Personal Identity without Persons

Jens David Ohlin

Submitted in partial fulfillment of the
Requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2002

ABSTRACT

Personal Identity without Persons

Jens David Ohlin

The project takes as its starting point our conflicting intuitions about personal identity exposed by Bernard Williams' thought experiment involving the switching of bodies in "The Self and the Future." The conflicted intuitions are identified as animalist and psychologist and correspond roughly with the two major approaches to personal identity. The traditional strategy to resolve the conflict—thought experiments—is critically examined and the project concludes that proper thought experiments will reveal the conflict but are unlikely to resolve it. A new reading of the conflict is proposed. The concept of the person is a cluster concept with distinct components: biological human beings, rational agency and psychological continuity, where the latter is construed as the temporal analog of phenomenological unity at a time. The project then suggests that moral theory is best pursued not by a naturalist conception of persons that unites the components of the cluster, but by a novel conception that separates them. This can be accomplished best by *eliminating* the concept of the person altogether. Objections that the concept of the person is ineliminable are considered and rejected, as are objections that the nature of conflict is not reason enough to abandon the concept. Personhood's centrality for value theory is questioned and the eliminativist strategy is defended on the grounds that responsibility, self-concern and moral rights are best analyzed with the component concepts instead of the cluster concept. Eliminativism is therefore *preferable* to competing accounts of personal identity because the former is better suited for moral theory. Among the advantages are the ability to: attribute responsibility to group agents without calling them persons, make sense of our conflicting demonstrations of self-concern without taking them as evidence for conflicting theories of personal identity, and attribute moral rights to entities who fail to meet traditional criteria for personhood but who are nonetheless entitled to moral respect.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

My interest in personal identity began when I took a seminar on persons offered by Carol Rovane at Columbia University in the spring of 1999 and later a seminar by Derek Parfit at New York University in 2001. Over three years Carol Rovane supervised my doctoral research on the subject. Akeel Bilgrami and Patricia Kitcher offered valuable criticisms of my work and served as members of my dissertation committee. Jeremy Waldron and Wayne Proudfoot sat as outside readers on my committee. I presented a paper summarizing my views on personal identity to the department's dissertation forum and received several suggestions from fellow graduate students. Throughout, Nancy Butcher was an invaluable sounding board for the plausibility of my philosophical positions and the clarity of their formulation.

This project is yet another addition to the growing body of literature about personal identity. In many ways, though, the journey charted in these pages represents a departure from past approaches. A major assumption underlying the investigation of personal identity will be questioned. That assumption is nothing less than the assumed importance of the concept of the person itself.

In that sense this project is an attempt at reformulating the terms of the debate, of switching the ground of inquiry to more hospitable quarters. As the argument unfolds, it will become clear that the classic debate in personal identity—between animalism and psychologism—is irresolvable in its present form. I will suggest that a coherent and convincing account of personal identity will not be found just so long as the inquiry unfolds around the concept of the person. I will depart from previous accounts of personal identity by arguing that the concept of the person is a cluster concept composed of distinct components: biological human beings, psychological continuity and rational agency. We have been unable to resolve the animalist-psychologist debate about personal identity precisely because the concept under investigation is a cluster concept housing distinct and sometimes competing components.

That being said, the prescription will inevitably turn to elimination of the concept of the person. It has become fashionable of late to reduce or eliminate concepts and phenomena or to argue against such moves because they are, for what ever reason, irreducible or ineliminable. These are familiar moves in analytic philosophy. But the account traced in these pages is not fashionable at all; indeed, it represents a break from the mode of investigation in personal identity going back at least to Locke. This investigation has been built around the assumption that the concept of the person is significant (a forensic concept, as Locke said) and that something essential would be lost if we could not offer a complete account of personal identity. More than anything it is this idea which is under attack in this project.

The project is divided into five chapters. The first chapter analyzes our conflicting intuitions about personal identity and argues that they stem from personhood's status as a cluster concept. The second chapter proposes elimination of the concept as a possible solution and responds to three categories of objection to the eliminativist strategy. The rest of the project investigates whether elimination is possible given that the concept of the person is central in many areas of our discourse. In that vein, chapter three through five investigate whether the concept of the person is essential for understanding responsibility, self-concern and rights respectively. These chapters also investigate whether eliminating the concept of the person might improve our understanding of these phenomena.

An important caveat about terminology: Throughout the project the term 'psychology' and 'psychological continuity' are used frequently. Psychology here is meant to invoke its phenomenological counterpart, in the sense in which psychological continuity might be considered the temporal analog to the unity of consciousness at any given moment. This phenomenological notion is crucial to a person's understanding of self as a conscious being persisting through time. That is, not only am I aware of myself at a time, but I am aware of myself over time. This allows me to think of myself as having a conscious point of view on the world, a unified *phenomenological* point of view. It is this sense of the expression 'psychological continuity' that was inaugurated by Locke during his construction of personal identity thought experiments concerned with consciousness and memory.

In many areas of contemporary philosophy 'psychology' is not construed so narrowly. These terms could be misleading because they are used in a variety of ways within philosophy and they are being used in a very restricted sense in this project. Understanding the sense of my terms 'psychology' and 'psychological continuity' is especially important when I advocate eliminating the concept of the person in favor of its components. It is impossible to get an accurate handle on the proposal unless I am clear about what is meant by psychological continuity—one of the components of the cluster. Later in this project I will argue that personhood is a cluster concept and that we should separate out the

component of psychological continuity from the other components of biological human beings and rational agency. In making this claim I am arguing that psychological continuity *qua* phenomenology ought to be separated out from the other two components. But someone might object that this only makes sense within a very limited and narrow understanding of psychology as an exclusively phenomenological concept, which some might reject.

There is a long list of possible meanings for the concept of psychology and psychological continuity and there is an entire department of the academy, separate from philosophy, devoted to its study. Even within that discipline, as indeed within the philosophy of mind, the sense of the term 'psychology' changes. From theoretical to experimental psychology, from abnormal and developmental psychology to cognitive and psycho-pharmaceutical psychology, the term is not held constant. Not only does the method of study change but what is understood to be the object of study changes. Indeed, what a behaviorist means by the psychology of an individual is not necessarily what a Freudian means by the psychology of her patient.

Many will resist my attempt to separate out the concept of psychological continuity from the other components of biological human beings and rational agency. In some sense, all of the components of the cluster can be seen as psychological in some way. Biological human beings are the physical basis for the emergence of a conscious point of view and agents have a distinctive rational psychology. According to the psychologist, all of these components come together and are part of the psychology of persons, which is the theoretical rubric under which these psychological elements are united. This is all part of a unified psychological theory of persons.

Furthermore, a naturalist about persons might think of psychology as being necessarily connected with human beings and agency, since they are all part of the normal application of biological human beings functioning properly as rational agents with a particular psychology—and that's exactly what it means to be a person, according to the

naturalist. For the naturalist, then, psychology ought to be broadly construed in this sense as applying to all of the components and perhaps providing a unified rubric for a theory of persons. Since in most cases all of these components go together, it will seem wrong to the naturalist to limit psychology to its phenomenological sense for eventual separation from the other components. A psychological theory of persons ought to be built around what the psychological naturalist refers to as the normal cases of personhood.

My answer is that I am perfectly willing to concede that human beings ought to be construed naturalistically and that a theory of human beings ought to be constructed around the "normal" paradigm of human beings. Furthermore, I am also willing to concede that all of the components of the cluster concept of the person have a biological basis. Psychological continuity can only be realized in some physical medium like the human brain and rational agency emerges from the same biological foundation. Large-scale disruptions to psychological continuity and rational agency will no doubt be the result of physical damage to a biological system. Comas, fugue states, and even multiple personalities—should they turn out to exist—can be described in explicitly physical language. But there is no reason to deny any of this—only a dualist would do so. What *is* being denied is that personhood is simply a synonym for human beings; its history demonstrates that, as Locked recognized, it is a ***moral*** notion. Personhood is a term shared by many areas of inquiry but most central for one—value theory. And that being the case, value theory has its own distinctive needs when it comes to the concept of the person.

This leads us to the central thesis of the dissertation. All of the smaller arguments contained with the project can be seen as building towards one central and revisionary conclusion: that it is not best for moral theory to construct a theory of persons around the so-called "normal" paradigm of human beings where biological, psychological and agency components coincide. It may turn out that there are many cases where it is best to keep them separate: amnesia, comas, multiple personalities, fugue states, group agents, etc.... Not only does this conclusion represent a direct departure from the standard treatments in the

literature, but it is also noteworthy for its consequences. Once the conclusion is accepted that a naturalist theory of persons is ill-suited for the demands of moral theory, the direction of the debate will take a genuinely novel turn. We will need to start asking which concepts are best suited for the demands of moral theory. I will argue that the demands are best served not by the concept of the person at all, but rather by its component concepts. That is the heart of the project's argument. It is the nature of this argument that it can only be defended by working through the demands of moral theory and how the component concepts might better meet those demands. This is the slow work of chapters three through five.

Take just one example. If multiple personalities turn out to exist (and this is an empirical question on which I must remain agnostic), our biological notion of human beings would be a poor foundation for moral theory. Multiple personalities present themselves as distinct agents with distinct deliberative structures. Since we are intuitively committed to the idea that agents are responsible for their actions because of that distinct deliberative structure, we would be inclined to hold each personality individually responsible for his or her actions. A naturalist theory of persons, built around the supposedly paradigmatic case where agents come one to a body, would be a poor foundation for moral theory. The naturalist might respond that this case is an exception that in no way entitles us to abandon the naturalist conception of persons in the normal cases. But what if abandoning the naturalist conception had no negative consequences for the normal cases? Indeed, what if the consequences were positive? This question will be pursued closely in chapter three.

This brings us back to the issue of restricting our understanding of psychology to its phenomenological interpretation. There are essentially two reasons for this decision. The first relevant reason to restrict psychology to its phenomenological element is precedent. The history of the personal identity literature, from Parfit in the present to Locke in the past, uses the term psychology in this restricted way. This goes all the way back to Locke, whose psychological account of personal identity centered around consciousness and memory. After all, it was the consciousness of the prince and the cobbler that were switched.

More importantly, though, we can envision a payoff received for this decision in the second half of the project, where moral theory is at issue. It is the phenomenological element of psychology that is most directly morally significant in the case of moral theories such as utilitarianism (e.g. in terms of the capacity to feel pain). And treating our restricted notion of psychological continuity independently of the concept of the person will be shown to be advantageous for moral theory. We are then left with a balancing test to judge the utility of eliminativism relative to a naturalist theory of persons and other accounts of personal identity. I argue that eliminativism is preferable to a naturalistic theory of persons because the former has better consequences for moral theory, which is the central reason for having the concept of personhood in the first place. I will demonstrate this by the slow accumulation of evidence in chapters three through five. In short, the evidence is improved accounts of responsibility, self-concern, and rights. The advantages might be summarized as follows: (i) we can attribute responsibility to atypical agents (like group or multiple agents) without having to call them persons and in doing so violating our biological or psychological intuitions about personal identity; (ii) we can make sense of our conflicting demonstrations of self-concern—and see them as rationally justified—without taking them as evidence for conflicting theories of personal identity; and (iii) we can recognize that moral rights track the components of the cluster instead of personhood itself, thus offering a solution to the moral status of marginal entities who lack one of the components. Taken individually each of these advantages may seem insufficient to justify the revisionary proposal of elimination. In other words, it might not seem worth the violence to our common sense to eliminate a central concept from our discourse. But when weighed together, the totality of evidence demonstrates that eliminativism is the best available option for moral theory. And that speaks to the proposal's utility, which is the standard of proof here.

