press.

White, Steven. "The Desire to Survive." *Philosophy and Phenomenological Research* Vol. LII, No. 1 (March 1992): 153-158.

Whiting, Jennifer (1986). "Friends and Future Selves." *Philosophical Review,* Vol. XCV, No. 4 (October 1986): 547-580.

Whiting, Jennifer (1991). "Impersonal Friends." *The Monist* 74: 3-29.

Wiggins, David (1980). *Sameness and Substance*. Cambridge, MA: Harvard University Press.

Wilkes, Kathleen V. (1988). *Real People: Personal Identity without Thought Experiments*. Oxford: Clarendon Press.

Williams, Bernard (1956/1973). “Personal Identity and Individuation.” Reprinted (with new pagination) in Williams (1973).

Williams, Bernard (1970a/1973). “Are Persons Bodies?” Reprinted (with new pagination) in Williams 1973.

Williams, Bernard (1970b/1973). “The Self and the Future.” Reprinted (with new pagination) in Williams 1973.

Williams, Bernard (1973). *Problems of the Self*. Cambridge: Cambridge University Press.

Wiser, Marianne and Susan Carey (1983). “When Heat and Temperature Were One.” In Gentner and Stevens (1983): 267-297.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. Oxford: Basil Blackwell.

Wolf, Susan (1986). “Self-Interest and Interest in Selves.” *Ethics* 96 (July 1986): 704-720.

Yablo, Stephen (1993). “Is Conceivability a Guide to Possibility?” *Philosophy and Phenomenological Research* LIII:1 (March 1993): 1-42.