Unfortunately it would be impossible to offer here complete accounts of responsibility, self-concern and rights. And it is even more daunting to offer a total theory of biological human beings, psychological continuity and rational agency—the component

concepts of the cluster. Various details must necessarily remain unsaid. But enough detail is given to support the eliminativist proposal. Far from being indispensable, the concept of the person might turn out to be just one concept among many, neither ordained with divine significance nor born with platonic purity and its existence in our conceptual apparatus is indeed subject to justification, a task that has gone unperformed for far too long. Once that justification is attempted its place in our conceptual apparatus will, I suspect, be questioned vigorously and a new avenue for pursuing personal identity, with different concepts at our disposal, will be revealed.

I will argue in this chapter that a series of thought experiments by Williams reveals our conflicting intuitions about personal identity and that there are no positive reasons to accept or reject either of the intuitions. Invoking Quine's skepticism about thought experiments I will suggest that this methodology is unlikely to resolve these conflicting intuitions in the future. This is evidence that our concept of the person is a cluster concept that tracks a variety of component concepts such as biological continuity, psychological continuity and rational agency. Finally, I will position this claim in relation to Parfit's well known views about personal identity.

This chapter takes as its starting point Locke's thought experiment about the prince and the cobbler, as updated by Shoemaker and Williams.[1] The striking aspect of Williams' version of the thought experiment was that it exposed our conflicting intuitions—intuitions which Locke hastily claimed were definitive. While there is no shortage of philosophers—and lay respondents—who have Lockean intuitions, Williams demonstrated that it is possible to elicit strong anti-Lockean intuitions. These results should not surprise us, for it is hardly surprising that we should have conflicting intuitions about what it means to be a person. We are committed to the idea that persons exhibit psychological continuity, we are committed to the idea that persons exhibit biological continuity, and we are committed to the idea of persons as rational agents.[2]

The problem with Williams' thought experiment is that it exposes our conflicting intuitions while providing little ground for resolution. While Williams forcibly demonstrated these conflicting intuitions, his solution to the conflict was to support one side over the other for positive reasons. Unfortunately those positive reasons were not adequate

---

[1] See John Locke, *An Essay Concerning Human Understanding*, Book II, Ch. XXVII; Sydney Shoemaker, "Persons and their Pasts" in *American Philosophical Quarterly* 7 (1970); and Bernard Williams, "The Self and the Future" in *Problems of the Self* (New York: Cambridge University Press, 1973).

[2] In taking these conflicting intuitions as my starting point I am following—in obvious addition to the work done by Williams—the diagnosis of the problem offered by Carol Rovane in *The Bounds of Agency: An Essay in Revisionary Metaphysics* (Princeton: Princeton University Press, 1998), see especially chapter two. However, unlike Rovane, who takes our conflicting intuitions as a warrant for a revisionary account of personhood, I will take the conflict as evidence that personhood is a cluster concept and as a justification for its eventual elimination.

to resolve the dispute, mainly because Williams argued that it was the animalist approach that was least vulnerable to an objection from branching. As will be demonstrated in this chapter, though, the animalist approach is *not* immune to a branching objection. Indeed the branching objection poses a major problem for almost any account of personal identity—psychological and animalist alike. (Chisholm, Wiggins, Shoemaker and Nozick present accounts that portend to deal with the branching problem, though as we shall see, they do so at considerable cost.) Consequently, the branching objection is irrelevant for the purpose of deciding between the competing animalist and psychological accounts.

Indeed, part of the frustration is with the methodology itself. Quine's objection is relevant and must be dealt with here. Quine's idea was that words have no meaning beyond which our current needs have invested them with.[3] Thought experiments that border on science fiction attempt to find a greater degree of precision in our everyday concepts than actually exists there. Others have offered criticisms of *Gedankenexperimente* methodology in the personal identity literature. Indeed, if any of these criticisms are correct and the methodology is unsound, that would provide full-blown proof that no thought experiment could resolve our conflicting intuitions. This chapter will stop short of providing such proof, however, although that will do no damage to the argument. It is not necessary to prove that thought experiments, by their very nature, will never provide a resolution. It is sufficient to note that no resolution is present, no resolution is forthcoming, and that ought to give us pause to search for a new strategy—in this case elimination of the concept of the person. The task will then be to show that the alleged costs of this strategy have been greatly exaggerated and that there are, in fact, advantages. Chapter two will be devoted exclusively to explaining the eliminativist strategy. The remaining chapters will be devoted to arguing for the strategy's plausibility.

---

[3] W.V. Quine, Review of *Identity and Individuation* (edited by Milton K. Munitz) in *The Journal of Philosophy* 69 (1972): 489.

Our conflicting intuitions about personhood are evidence that 'person' is a cluster concept—not a discrete concept—composed of several sub-concepts, among them psychological continuity, biological continuity, and rational agency. As we shall see, the fact that personhood is a cluster concept makes it ripe for elimination in favor of its components. Eliminating the concept of the person is a similar strategy to Parfit's reductionism. Parfit noted that questions about personal identity are sometimes empty and he distinguished two ways that a question can be empty. Parfit took this to be evidence that facts about personal identity could be reduced to certain other facts. This chapter will note the similarities between Parfit's reductionism and my claim that the concept of the person is a cluster concept. Although my claim entails Parfit's claim—that some questions about persons and personal identity are empty—the reverse is certainly not true, i.e. his claim certainly does not entail mine. My claim that 'person' is a cluster concept goes far beyond Parfit's reductionism and suggests the eliminativist strategy. The comparison is relevant because it shows that readers already sympathetic to Parfit's reductionism have a shorter jump to embrace my claim that 'person' is a cluster concept ripe for elimination.

To explain the distinction between eliminativism and reductionism we will consider Parfit's example of nations. Parfit introduces this case to explain reductionism. He says that all facts about nations can be reduced to certain other facts about its members and their activities. This is what it means to be a reductionist about nations. But you could be an eliminativist about nations. One might claim that 'nation' is a cluster concept that tracks various components such as country, people, state, geographic area, etc. Our intuitions might conflict about what it means to be a nation because the components do not always go together (indeed the whole point of nationalism is to get these things to line up when they don't), so one could argue that we ought to consider eliminating our concept of the nation and replacing it with its components. This kind of elimination is also a common strategy in medicine when syndromes that track a cluster of symptoms are eliminated in favor of individual diseases with discrete causes.

This is the same general story that I will tell about the concept of the person—that it is a cluster-concept because one term is loosely attached to a whole cluster of related ideas and concepts; we are unable to understand what it means to be a 'person' because the sub-concepts often conflict with each other. In the final section of this chapter I will explore briefly how this cluster of related ideas came to be housed under one term. As it turns out, the history of personhood is the history of a concept that was designed to house competing ideas under one umbrella—i.e. it has *always* been a cluster concept. This was the case in Roman Law (when the Latin word *persona* came into this context) and this was the case in dualist conceptions of the person that included an earthly body and an eternal soul deserving of judgment in the afterlife. These were all cases where a single concept was needed to track diverse components.

§1.1 INCONSISTENT RESPONSES

Williams introduced a significant advance to the traditional Lockean thought experiment by injecting Shoemaker's version of it with a dose of self-concern. The angle of self-concern was genuinely novel and introduced a first-person point of view to a thought experiment which had previously been explored exclusively from a third-person point of view. The move to the first-person point of view—and specifically the addition of self-concern as an investigative tool for personal identity—was a prescient one, because it revealed our conflicting intuitions about our identity: was it psychological or biological? Shoemaker had already done the work of cleaning up Locke's thought experiment and rendering its premises palatable to the contemporary reader, exchanging Locke's vague language about the switching of consciousnesses with a more scientific account of brain transplants. The former story lacked the details necessary to engage the scientifically sophisticated modern reader; the latter story included the necessary details although the details seemed to some to strain credulity. But concerns about the plausibility of a brain

transplant were dismissed as "merely technical" impossibilities—not the deeper logical impossibilities that might derail the experiment's coherence.

Williams recognized that the brain transplant was an unnecessarily physical way of achieving Locke's switching of consciousnesses. His thought experiment involved a brain zap in which the brain was effectively "wiped" of its psychological content by means of physical tampering; through the same process the brain could receive a new set of psychological dispositions, memories, beliefs, etc.... This process produced the same results as the brain transplant without actually moving the brain, which in turn could be consistently reidentified as having not moved on the basis of spatio-temporal continuity. This change also turned out to be significant for the results of the thought experiment.

In the first trial of the thought experiment, persons A and B each undergo the procedure described above, with the result that the character that was once expressed in A's body is now expressed in B's body and vice versa; the character expressed in B's body is expressed, after the operation is completed, in A's body. It is at this point that self-concern enters the picture. One post-operative body will be tortured and the other body will receive a large sum of money and the participants, before the operation, are each asked which body ought to receive the torture. It seems rational to think that before the operation the A-bodied person ought to argue, in self-interest, that the A-bodied person produced by the brain zap ought to receive the torture. After all, his beliefs and desires, his memories and psychological dispositions, indeed his entire character, will now be housed in the B body. He cares not about the body he *used* to inhabit but rather the body he *will* be inhabiting when the torture is to be administered. Conversely, the B-bodied person before the operation would be rational in asking for the B-bodied person produced by the brain zap to receive the torture. After all, his beliefs and desires, his memories and psychological dispositions, indeed his entire character, will now be in the other body. In short, if you switch bodies, you want life to go well for your *new* body.

This result can be taken as hard evidence for a psychological account of personal identity. Unfortunately, Williams demonstrated that our responses to the thought experiment are not constant. A second trial of the thought experiment produces contrary results. Consider the following: you are informed that tomorrow you will undergo a painful medical procedure that is not to your benefit. You will be tortured. The news scares you and you go to bed dreading tomorrow. You are also informed that before the torture your brain will be physically tampered with such that you will not even remember the advance notice that you received a day before the procedure. Indeed, you will wake up with memories and character traits that you did not have before the procedure. But this does little—perhaps nothing—to soothe your anxiety at the coming of tomorrow and the painful operation.

Williams notes that what distinguishes the second trial from the first is simply the withholding of information about what happens to the other body. In other words, the difference is that the experimental subject is not informed of a second body who will receive memories, beliefs, desires, psychological dispositions and character traits that are qualitatively similar to his. But why should this omission be significant? Why should the appearance of another body, with psychological dispositions just like yours, be relevant to your self-concern about the impending torture? How could it be rational for this knowledge to mitigate your anxiety? In this case it is clear that it is not. But this is precisely what was described in the first trial of the experiment—a second body that (question-beggingly) "receives" your psychology.

Consequently we have two versions of the very same thought experiment that produce radically different expressions of self-concern. The first can be taken as evidence for the psychological criterion; the second can be taken as evidence for an animalist account. Williams used this situation to justify his claim that no thought experiment about self-concern, or any similar neo-Lockean thought experiment, will help resolve the debate about personal identity. The decision between the accounts must be made on alternate grounds.

Those alternative grounds rest on an objection from branching.[4] The issue has become such a staple of the literature that I shall avoid rehashing the specifics, such as defective teletransportation and the splitting of a brain into two hemispheres and their subsequent transplant into qualitatively similar bodies. What I will concentrate on is Williams' claim that the animalist account is less vulnerable to an objection from branching and is consequently the best candidate for a theory of personal identity. But it is unclear whether animalism is any more immune from branching problems than a psychological account of personal identity. Indeed, it seems that most theories of personal identity are vulnerable to branching—with the exception of essentialists such as Wiggins who rule it out as logically impossible or neo-Lockeans such as Shoemaker who restrict their criteria to non-branching worlds. While we will deal with Wiggins' theory of natural kinds in the next chapter, Shoemaker's approach of biting the bullet and restricting the criteria to non-branching worlds violates our intuition that the identity relation must be intrinsic. And Nozick's Closest Continuer theory of personal identity—which also provides a response to branching problems—violates the very same constraint.[5] Hence the branching objection is solved but it leaves in its wake an even bigger problem: showing how the identity relation need not be intrinsic. Chisholm gets around the branching objection because he believes in what Parfit calls a "further fact" to personal identity—although he cannot quite articulate what it is and why it must be the case that there is a determinate answer to the question of which of the two resulting persons I will be after branching. He simply says it must be so.[6] The point here is that the branching objection, as well as the responses from Shoemaker, Wiggins, Nozick and Chisholm, provide no reason to prioritize our animalist intuitions over our psychological intuitions—or vice versa. Consequently, the issue of branching by itself

---

[4] Bernard Williams, "Bodily continuity and personal identity" in *Problems of the Self.*

[5] Robert Nozick, *Philosophical Explanations* (Cambridge: Harvard University Press, 1981).

[6] Roderick M. Chisholm, "Identity Through Time" in *Language, Belief and Metaphysics*, edited by Howard Kiefer and Milton K. Munitz (Albany: State University of New York Press, 1970).

provides no independent reason to resolve our conflicting intuitions by favoring one over the other.

But we should look at Williams' objection more closely to see why he thinks that branching provides an extra reason to adopt an animalist view of personal identity. Consider the following case of physical duplication via amoeba-style splitting:

> It is possible to imagine a man splitting, amoeba-like, into two simulacra of himself. If this happened, it must of course follow from my original argument that it would not be reasonable to say that either of the resultant men was identical with the original one: they could not both be, because they are not identical with each other, and it would not be reasonable to choose one rather than the other to be identical with the original. Hence it would seem that by my requirements, not even spatio-temporal continuity would serve as a criterion of identity: hence the requirements are too high.[7]

One might think that this is a bona fide case of duplication and therefore the physical criterion is in the same boat as its competitors. But Williams argues that there is a relevant difference between this style of duplication and the variety faced by the psychological criterion. To make use of the physical criterion in deciding identity is to "chart an historical course" of the entity in question which will, if successful, reveal the existence of the duplication. "This consideration puts the spatio-temporal continuity criterion into a different situation from the others discussed; for in this case, but evidently not in others, a thorough application of the criterion would itself reveal the existence of the reduplication situation, and so enable us to answer (negatively) the original identity question."[8] By this Williams means to say that we would be justified in saying that the entity does not survive the split. For persons, we would presumably say that physical branching equals *death*. The analogy to amoebas helps Williams here, for our intuitions tell us that when an amoeba splits, *two new amoebas* take its place.

Williams' conclusion that physical branching equals death hinges on his claim that "in the strict sense" the results of the branching would not be spatio-temporally continuous with the pre-fission original. This claim rests on several assumptions, some of which might be denied: that no entity can have mereologically detached parts, that the same cannot be

---

[7] Williams, "Bodily continuity and personal identity," p. 23.

[8] Williams, p. 24.

said of the other criteria, etc. First, it is unclear why "in a strict sense" the resultants are not spatio-temporally continuous with the original. Presumably, this is because the task of historical tracking will be irresolvably deadlocked at the moment of branching. While looking for the next "stage" of the continuant, there will suddenly be two, equally viable candidates for the next stage of the continuant. Unable to find a non-arbitrary reason to choose one of the two stages, the process of historical tracking must come to an end. Consequently duplication poses no problem to the physical criterion because it yields a definite answer to the identity question: entities never survive branching.

If this is the reason why the results of the split are not "strictly speaking" spatio-temporally continuous, it is unclear why the same is not true of psychological fission. For in that case, the same kind of historical tracking can identify the history of psychological continuity. At any one stage in time one looks at the psychological dispositions of the entity in question and then looks, at the next moment, for the psychological dispositions which are appropriately related in a causal way to the previous set. The two sets will be closely psychologically connected to each other. More importantly, there will be a causal nature to these connections. It will be the task of the historical inquiry to track these causal connections. In a case of branching, at one moment there will be a set of psychological dispositions which are followed, at a subsequent moment, by *two sets* of psychological dispositions which are equally related and both qualitatively similar to the original set. Moreover, the same causal mechanism is responsible for one set as is responsible for the second set. There being no non-arbitrary criteria upon which to choose between one set or the other, the process of historical tracking is irresolvably deadlocked and must come to an end. This is precisely the same situation as we saw in Williams' physical criterion.

What could be the source of the alleged or perceived asymmetry that animates Williams' argument? I can think of none, except that some bias would lead the supporter of the physical criterion to assume that theirs is a process of historical enquiry and the psychological criterion is somehow not. But this is clearly not the case—both are equally a

case of historical enquiry. Perhaps the asymmetry stems from Williams' idea that questions about psychologically continuity and connectivity are not questions about numerical identity but are rather about qualitative similarity. Evidence for that would be that someone has my psychological dispositions if they have psychological dispositions *exactly like mine.* But this is not a question of bare identity, according to Williams. Questions about spatio-temporal continuity in the physical realm *are* about numerical identity, according to Williams.

Even if this is the case, it still does not add up to the claim that historical enquiry is the sole purview of the physical criterion. That is because it is not numerical identity which is the hallmark of historical enquiry. Rather, the real hallmark of the enquiry is the identification of causes, of causal chains and causal pathways. The tracing through of these causal chains is the real task of historical enquiry and it poses just as much problem for physical criterion as it does for the psychological criterion.

Perhaps the asymmetry stems from his claim that "the normal application of the concept of continuity is interfered with by the fact of fission, a fact which would itself be discovered by the verification procedure tied to the application of the concept."[9] First of all, some substantive account of "normality" is required to show that the fact of fission "interferes" with the concept of continuity. On the face of things it is not immediately obvious why this should be so, unless the concept of continuity precludes, by definition, the possibility of branching. If this is so, one must ask why and with what warrant such possibilities are prohibited. A substantive account ought to be given here, for it is unclear why duplication is "unnatural" except by appeal to some substantive account of the "natural" development and purpose of the entity involved.

This substantive account might take the form of a Natural Law account of persons. According to Natural Law it is not possible for biological bodies to branch. This approach would emphasize that our notion of what constitutes a biological human body is fixed in

[9] Williams, p. 24.

relation to the physical and scientific laws which govern the creation and development of the biological human bodies in the world. Those scientific laws are the biological laws which codify the process by which human bodies are generated through reproduction, cell division and the development of the embryo. These are the *natural* processes by which human beings come into existence and develop through maturity. It is therefore *unnatural* for biological human bodies to "branch". What is the basis for this conclusion about what is natural? The basis is simply the nomological regularities by which human beings develop. The point is not simply that biological human beings usually begin and develop in this way. Rather, the point is a stronger one: that the creation and development of bodies is so regular that it counts as nomological. It can be expressed in the form of a *law*. This fact alone justifies the quasi-normative claim about what is natural and what is unnatural. Consequently we are spared the burden of having to consider the branching of biological human bodies. According to this account, at least, it violates *natural law*.

I will not offer here a direct repudiation of this line of thinking although I will discuss the subject in some grater depth in the next chapter during a consideration of the essentialism of Wiggins. But suffice it to say that *if* a natural law account here is plausible I see no reason why it should be plausible in the case of animalism but not plausible in the case of psychologism. If one can go the natural law route with regards to the branching of physical bodies one can go the natural law route with regards to the branching of psychological continuity. The point is not whether the natural law account is correct. The point is that regardless of whether it is correct or not, it does not provide an independent reason for elevating our animalist intuitions over our psychological intuitions or vice versa. Branching of psychological continuity would be just as nomologically irregular as the branching of physical bodies.

To put the point in the terms used by Williams, if the normal application of the concept of physical continuity is interfered with by the fact of fission, it is unclear why the same application in the case of psychological continuity is not similarly interfered with.

Presumably the interference stems from the existence of an outside force that causes the duplication, i.e. it is an "interference" because it is not internally generated by the entity in question. But again, there is no asymmetry here with the case of psychological continuity, where the concept is equally violated by the existence of the outside interference. Regardless of the nature of the interference, if the concept is violated for the case of physical continuity it is violated for the case of psychological continuity. The central issue for either kind of continuity is not just identity but *causation*—and an "abnormal" causal chain is possible for either kind of continuity. Once again there seems little reason to suspect a relevant asymmetry between the criteria.

Without the relevant asymmetry at hand, we are left with no convincing reason to support the physical criteria over the psychological criteria. Two thought experiments have yielded conflicting results, each of which is supported by our intuition, but which yield incompatible conclusions about the nature of personal identity. Also, the alleged "objection" of duplication, while certainly problematic for the psychological criterion, is equally problematic for the physical criterion. And if it is not problematic for the physical criterion, the very reasons that render it unproblematic also render the objection unproblematic for the psychological criterion. Regardless of which fork in the road is taken, the destination is indistinguishable: the lack of a relevant difference between the two accounts. Therefore, we have no positive reasons for supporting one account over the other. Although nothing in this argument precludes the logical possibility of finding such reasons in the future, it is sufficient for my purposes to point out that intuition supports *both* accounts (depending on which thought experiment is being considered) and no new positive reason has emerged on the horizon.

One strategy would be to proactively look for these reasons, perhaps by fine-tuning our thought experiments, with renewed sophistication, in order to find deeper and more primary intuitions with the hope that they will be less contradictory than the surface intuitions we have already identified. One strategy here would be to review the state of the

literature on personal identity thought experiments. That would be too onerous a task and would take us too far afield from the main line of the argument. There are literally hundreds of these experiments and the volume increases annually. In any case the review is unnecessary because the same results may be achieved by looking, in an abstract and general way, at *Gedankenexperimente* methodology and evaluating the probability that a refined experiment will expose one of our intuitions as chimerical (despite what we previously thought) or reveal deeper level intuitions which are not mutually inconsistent.

Before we continue any further, it is important to distinguish two different but related issues: personal identity and an account of persons. The issue needs to be addressed because I have taken our inconsistent responses in the Williams thought experiment about personal identity as evidence that we are conflicted not just about the nature of personal identity but also about the nature of personhood itself. I think a limited conflation of the two issues is appropriate for the following reasons. This can be shown by first looking at how they are different.

A theory of personal identity and a theory of persons are different because the first is an account of the conditions under which a person continues to be—or ceases to be—the same person over time (where "sameness" is cashed out in terms of numerical identity), while the second is an account of what counts as a person *simpliciter*. Here's one way of putting the difference: the first theory tries to individuate individual persons from each another and the second tries to individuate persons as a class from non-persons. While the first theory (personal identity) tells you when a particular person is still the *same* person over time, the second theory (a theory of persons) tells you when a being doesn't count as a person at all.

Although it is important to remember that these are separate questions ("What counts as a person?" versus "What counts as the same person?"), in several important respects these two inquiries are deeply related. And here's the relationship; in providing an account

of personal identity one naturally provides, either explicitly or implicitly, at least a partial theory of personhood. Put simply, in telling me when a particular person is the same person, you inevitably tell me something about what counts as a person. This for the simple reason that when a person ceases to be a person at all, this logically entails that it stops being a particular person too.

An analogy might help. In developing a theory of individuation for, say, bottles, you naturally also say something about what counts as a bottle. To put it bluntly, if you develop a spatio-temporal criterion for bottle identity you implicitly assert a claim about the physical nature of bottles. Similarly, the claim that personal identity is given by the spatio-temporal continuity of the biological body implies a certain claim about the physical nature of persons. However, it is not the case that a theory of personal identity entails a *complete* theory of persons, just as a theory of bottle identity does not logically entail a total theory of bottles. (For example, to offer a spatio-temporal criterion of personal identity is not to deny that persons lead psychological lives.) But that does not matter. What does matter is that the two theories are tightly connected. For example, if one were to eliminate the concept of the person, the issue of *personal* identity would disappear. Or perhaps it would be better to say that the issue of personal identity would be *transformed* (depending on what concepts personhood was being replaced with). So for our purposes it is appropriate to discuss personal identity and a general theory of persons in the same breath.

§1.2 QUINE'S OBJECTION

Quine had a distinctive rationale for disliking thought experiments and his view ought to be examined closely. He argued that our concepts, and the words we use to describe them, do not apply outside the empirical contingencies that gave rise to them. Our notion of the person only has an application when all of the empirical regularities come together: i.e. when human beings have brains, consciousness, bodies and agency. It is precisely these normal situations that gave birth to our concept of the person, so to ask in a thought

experiment what happens to our concept of the person when these empirical contingencies fail to hold is to ask our concept of the person to do more than it was ever designed to do. "To seek what is 'logically required' for sameness of person under *unprecedented* circumstances is to suggest that words have some logical force beyond what our past needs have invested them with," Quine wrote (emphasis added).[10]

Quine's objection is a serious one and stems from his doctrine of meaning holism, in which our beliefs are not independent, but rather interrelated in a complex web whose individual components are mutually related and supporting. Beliefs are not defined in isolation but in relation to each other. As a consequence, belief revision is more than just a change in an isolated belief. Rather, belief revision is the adjustment process of one's current doctrine. Like ripples in a pond, alterations in one belief affect other beliefs whose meaning they rely upon. To take just the most basic example: changing one's concept of 'man' might require not only changing one's understanding of 'woman', 'boy', 'girl', 'husband' and 'wife', but every other gendered concept whose meaning relies on an understanding of the concept of 'man'. Changing one's belief about 'man' requires a wholesale adjustment of one's current doctrine.

Quine's meaning holism poses a problem for some thought experiments in the personal identity literature. Prying our concept of the person away from the empirical contingencies under which it was formed is the same as asking us to revise our concept of the person. However, such an isolated revision is impossible. Altering our concept of the person for the sake of a thought experiment (e.g. stipulating that all persons could perform fission and fusion at will) would require us to alter an almost infinite number of supporting beliefs in our current doctrine that rely on: a) the fact that persons reproduce only through mating; b) the fact that persons only have their own quasi-memories, not those of others; and c) countless other beliefs. A simple thought experiment would be insufficient to track this complex adjustment. If some of these changes *do* occur, we might be forced to go through

[10] Quine, p. 489.

this process anyway. One current example might be human cloning. If this process—already executed in sheep and other farm animals—is extended to humans in the future, we might be required to alter our understanding of our concept of the person and revise our concepts to meet our new needs. But a thought experiment would be legitimate only if it recognized that our adjustments will ripple outwards from the center of our current doctrine. Many thought experiments do not. A thought experimenter who makes a radical supposition and then asks how that would change *just one belief* is not being faithful to Quine's meaning holism. Alterations to the concepts cannot be performed atomistically. Change one concept and the consequences ripple outward across the network. To change one concept without making the necessary alterations to the other concepts in the network causes *incoherence* to ripple outward in the network. The result is a network of concepts from which one can derive no meaningful conclusions. The rub is that this is precisely what thought experiments, especially extreme ones, try to accomplish.

Quine reminded us that we should not expect our words to have more meaning than our present needs have invested them with. This is a general attack on the methodology. Quine argued that thought experiments often push our concepts to the breaking point. To ask how our present concepts would work in a radically different world was to ask a nonsensical question because our concepts were *custom-designed* for the actual world. We have no reliable intuition to draw upon when asking how our current concepts would function if the world was radically different. The lack of a reliable intuition stems, in part, from the fact that we are in the actual world, not the possible world being imagined by the thought experiment. The result is that we often substitute *legislation* for *discovery*. In the absence of information to discover about a possible world, we begin by legislating a coherent account of how our concepts might function. But this is a far cry from what most thought experiments aim for.

§1.3 THE LIMITS OF THOUGHT EXPERIMENTS

Critics of thought experiments in personal identity have picked up on Quine's point—which is a good one—and have used it to question the dominant methodology of the field. Wilkes notes that "if the world were indeed radically different, then practically nothing is left for us to rely on, and it is doubtful that any of our concepts would remain secure. Certainly we cannot claim that just the notion of a person alone would (or would not) need to vary... in a world indeterminately different we do not know what we would want to say about anything."[11]

The idea Wilkes is invoking is not just a Quinean one but a Wittgensteinian one. If the world were radically different, our concepts would be too. But in that case, the concepts being different, it does not tell us much about our concepts in the actual world. And moreover, although we can imagine that our concepts might have been different, this does not necessarily mean that we can fully explore how those concepts would have taken shape. Wilkes puts the point this way:

> If we adopt this recommendation we can indeed understand how intelligent beings might have devised quite different concepts from our own. What we cannot do (and Wittgenstein never says that we can) is imagine what these concepts would look like; and this is absolutely crucial for the thought-experimenter. We cannot, that is, couch them in our own terms: there is no background common to the two worlds, since the 'very general facts of nature' are left indeterminate and undetermined. In a world where these general facts of nature were so unlike our own as to allow amoeba-like Martians to flourish, we just have no idea what we should say about it.[12]

Although Wilkes talks about the lack of a common background to the two worlds, I do not believe that the point relies on a claim about conceptual schemes or paradigm incommensurability. Rather, the point is simply the common-sense idea that concepts are *custom designed* for the world and that radically different realities would require radically different concepts. And it is unclear what those radically different concepts would tell us about our use of our own concepts. And furthermore, it is unclear what empirical or conceptual access we have to these radically different concepts anyway.

---

[11] Kathleen Wilkes, *Real People: Personal Identity without Thought Experiments* (New York: Oxford University Press, 1988), p. 46.

[12] Wilkes, p. 48.

Because of this quite legitimate Quinean and Wittgensteinian anxiety, it is important to place severe restrictions on the conduct of thought experiments. (This restriction flows naturally from meaning holism. I will not offer an independent defense of the doctrine of meaning holism here because it would take us too far away from the main argument. I concede that the point relies on such a defense and that those who do not subscribe to the doctrine would be unimpressed by the alleged anxiety.) If the problem is the so-called "ripple" effect in the network of concepts, the burden falls upon the thought experimenter to fill out the relevant alterations in *all* of the affected concepts. If the thought experiment is about a major concept which occupies a large node in the network, then the number of potential revisions to the network could be daunting. The burden falls upon the thought experimenter to explicitly list the concepts that are to be changed in accordance with the new reality and those concepts that are meant to be held fixed. Failure to keep these two separate can lead to inconsistent responses—the Achilles' heel of the thought experiment. However, explicitly listing the concepts that are to be replaced with their counterparts in the possible world and those concepts that are to remain fixed is such a prohibitive requirement that it may not be possible in all cases. Indeed, thought experiments that imagine a possible world *radically* different from the actual one may require rewriting such a large percentage of the network as to be unworkable. And that reworking is important if the thought experiment is to guard against inconsistent responses.

One might object that it would be impossible to list *every* change that flows from even a simple thought experiment. Consequently the constraint is much too restrictive because it would prohibit all thought experiments—even benign ones. But the answer is quite simple: it is not necessary to explicitly list *every* change in the network of concepts. This is indeed an impossible task. What is required is an honest attempt to trace the ramifications of the proposed change just enough to demonstrate that the thought experiment is coherent and, more importantly, just enough for the subject of the

Gedankenexperiment to get a realistic picture of the possible world that she is being asked to consider.

Here is one example of the problems associated with this kind of failure: Newton imagined watching, in an otherwise empty universe, the unwinding of a water-filled bucket suspended by a twisted rope.[13] At the moment before the bucket is released the water is level in the bucket. At the moment just after the bucket is released there is relative motion between the water and the bucket. After several moments have elapsed, the water and the bucket are at rest but the water is no longer level as it was at the beginning. Newton asked why the water is level in the first moment but not in the final moment. His answer was that in the final moment the water and bucket are in motion relative to absolute space, which they were not at the first moment. Newton took this thought experiment to be evidence for absolute space.

Physicists later questioned Newton's thought experiment by denying the alleged observation of the third moment, saying in effect that if the universe were truly empty (and consequently empty of all gravitational influences such as the fixed stars), the water and the bucket would *still* be level in the final moment. Hence the route to Newton's conclusion is invalid.

What is the source of the inconsistent response to Newton's thought experiment? The trouble stems from the fact that the subject of the thought experiment was not given enough information about the concepts of this radically different universe. This universe is so different that it only includes one thing: a water-filled bucket suspended by a twisted rope. Our only experience with buckets stems from our actual world—a world which includes many things, the least of which is the fixed stars which compose a massive amount of matter exerting gravitational influence.

---

[13] This example is taken from Roy A. Sorenson, *Thought Experiments* (New York: Oxford University Press, 1992).

In order for the thought experiment to yield consistent responses, all of the concepts in the network that were affected by the fact that the universe was virtually empty should have been appropriately revised. And this is a daunting task because all of the concepts in modern physics were custom designed to help us understand a world which is not empty. It would take a lengthy revision to alter our current concepts to be appropriate for a world with virtually no matter. Of course, only the concepts in the network that are relevant to the thought experiment at hand need be revised. But this is precisely the mistake that thought experimenters most often fall victim to—they radically underestimate the number of ancillary concepts that turn out to be relevant to the experiment. And this was Newton's mistake: he did not realize that the lack of matter would dramatically affect the behavior of the water in the bucket. And this was so because all of his (and our) experiences with water and buckets took place in a universe with much more matter than his imagined universe. He had, by definition, no empirical experience in his imagined universe to draw upon, and his empirical experience in the actual universe was extended inappropriately and without warrant to the imagined universe.

None of this means that thought experiments cannot be used. It only means that a more rigorous experimental method for thought experiments needs to be enforced, as is the case with the scientific method.

Another common criticism about thought experiments in general (and specifically the outlandish ones that dominant the field of personal identity) is that they ask us to draw conclusions from what may be impossible situations—brain rewiring, brain transplants, teletransportation to Mars, etc.... And what value could these conclusions have if they depend on assumptions which turn out to be impossible? This objection to the methodology is less viable because there are multiple responses to it.

Take the much discussed case of brain transplants. Wilkes argues that the practice is impossible, despite the fact that it is based on actual cases of brain bisection performed on epilepsy patients. The data, brought to the attention of the wider philosophical community

by Nagel, suggested that just half of a brain could support continuity of consciousness.[14] In the cases cited by Nagel, doctors performed commissurotomies (severing the neuro-connections between the hemispheres) with the result that the patients seemed to exhibit evidence of multiple streams of consciousness corresponding roughly to each hemisphere. This bit of empirical data was taken as evidence that each hemisphere was capable of supporting an independent stream of consciousness and subsequently became the basis for an entire genre of thought experiments.

Wilkes has argued that brain transplants are, in fact, impossible, or if not necessarily impossible, then a lot more difficult than thought experimenters have taken them to be. Although neurosurgeons have severed the hemispheres, they have never split the lower brain, and Wilkes claims that it is unlikely to ever happen since the brain stem is not naturally divided into hemispheres like the upper brain. Secondly, she questions the likelihood of severing a hemisphere from one brain stem and reattaching it to a second brain stem. It is unclear how this could be performed without permanent damage to the brain. Finally, thought experimenters underestimate the degree to which the brain stem and the rest of the nervous system carry personality traits. Consequently, a full exchange of personalities would require not just the transplanting of a cerebral hemisphere but the transplanting of the *entire* central nervous system throughout the body, a next-to-impossible medical feat.

Does any of this medical skepticism invalidate thought experiments inspired by brain bisection? Parfit's response to the objection has been to distinguish between technical and logical impossibilities.[15] The distinction is a good one because it explains the different categories of *possibilia*. Thought experiments that imagine technically impossible situations need not be discarded, for they merely take for granted technological advances that have not yet happened. And why should it matter to the thought experiment if the technological advances are in the future instead of the present? On the other hand, logically impossible

---

[14] Thomas Nagel, "Brain Bisection and the Unity of Consciousness" in *Mortal Questions* (New York: Cambridge University Press, 1979).

[15] Derek Parfit, *Reasons and Persons* (New York: Oxford University Press, 1984), ch. 12.

scenarios are *prima facie* irrelevant for philosophy because there is no logically possible world where the scenario could take place. And in fact, it would be impossible to coherently describe and explore such a scenario if it is indeed logically impossible. According to Parfit, so long as a thought experiment does not trade on the logically impossible, it is valid:

Sorensen pursues a similar line. He defends thought experiments by drawing an analogy to scientific experiments. His general strategy is to justify thought experiments by showing that they differ from scientific experiments by degree and not by kind. Thought experiments are to be classified as a limited case of experiments *simpliciter*. Although scientific experiments have an empirical component, they are not exclusively empirical affairs. They include rational and conceptual components that function as persuasion. Thought experiments include these conceptual components while forsaking the empirical components, usually because the empirical portion cannot be performed. Sorensen's general strategy of justifying thought experiments by analogy to scientific experiments (arguing that if the latter is valid, so is the former) is similar to Parfit's point. Both prioritize logical possibility over technical possibility. When Sorensen says that thought experiments differ from scientific experiments in degree and not in kind, this distinction tracks Parfit's idea that valid scientific experiments and valid thought experiments are both logically possible. The fact that the former may be technically "possible" while the latter may be technically "impossible" is a matter of degree—not kind. It does not put the experiment into a different category.

But the distinction demands a deeper analysis than Parfit provides us. What about imagining a scenario that is somewhere in between the technically impossible and the logically impossible? Consider, for example, the contingently impossible. Perhaps we could imagine that biological humans were, in some fundamental way, different. This is not logically impossible, but nor does it trade on some technological advance far in the future. A thought experiment might ask us to imagine ourselves as different creatures than we actually

are. Most philosophers writing in the field of personal identity see no problem with this kind of thought experiment because it does not violate the logically possible restraint.

The problem with imagining ourselves differently than we really are is that our concepts were custom-designed to meet our present needs, and our present needs do not extend to cases where we imagine ourselves as radically different entities, even if the fact that we are not such radically different entities is a purely contingent fact. This shows the depth to which the Quinean objection penetrates. What reason do we have to think that our concepts would work even reasonably well in a world in which we were radically different creatures? Those concepts were created to deal with the kind of creatures we actually are, living in the world we are actually living in. If we were different creatures with different lives we might have different concepts. This is all that we can conclude from a thought experiment based on this kind of conjecture.

If this kind of conjecture violates the Quinean point, then perhaps even the other cases of technical impossibilities violate it, because even technological advances are relevant to the development of our concepts. What does it matter to concept-formation whether the facts of the present world are a function of technological limitations or contingent facts which might have been different? In both cases we must admit that concepts were designed in accordance with the way the world is, regardless of the cause of that situation. One might object that the difference is between a possible world and the future of the actual world, i.e. that imagining that we had been different creatures is a question about a possible world, while imagining a technological advancement would be a question about the future of the actual world. And only the former is a case where our concepts would not apply because the latter is a case where presumably the same concepts would apply.

Naturally, this distinction eventually breaks down. We might *become* radically different creatures *because* of a technological advance. For example, if brain bisection and brain transplants were technologically feasible, and the public at large thought the procedure desirable, we might become creatures who divided periodically into multiple creatures to

which we were psychologically connected. If the technology allowed it (and one could easily amend the thought experiment to make it possible), we might divide periodically, say every year, and produce not just two psychologically continuous descendants, but perhaps dozens, or a near infinite number. The creatures in question would then be radically different from the kind of creatures that we are—so different that our current concepts would no longer be applicable. And all of this from a mere "technological" change. Therefore, the bright line between possible worlds and the future of our actual world is not so bright. The future of our own world could conceivably be so different as to render our current concepts unusable.

This reveals the strength of the Quinean point. Our concepts are not just custom designed for the actual world (as opposed to other possible worlds), concepts are custom designed to respond to our *current* needs. If the future is radically different from the present, then we will respond to the occasion by adding precision to vague concepts, amending other concepts, and introducing new concepts where the previous two strategies fail to adequately meet our needs. But the idea here is that the Quinean point applies not just to the actual world, but to the actual world *at the present*. Creatures in a different circumstance might have different concepts. It does not matter if the creatures are in a possible world or in the future of the actual world. The point is just that concept formation responds to need.

None of this is meant to suggest that Parfit's distinction between technical and logical impossibility is not a good one. Rather, the argument is meant to show that the Quinean point is prior to—and deeper than—the distinction between technical and logical impossibility. Just because a situation is not logically impossible does not necessarily mean that it is valid; it still bears the burden of meeting the Quinean constraint. And the Quinean constraint does not rule out the thought experiment entirely. It simply sets the following constraint: the background conditions must be thoroughly spelled out, concepts across the entire network need to be altered in accordance with the scenario being imagined, and the factors that are being held constant need to be identified. These conditions ought to be adopted as a generalizable experimental method for *Gedankenexperimente*. Only if these

conditions are fulfilled does one have a well-executed thought experiment that could yield consistent responses.

I have not rejected the entire methodology of thought experiments; neither have I uncritically accepted it. Thought experiments can be flawed, they can be misused, and they can be poorly constructed. But this stops short of saying that the practice ought to be abandoned. I have marshaled no conclusive, knock-down argument justifying this prescription. What I have marshaled is powerful evidence for *skepticism* about thought experiments.

Given this conclusion that thought experiments are legitimate (though they ought to be restrained and placed under severe methodological constraints), where does this leave us vis-a-vis the possibility of resolving our conflicting intuitions about what it means to be a person? The current strategy in the field of personal identity is to conduct new thought experiments with the hope of resolving these conflicting intuitions. I will argue that this strategy should be abandoned. To do this I do not need to assert that thought experiments will *never*, by definition, resolve the dilemma of our conflicting intuitions. I am not suggesting that it is impossible to come up with a coherent account of personhood. This is not necessary for my argument to succeed. All I need to show is that we have these conflicting intuitions and that the strategies on the horizon involve denying one of the two intuitions. But one way to resolve this conflict without denying either intuition would be to eliminate the concept of the person.

Traditionally, the response has been that the concept of the person was indispensable, i.e. that personhood matters. This is precisely the assumption that this project aims to debunk. If personhood matters, it is only because its components matter. That being the case, we ought to stick with the components and eliminate the concept of the person.

A pragmatist inspiration looms large here. Deciding what it means to be a person is a purely metaphysical concern. According to my argument, resolving our conflicted intuitions is not necessary for inquiry. We can eliminate the cluster concept and replace it with its sub-

concepts and the project of inquiry will not be threatened. In fact, it is the aim of this project to demonstrate how inquiry might be *improved* by making this change. A good pragmatist would suggest that if it makes no difference to inquiry, it should make no difference to philosophy. This project will demonstrate that, despite what common-sense tells us, it makes no difference to inquiry what it means to be a person.

This conclusion is inspired by the economy of intellectual resources. Philosophical inquiry requires time, money and resources—just like scientific inquiry. Decisions must be made about fruitful areas of inquiry and lucrative avenues of investigation. This is just as true in philosophy as it is in science, where corporations and governments must decide where to spend their limited financial resources. In philosophy, we need to identify avenues of investigation that are likely to yield progress. It is my contention that continuing the present-day strategy of refining thought experiments, in the hope of resolving our conflicting intuitions about personhood, is not a fruitful expenditure of philosophical resources. Although I have offered no proof that thought experiments will *necessarily* fail in this task, I have offered enough reasons to support skepticism that this strategy will yield a resolution. And why should we continue to pursue a course that has not yielded results and is unlikely to yield results in the near future? Furthermore, responsible methodology severely limits our attempts at finding a resolution through thought experiments.

What if someone appears on the scene in the future with a resolution to these conflicting intuitions? I have not claimed that this is impossible. If it does happen we should evaluate the results and compare them to the eliminativist strategy. I will attempt to show in this project the virtues of the eliminativist strategy. In short, the virtues come down to a resolution in our metaphysical intuitions about personhood without doing any damage to our value theory. In fact, it may even be the case that our value theory will be the better for it. But if a new strategy to resolve the conflict were to appear on the stage in the future, there is nothing in the eliminativist strategy which would prevent us from evaluating both proposals by the same pragmatist criteria. Should the new proposal fare better by those lights

we could always abandon the eliminativist strategy. But until that day comes—and I am skeptical that it will for the reasons already articulated—we are justified in pursuing the eliminativist line. It is not my burden to demonstrate that no other plausible strategy will ever present itself. Rather, my burden is simply to demonstrate that the eliminativist strategy is better than the options presently available and the options on the conceptual horizon.

Having explained the potential for thought experiments to resolve the conflict, let me now address the role thought experiments played in *revealing* the conflict. The burden on my argument is to defend thought experiments just enough to justify their use in revealing the conflict. Otherwise, if the methodology was faulty, the conflict could be dismissed as chimerical. I will stop short of offering a knock-down argument justifying *Gedankenexperimente* in this case. Rather I will explain what is going on in the Williams thought experiment, with some help from Kuhn's story of how thought experiments can lead to theory change.[16]

Initially, a concept is used without complication. It "fits" the world—or at least the world as the users of the concept see it. Put another way, the concept, and the conceptual framework it represents, *works well.* But there's some small piece of information or experience which cannot be assimilated by the conceptual framework or cannot be expressed by the concept. But it is important to remember that the concept works fine, not just for the majority of the cases but for *virtually all* of the cases. In our case, this happens when the person concept has difficulty dealing with the marginal cases of everyday life—fetuses, Alzheimer sufferers, brain-dead patients, etc., which the conceptual framework has difficulty assimilating. According to Kuhn, the users of the concept have all of the information available, and they are theoretically capable of revising their concepts, but they have difficulty organizing the information in the appropriate fashion, or they have difficulty seeing some of its logical entailments. Because of the failure of the concept its users may be

[16] T.S. Kuhn, "A Function for Thought Experiments" in *The Essential Tension* (Chicago: University of Chicago Press, 1977).

left with a feeling of "uneasiness" according to Kuhn. This is a perfect description of our everyday feelings about the concept of the person. Although we feel that the concept is generally adequate, we (even non-philosophers) have a faint recognition that it is less successful in the marginal cases, though we are not *fully* aware of the depth or significance of the problem. Indeed, some non-philosophers humbly attribute the problem not to a difficulty in the concept fitting the facts of the world, but instead see the problem as one of conceptual or linguistic competence—i.e. their inability to use the concept properly. They think like Julius Caesar and attribute the fault, not to the concept, but to themselves.

Kuhn argued that a good thought experiment functions to bring the confusion (and the concept's inadequacy) to the attention of the concept user. This, I submit, is the function of the Williams thought experiment. It pushes to the foreground our anxieties about the concept that until that moment have lingered in the background. In the process an implicit confusion about our use of the word 'person' is rendered explicit. We realize, on a conscious level that our understanding of the term 'person' may no longer suffice in the near future.

Kuhn is very instructive on this point about the experimental conditions that can be used to reveal the concept's problems. He argues that impossible scenarios dreamed up from possible worlds are problematic because our concepts "were never intended to apply in such a case." Speaking of Galileo, Kuhn writes that

> if this sort of thought experiment is to be effective, it must allow those who perform or study it to employ concepts in the same ways they have been employed before. Only if that condition is met can the thought experiment confront its audience with unanticipated consequences of their normal conceptual operations.[17]

This constraint is the same as the Quinean constraint that I introduced above, although presented in different language. If the situation imagined is too disconnected from one that might be experienced in everyday life, then we have little reason to expect our current concepts to work in that possible world. Our concepts would be different.

---

[17] Kuhn, p. 252.

The Williams thought experiment meets this constraint. Although the thought experiment involves brain transplants—which are technically impossible—we can defend an expectation that our concepts should be able to handle this case. And why? Because the thought experiment simply brings to our attention an anxiety or confusion that already manifests itself subtly in everyday discourse. The concept already fails us in marginal cases in everyday life where one component of the cluster is missing, say an Alzheimer's patient who lacks psychological continuity because he can't remember who he is but who still exhibits physical continuity, an organ donor who lacks all higher cognitive functions and can't be said to "think" at all, or a pair of Siamese twins who share one connected brain but have two bodies. In each of these examples the normal application of our concept of the person is violated and our intuitions are conflicted. The Williams thought experiment does not just bring these conflicts to our attention, *it explains to us what the conflict is.* Before the experimental situation we are only aware that the concept does not always work, but we do not know why. The Williams thought experiment shows that we have competing ideas about what it means to be a person. Without the thought experiment the conflict stays in the background because bodies, minds and agency almost always go together in everyday life.

The task is then to suggest ways in which the concept—and the conceptual framework it represents—can be revised to better cohere with other facts about the world: I want to argue that this can be accomplished by realizing that 'person' is not one concept but a cluster of concepts brought under the umbrella of one term. For a long time this arrangement sufficed. But our uneasiness, Williams' thought experiment, and the entire history of the field of personal identity demonstrate that it no longer suffices. An analogy to a scientific case discussed by Kuhn might help.

Kuhn presents a thought experiment by Galileo that demonstrates difficulties with Aristotle's theory of motion. The experiment centers around our understanding of the concept "speed" or "faster". Aristotle took this to be a discrete concept, but only because he only considered a world with uniform motion. In our world, of course, not all motion is

uniform. For an object that is accelerating (or decelerating), its speed at a particular instant will be different from its speed as measured by the time it takes to transverse a longer distance, say several meters. Galileo's thought experiment exposed this difficulty by drawing our attention to a confusion which lingers implicitly in the real world. Kuhn draws an analogy between Galileo's thought experiment and the way Piaget brings this conflict to the attention of young children in his experiments on childhood cognitive development. Galileo's thought experiment carefully picks out an experimental situation where objects are accelerating, thus producing conflicting responses about which object is faster. The realization is an occasion to revise our concept of "faster" and our theory of motion. Galileo, having brought the conflict to the attention of his experimental subjects, suggests that our everyday use of the term "faster" actually masks two related yet distinct concepts: instantaneous speed and average speed. The thought experiment forces us to realize that there are actually two concepts here, not one.

So it is with the Williams thought experiment. It forces us to realize that there are multiple concepts here, not one. The analogies are striking. Both 'faster' and 'person' worked remarkably well for most situations. But concept users were anxious about both concepts because some information did not fit the accompanying conceptual framework. A thought experiment in both cases explained *why* the concept failed to cohere with our other intuitions about the world. And finally, the conflict is resolved once we recognize that both 'faster' and 'person' are cluster concepts that are composed of distinct component concepts.

§1.4 'PERSON' AS CLUSTER-CONCEPT

To review: we have identified that our intuitions about what it means to be a person are deeply conflicted. This was shown by Williams' work on self-concern. *Pace* Williams there are no overwhelming reasons to support one intuition over the other; his claim that the physical criterion is less vulnerable to an objection from branching fails to offer a positive reason for supporting one side over the other. The standard strategy of dealing with this

conflict is to refine our thought experiments with the hope that one of our intuitions will be decidedly rejected. With the help of Quine's meaning holism, I expressed skepticism about the validity of outlandish thought experiments, without rejecting the method altogether. Severe restraints ought to be placed on *Gedankenexperimente* in order to meet the Quinean complaint, although in doing so it appears less likely that a resolution to our conflict is on the horizon. Given the unlikelihood of a resolution, what should we do? Should we consider a change in strategy?

One strategy is a new and genuinely innovative reading of the conflict. One way to read the conflict is to see it as an incoherence in the concept of the person which would be resolved if we favored one of the intuitions over the other. This would give us a more precise concept to deal with. But there's another strategy. We can not only admit that we have these conflicting intuitions, *we can embrace them.* This is a strategy from common-sense, because common sense tells us that persons are biological animals, exhibiting continuity of consciousness and rational agency. In our day-to-day lives, these facts of personhood almost always go together, and it is difficult (if not impossible) to pick or prioritize between these facts. In normal, non-philosophical life, no prioritization is required. But one can always construct a thought experiment that peels away or separates these strains of personhood, with the goal of identifying the one that is a more fundamental—more necessary—element of persons. (This strategy has not yielded a resolution.) Following Quine, we should recognize that our concept of the person was designed with our daily needs in mind and for that the concept works rather well.

Let me suggest a new reading of the problem: our concept of the person is a cluster-concept. In a postscript to the anthology on personal identity that she edited, Amélie Rorty wrote that

> Some current controversies about criteria for personal identity, for characterizing and reidentifying human individuals, are impasses because the parties in the dispute have each

> selected distinct strands in a concept that has undergone dramatic historical changes; each has tried to make his strand serve as the central continuous thread.[18]

Rorty is suggesting here that our concept of the person has tracked different sub-concepts throughout time, serving different purposes as our needs required. These sub-concepts include 'heroes,' 'characters,' 'protagonists,' 'actors,' 'agents,' 'persons,' 'souls,' 'selves,' 'figures,' and 'individuals.' In her concluding section, Rorty writes that

> Our philosophical intuitions—the intuitions that guide our analyses of criteria for personal identity—have been formed by all these notions: they are the archaeological layers on which our practices rest. As is obvious, they are latently in conflict; if we try to be all of them, conceiving of each as having the final obligation over us, we shall indeed be torn.[19]

I want to borrow Rorty's suggestion that personhood is a cluster concept, although I wish to alter her suggestion in two significant ways. First, while Rorty's presentation is largely historical—she tracks the use of personhood's sub-concepts *over* time—I want to suggest that personhood tracks several sub-concepts *at* a time. It is not just that the term 'person' has meant many things to many people over a long period of time, it is my contention that the concept *continues* to mean many things to many people, and often it means many things to just one person at a single moment in time. This is a result of our conflicting intuitions that were exposed in the previous section. We have several things in mind when we think or say that someone is a person.

Secondly, I have no interest in dissolving the concept of the person into the sub-concepts that Rorty used as the basis of her discussion. Her exegesis was both historical and literary, and neither of those qualities serve my purpose here. I want to dissolve the concept of the person into the sub-concepts that were exposed as being at the heart of our conflicting intuitions about personal identity: the body, psychological continuity and connectedness, the brain and the mind, and agency. There might be more sub-concepts lurking within our notion of the person. I do not deny this. But these are the central concepts which concern me most.

---

[18] Amélie Oksenberg Rorty, "A Literary Postscript: Characters, Persons, Selves, Individuals" in *The Identities of Persons* (Berkeley: University of California Press, 1975), p.302.

[19] Rorty, p. 319.

What exactly does it mean to say that our concept of the person is a "cluster concept"? It is to say that 'person' is a colloquialism that does not belong in our ultimate ontology. It is to say that the term is a common shorthand for a collection of different philosophical and more premise—but sometimes cumbersome—concepts. Our use of this cluster concept has paved over our conflicting intuitions about its source, its use, its application. This paving over disguises deep disagreements about what it means to be a person and has led to cross-talking among those who would prioritize different strands of the concept to use as their criteria for identity over time. To say that personhood is a cluster concept is to say that the term sometimes serves its purpose but ultimately lacks something in specificity, i.e. that a philosopher concerned with accuracy ought to shun its use in favor of its more precise sub-concepts. To say that it "serves its purposes" is to say that it sometimes successfully tracks some of its component concepts. Furthermore, use of the sub-concepts in favor of the cluster concept promotes clearer analysis of the issues involved.

The claim here is that the term 'person' stands for a whole cluster of concepts about personhood—which is precisely why we have conflicting intuitions about it. We use the word 'person' to pick out these component concepts, though *which* concept we are picking out with the word depends on the situation. Sometimes we use the word 'person' to talk about agents. Sometimes we use the word 'person' to emphasize continuity of consciousness. Sometimes our use of the word 'person' coincides with biological animals. And often we use the word to highlight two or more of these discrete concepts. But the point is that these discrete concepts do not *necessarily* go together (as thought experiments have demonstrated); they go together *contingently*. It is an accidental fact of our existence that these facts go together. It might have been different. For our present needs, then, the word 'person' does remarkably well for us. It identifies a cluster of concepts that in quotidian discourse we have little need to distinguish between. We talk about persons when we want to identify 'people' in the vulgar sense of the expression. It does not matter if we are thinking about bodies, minds or agents, because people are all of these things. So the cluster concept works fine. But

when put under extreme pressure by a thought experiment that pits the components against each other, the word 'person' becomes problematic. That's because it does not stand for just *one* concept, it stands for a *cluster* of concepts that in the actual world are linked by contingent fact.

I need to say more about what I mean by the term 'cluster concept'. What exactly do I mean when I say that personhood is a *cluster* concept, as opposed to some other kind of concept? What does it mean for a concept in general to be a cluster concept? The term 'cluster concept' has been used widely to different effect. An objector might note that there are many harmless cluster concepts out there that need not be eliminated. And if anything in my account suggests that they ought to be eliminated (if, say, the same argument about personhood applies to them as well), then this can be taken as a *reductio ad absurdum* of my argument. For example, the concept of a forward in a hockey game is a cluster concept composed of three concepts: a left-winger, a center, and a right-winger. This is a cluster concept too and it is not in need of elimination, according to the objector.

This notion of a cluster concept is certainly not what I had in mind when I pointed out that personhood is a cluster concept. My idea of a cluster concept has some affinity with Wittgenstein 's characterization of the concept of a game: it has blurry edges and can be identified only by "family resemblance."[20] The relevant difference between the concept of the forward and the concept of person is that the components of the latter compete with each other. And by that I mean that our intuitions are genuinely conflicted about what it means to be a person and those conflicting intuitions tend to track the components. By contrast, we are by and large *not* conflicted about the concept of a forward. It is an uncontroversial exemplar of a concept with a disjunctive definition: a left-winger or a center or a right-winger qualify as a forward. The components here are *not* in conflict with each other.

---

[20] Ludwig Wittgenstein, *Philosophical Investigations*, translated by G.E.M. Anscombe (New York: Macmillan Publishing Co., 1958), §67-71.

In any case, it is important not to misunderstand the structure of the argument that is to come. The structure is *not*: (1) Person is a cluster concept; (2) all cluster concepts must be eliminated; therefore (3) the concept person must be eliminated. This summary does not adequately represent my argument because I am not making a general claim about all cluster concepts or the nature of concepts in general. My account is restricted to one *particular* concept, of which I am claiming that our intuitions are deeply conflicted. This is no surprise since the concept is a cluster concept, although this by no means entails that our intuitions are conflicted about all cluster concepts or that all cluster concepts stand in need of revision or elimination. My claim is specifically about 'persons' and its inadequacy as a concept.

The word 'person' is inadequate in our daily lives when we talk about special cases where the component concepts fail to go together. Consider a hospital patient who falls into an unrecoverable coma or a patient who is brain dead. But their bodies continue to function. One of the components is clearly present while the other is not. A biological animal continues to breathe (with mechanical intervention, but it breathes nonetheless) but there is no evidence of consciousness. We ask ourselves: Is this a person? Or consider a fetus. There is a body and some animal functions, but there is little to no psychological continuity and little to no agency. We again ask ourselves: Is this a person? Indeed, we torment ourselves with these questions because we take the answers to be morally relevant. In fact, we consider it the central job of morality to answer such questions.

However, these cases are exceptions and reside at the margins. This is important to remember because it explains why the term 'person' has existed for so long. If the Quinean point is correct, the term only has as much precision as our current needs have required. If the term 'person' had broken down in most cases, it would have been revised, altered or eliminated. It was not, which suggests that in the majority of quotidian uses, the term is adequate for our purposes.

But the mere fact that the concept breaks down in a minority of cases fills us with unease. And putting pressure on the term, through thought experiments, exposes our very

conflicted intuitions about what the term fundamentally means. I suggest the following explanation for how our intuitions could be so conflicted: the term 'person' does not stand for a single concept but rather a cluster of concepts that are unified contingently in our daily lives. And our intuitions track the component concepts. So when the respondent to the Williams thought experiment expresses self-concern for bodily continuity, her responses are tracking one component of the cluster. And when the respondent identifies with her psychological continuant, her responses are tracking another component. In daily life these components usually stick together, which explains why the concept has lasted so long. But in thought experiments—and marginal cases in real life—the components do not stick together. And it is for that reason that we must recognize that 'person' is a cluster concept composed of distinct sub-concepts.

It has been my contention that our intuitions about personhood are deeply conflicted. The explanation is that we are naturally conflicted because the object of our conflict is not a single concept but a cluster of concepts masquerading as a single and coherent philosophical idea. But are we any less conflicted about the component concepts? Aren't we conflicted about what it means to be a biological animal, to exhibit continuity of consciousness, or to be a rational agent? To continue an example from above, we have conflicting intuitions about the status of a fetus—independent of any issues of personhood. Is the fetus a separate biological entity from its mother or are they a single biological animal? And an entire field of inquiry is devoted to analyzing conflicts about the nature and behavior of rational agents.

My argument does not demand that I prove that the component concepts are free from vagueness, imprecision, or outright incoherence. All that is required is that the removal of the cluster concept in favor of its components improves—rather than impinges—the goal of coherence. It seems clear that our intuitions about personhood are deeply conflicted (to the point of being mutually inconsistent, as Williams demonstrated). If we eliminate personhood then we are left with our intuitions about physical and psychological continuity

and agency. This does not mean that our intuitions about the component concepts are free from conflict. Of course there are pre-existing conceptual problems within the component concepts, problems which are under investigation by the philosophy of biology, mind and action respectively. But in eliminating the concept of the person we would remove one source of the conflict—the conflict *between* the components. Why is this the case? Cluster concepts are ripe for internal conflict because often the components are at odds with each other. And while the component concepts might suffer from their own problems (local inconsistencies, vagueness, etc.), they are nonetheless *discrete* concepts. And discrete concepts, though by no means free from problems, are free from the kind of conflict that can plague a cluster concept.

Consider our intuitions about what it means to be a biological animal. The issue seems plagued with vagueness. Parfit's physical spectrum demonstrates that it may be indeterminate at what point a physical body ceases to be the same physical body. Consider a body that has an arm replaced—it is clearly still the same body. And if the body has all limbs replaced it is probably still the same body, though the question may be more contentious. What if almost all of its body is replaced? Finally, when all of the body is replaced we are likely to think that we are no longer dealing with the same body (depending on how you answer the Ship of Theseus thought experiment). We have intuitions about each end of the spectrum but in the middle of the spectrum we have difficulty answering the question. The difficulty is not with our linguistic competence with the concept, rather it is the concept's deficiency in fitting these marginal cases. A well designed and executed thought experiment can diagnose the specific deficiency. In this case, it is not that we do not know how to use the concept in the middle of the spectrum, it is rather that the concept is vague and in the middle of the spectrum the question is empty.

So the component concept is not perfect. But this is not evidence that person is not a cluster concept. It is not necessary that the components be free from vagueness or any other local inconsistency. The only requirement is that in recognizing 'person' as a cluster

concept we remove one source of inconsistency. This is certainly accomplished. The vagueness within the component concept is not *created* by recognizing that person is a cluster concept. The vagueness was already there implicitly; we are just now seeing it explicitly. So recognizing that person is a cluster concept improves the situation (by reducing inconsistencies), without necessarily making it perfect.

§1.5 PARFITIAN REDUCTIONISM

My claim that personhood is a cluster concept is very similar to Parfit's reductionism—though it is more radical. To *eliminate* the concept of the person is more than just *reducing* facts about personal identity to certain other facts, though the two positions are related. In the next two sections I hope to explain this point. But in order to do this I must first explain Parfit's reductionist claims about personal identity. I will then contrast the reductionist and eliminativist strategies with an example about the concept 'nation' in order to explain how these strategies differ with respect to the concept 'person'.

Parfit wrote in *Reasons and Persons* that identity can sometimes be indeterminate, especially during the middle sections of the spectra that he calls the physical, psychological and combined spectra. In such cases, the question of identity may be empty. I borrow Parfit's claim that such questions may be empty, but I also extend the claim to such cases where we must *choose* between competing accounts of personal identity—i.e. cases which highlight our conflicting intuitions. In claiming that "person" is a cluster concept, I claim that these questions are empty too. Since the first edition of *Reasons and Persons* (in which Parfit supported the Wide Psychological Criterion), Parfit has extended his claim about empty questions and said that we should not try to choose between the criteria. This brings his position closer to my claim that 'person' is a cluster concept. However, Parfit still believes that empty questions can sometimes have better or worse answers, and that in many cases the psychological criterion provides the best answer.

I need to explain how questions about identity can be empty and how empty questions can sometimes have better or worse answers. At one point in *Reasons and Persons*, Parfit makes a crucial distinction between two ways that a question might be empty. The first:

> About some questions we should claim both that they are empty, and that they have no answers. We could decide to *give* these questions answers. But it might be true that any possible answer would be arbitrary. If this is so, it would be pointless and might be misleading to give such an answer.[21]

And the second:

> There is another kind of case in which a question may be empty. In such case this question has, in a sense, an answer. The question is empty because it does not describe different possibilities, any of which might be true, and one of which must be true. The question merely gives us different descriptions of the same outcome. We could know the full truth about this outcome without choosing one of these descriptions. But, if we do decide to give an answer to this empty question, *one of these descriptions is better than the others*. Since this is so, we can claim that this description is the answer to this question [emphasis added].[22]

A reductionist believes that some questions about personal identity might be empty. However, the point of the second passage is that different descriptions might be more or less appropriate, even if the question is itself empty. That is, sometimes all descriptions are not created equal.

Parfit believes that questions about personal identity can sometimes be one of these 'empty' questions that he describes above. Parfit's argument for indeterminacy consists mostly of three thought experiments—the physical, psychological and combined spectra—that make heavy use of the Sorites paradox. In the Psychological Spectrum, a mad neurosurgeon is determined to change your psychological makeup by implanting beliefs, desires and character changes. There is a spectrum of such changes with nodes representing increasingly significant psychological changes created by the mad neurosurgeon. On the closest end of the spectrum, the neurosurgeon does nothing and your psychological profile remains unchanged. In the far end of the spectrum, the neurosurgeon replaces all of your beliefs, desires and character dispositions, thus making an individual who is fully

---

[21] Parfit, p. 260.
[22] Parfit, p. 260.

psychologically discontinuous with you. In the middle end of the spectrum, the neurosurgeon replaces half of your beliefs, desires and character dispositions, thus creating someone who is partly continuous, partly discontinuous with you. While it seems clear that the resulting person at the near end of the spectrum *is* you, and while the resulting person at the far end of the spectrum is obviously *not* you, the status of the resulting persons in the middle of the spectrum is unclear. Parfit suggests that there might be no borderline, no answer to this question in the middle region of the spectrum. Asking ourselves: 'Will the resulting person still be me?' might be an empty question. Questions about our personal identity can sometimes be indeterminate.

The same argument can be run with the physical spectrum, where a mad surgeon replaces increasingly larger portions of your brain and body with replicas. Each node on the spectrum represents a case where the surgeon replaces one more cell in your body. The result at the near end of the spectrum is someone whose body is exactly like yours (it remains unchanged), and the result at the far end of the spectrum is someone who looks like you, but who is actually a complete *replica* of you. Where is the borderline?

Both the physical and psychological spectrum make heavy use of the Sorites paradox. The spectra involve seemingly imperceptible changes—one cell for the physical spectrum, one belief or desire for the psychological spectrum—that lead to incompatible results. If the small changes are irrelevant, we must admit either that the result at the far end of the spectrum is still me, or the result at the near of the spectrum is not me. Both seem equally wrong. "We are led there," Parfit writes, "by what seem innocent steps, to absurd conclusions."[23]

The solution to the Sorites paradox here is to admit that the question is empty and that there is no sharp borderline on the spectra. Parfit makes a further claim: even if there were a borderline, we could never discover it. "Such a view is not incoherent. But it is hard to believe... We could not *discover* what the critical percentage is, by carrying out some of

[23] Parfit, p. 231.

the cases in this imagined spectrum."[24] No matter what, the person will *think* that he is me and will report his conclusions as such. There are therefore two interpretations of this—one epistemic, one metaphysical. One says that there might be a borderline, but we can never know it. The other says that there is no borderline *simpliciter*. The difference is rather minor, if you are a pragmatist, because a borderline that you can never discover is not relevant to human inquiry. However, Parfit seems to side with the metaphysical interpretation.

There are other cases where questions about personal identity can be empty. For example, Parfit claims that in a case of branching questions about identity are empty. There are different answers, but these different answers do not represent different outcomes. Rather, they represent different descriptions of the same outcome. That being said, one of the descriptions might be better than the others because it is less misleading. Parfit suggests that the best description might be that we do *not* survive branching but that everything that matters to us in normal survival is preserved in branching. But this is just a description—not a different outcome.

Another example of an empty question is what happens in a brain transplant case, e.g. as conceived by Shoemaker or Williams. An answer to this identity question forces us to choose one criterion of personal identity over another. As I stated earlier, Parfit now suggests that we should not force ourselves to choose because the question is empty. My claim that 'person' is a cluster concept suggests the same thing. Like before, though, Parfit prefers one description over another. In this case, Parfit thinks that the psychological criterion must accurately captures our intuitions. This is where my claim diverges from Parfit's account. I have claimed in this chapter that the psychological criterion captures some of our intuitions—but not others. So our disagreement stems from the fact that Parfit does not believe that our intuitions are as conflicted as I believe them to be. These conflicting intuitions are my best evidence for suggesting that personhood is a cluster concept.

---

[24] Parfit, p. 234.

An example might best demonstrate the difference between the eliminativist and reductionist strategies. Consider nations. Because of the arguments presented above, a reductionist about nations might claim that facts about nations can be reduced to certain other facts, say facts about citizens and their actions. This reductionist might claim that all facts about nations can be redescribed without making reference to nations at all. Instead the reference can be made to citizens on a particular piece of land, organized in a certain way, performing certain actions. A reductionist might also claim that because all facts about nations can be reduced to certain other facts, the identity of a nation need not be determinate.[25] If a nation splits in two there may be no answer to the question of which nation (or neither, or both) is the same as the pre-fission nation. That is because the fact of a nation's existence is not something above and beyond the existence of certain other facts. Although there *may* be an answer to the question of what happens to a nation when it splits, there doesn't have to be an answer and there isn't always one. This is what a reductionist about nations might believe.

In contrast, an eliminativist about nations would do more than claim that facts about nations can be reduced to certain other facts. She might claim that 'nation' is a cluster concept because it houses various sub-concepts. Our intuitions about what the word 'nation' refers to are deeply conflicted, and we often use the words in competing situations, as anyone familiar with the literature on nations and nationalism will tell you. Sometimes the word 'nation' is used to refer to a people, sometimes it is used to refer to a topologically distinct geographic region. Sometimes it is used to refer to a country under a certain form of government. Sometimes it is used to refer to a group of individuals who associate themselves by common ethnic, religious or civic ties. Sometimes the word 'nation' is used to refer to a purely political entity. In the literature on nations and nationalism, the word is usually

---

[25] Parfit argues that both persons and nations are distinct entities from the things that compose them, although he does not believe that they can exist separately from their components. In other words, both persons and nations are distinct from their components but they cannot exist apart from them, according to Parfit. But it is possible for a reductionist to hold the stronger view that persons and nations are neither separate nor distinct from their components .

defined at the outset so that the reader will be clear which notion the writer is making reference to. This might be more than a problem about our definitions, this might be a problem about our conflicting intuitions about nationhood, the eliminativist might claim.

Why do we have to decide whether a nation is a people, a state, a government, or a country, or some combination thereof? Why not stick to the component concepts? What hinges on deciding between competing accounts of nationhood? Why not just recognize that it is a cluster concept? We already have quotidian experiences that we can't quite assimilate into the conceptual framework of 'nations'. These are the marginal cases where the components do not go together—cases where peoples, governments, countries and states do not line up one-to-one but rather overlap. In these cases there may be nationalistic movements that seek to realign these concepts in a one-to-one fashion. In these overlapping situations, our 'nation' talk begins to break down. We are uneasy about our inability to adequately deal with such situations with the concept at hand. A thought experiment might explain our difficulty (by choosing an experimental situation that led to conflicting responses) and provide an opportunity for revision. This revision might include recognizing that 'nation' is a cluster concept that tracks several discrete component concepts. Although the component concepts are not free from problems, the conceptual landscape would be improved by eliminating the cluster concept. This is a stronger claim than the reductionist claim that facts about nations can be reduced to certain other facts.

An eliminativist about persons claims that personhood is a cluster concept just as the hypothetical eliminativist above claims the same about 'nation'—i.e. that the term tracks a variety of uses and meanings that reflect a manifold of constituent concepts. This is because we have conflicting intuitions about our concept of the person just as our hypothetical eliminativist claims that we have conflicting intuitions about the concept of the nation. Perhaps a second example will give color to the idea of a cluster concept and contrast this claim with Parfit's reductionism.

Cluster concepts appear in the medical sciences. Initially doctors and medical professionals are bombarded with a list of unexplained symptoms. On the assumption that these symptoms will turn out to be the result of a discrete cause, doctors create one label for this new syndrome. Patients are diagnosed with the new syndrome when they present with a certain critical number of symptoms from the list and the symptoms cannot be attributed to any other known disease. (Examples of this classificatory strategy include Gulf War Syndrome and Chronic Fatigue Syndrome. In both cases doctors and researches have a working hypothesis that the syndrome is the result of a single cause.) Research is then conducted in an attempt to identify the cause of the new syndrome. But sometimes the research indicates that the classification based on symptomatology has actually clustered together several *different* diseases with *distinct* causes. The diseases were originally clustered together under one cluster concept because the symptoms were similar and overlapping. However, once the components are identified, the cluster concept under which they are housed becomes ripe for elimination. The justification for the elimination is largely pragmatic: the classification of diseases should facilitate the identification of a cure—which is ultimately dependent on the cause of the disease. If the different components of the cluster have different causes—even if in most cases they have similar or overlapping symptoms—then eliminativism is an appropriate strategy. Although the components may seem to intersect on the manifest level (the symptoms), they widely diverge on a lower level (biological causes). This set of claims goes far beyond what a reductionist might claim: that facts about the disease can be reduced to certain other facts. In just the same way that a medical term might be a cluster concept with competing components, so too our concept of the person might be a cluster concept.

The conflict between Parfit's reductionism and my claim that personhood is a cluster concept goes back to his distinction between two ways a question can be empty. Even though a question is empty, it is still sometimes important to choose among different descriptions and different concepts. These descriptions are not arbitrary (we are not just

giving answers here), so some of them fit the facts better than others, although none of them, strictly speaking, are inaccurate. This is where my claim departs from Parfit's reductionism. We have conflicting intuitions about personhood because it is a cluster concept. So in the next chapter I will argue that we ought to consider eliminating it. This eliminativism will be plausible if we can show that it does no harm to our value theory.

That being said, it is important to remember that this disagreement is about our concepts. The disagreement takes place against our shared belief that regardless of which description is chosen to answer an empty question, we already know the facts and are simply arguing over the best description. I argue that our intuitions about persons are so conflicted that we are not dealing with just one simple concept but rather a whole cluster of concepts. I will argue in the next chapter that this makes personhood ripe for elimination. I will spend the remainder of the project arguing that despite the prevailing wisdom, the concept is dispensable. I will argue that moral and political philosophy, far from being impoverished by this elimination, will be strengthened and invigorated by it. That is why it is possible for us to be eliminativists about persons.

§1.6 THE ORIGINS OF THE CLUSTER CONCEPT

One final question remains before we proceed to the next chapter: how did one term—'person'—come to house such a diverse cluster of sub-concepts? While the question is a complex one that can only be answered by a thorough investigation into the history of ideas, a more cursory inquiry might contain a few revelations. The history of personhood is the history of a concept that has housed—from the beginning—diverse and potentially conflicting notions under a single umbrella. From the many varieties of personhood in Roman Law to the religious notion of a person as body and soul, the term was introduced to bring together competing components under a single conceptual rubric.

Let's start with Roman Law, the foundation for most European and European influenced legal systems. It was Gaius who noted that Roman Law could be reduced to

persons, things and actions.[26] The status of persons in Roman Law was complicated and seemed well-suited for the introduction of a cluster concept. The term had to be sufficiently wide to incorporate manifold distinctions: slave versus freeman and citizen versus foreigner, to say nothing of other dependent persons such as children and wives. Both slave and freeman were persons—although not equal persons—because both were the subject of rights and privileges. Although there were very few restrictions on the treatment of slaves—and even these were rarely enforced—there were nonetheless some restrictions. As such, slaves could not simply be *things*, because *things* are, in themselves, not objects of any legal or moral respect. So both slave and freeman were persons.

The etymology of the term person is then no surprise; it comes from the Latin *persona*. Before it became enshrined in the roman legal context described above, it signified an actor's role, character, or a mask worn in a play (as in *dramatis personae*). This move from the dramatic context to the legal context makes sense because the notion of legal personhood is essentially the same: the 'role' one plays within society and its political and legal structure. And of course one could *change* roles. Roman Law delineated the procedures under which a slave might become a freeman and vice versa. The roles were not permanent. So from the very beginning of the term 'person' there is an attempt to house competing components within a single cluster concept.

I suspect that the genesis of the cluster concept is also related to the religious dualism of body and soul and the demands of the Final Judgment. It is rather important to construct a theological doctrine such that the soul in the afterlife is able to receive rewards and punishments for our body's actions on earth. This is only possible if there is an identity relation between the soul and its earthly counterpart. As Locke puts it: "But in the Great Day, wherein the secrets of all hearts shall be laid open, it may be reasonable to think, no one shall be made to answer for what he knows nothing of; but shall receive his doom, his

---

[26] Gaius, *The Institutes*, translated by W.M. Gordon and O.F. Robinson (Ithaca: Cornell University Press, 1988).

conscience accusing or excusing him."[27] One way to explain how a heavenly soul can be responsible for the actions of its earthly counterpart is to subsume both components under a single cluster concept. So personhood becomes the concept that brings together two different forms of substance—to borrow Descartes' language—under one term. It provides a concept under which both body and soul reside. And it cannot simply be reduced to one or the other of its components. To reduce personhood to the soul would run the risk of denying the importance of our earthly existence (which is what distinguishes us from God and why we worship Him), and to reduce personhood to the body would run the risk of denying the importance of our heavenly destination (the ultimate goal). Hence the need for the cluster concept.

In both cases—Roman Law and Final Judgment—the concept of the person is used precisely because it is a cluster concept. It is capable of subsuming under one rubric an entire cluster of sub-concepts. At times those sub-concepts are at harmony with each other. But it is possible for these components to swing apart from each other. Indeed, the entire logic of the story requires it to be the case. The whole idea in the case of the Final Judgment is that one day a person will inevitably be a soul but not a body.

As a final note, it is important not to misunderstand the role this historical analysis plays within the context of the eliminativist strategy. The historical analysis of this section provides no justification for my prescription. Such prescriptions can only be justified by other means. Nor does the history of a word provide a justification for normative conclusions about the politics and culture of the present. This is the methodology of Nietzsche and Foucault, who both believe that the genealogy of words will yield great theoretical advances. I make no such claim here and this is not the methodology of this section. Rather, the historical account is offered simply as a historical backdrop against which to read my claim that the concept of the person is a cluster concept. Given the history

---

[27] John Locke, *An Essay Concerning Human Understanding*, edited by Alexander Campbell Fraser (New York: Dover Publications, 1959), Book II, Ch. XXVII, Section 22.

of the concept of the person, it should surprise no one that the theoretical analysis I have offered should produce the claim that it is a cluster concept. But none of this is *justification* for the philosophical prescription of elimination.

Regardless of whether this historical account of the concept of the person is correct, the factors that led to the concept's creation no longer apply. We no longer need the concept of the person to perform the same functions it performed in the past. And whatever those long lost functions were, they came with serious costs. I have attempted to show in this chapter that our intuitions are deeply conflicted about personal identity and that the concept is surrounded by serious metaphysical confusion. If eliminating the concept in favor of its components is a viable strategy, we ought to consider it. Investigating this strategy is the task of the second chapter.

I argued in the first chapter that a series of thought experiments by Williams revealed our conflicting intuitions about personhood. I argued that there were no positive reasons to accept or reject either of the intuitions. Following Quine I expressed skepticism about the methodology of thought experiments and concluded that this methodology is unlikely to resolve these conflicting intuitions in the future. I concluded that our concept of the person is a cluster concept that tracks a variety of component concepts such as biological human beings, psychological continuity and rational agency. Finally, I made use of Parfit's idea that questions about personal identity can sometimes be empty. I showed why a Parfitian reductionist might be willing to adopt my position while also showing how my position is nonetheless more radical.

Having proposed this new reading—that personhood is a cluster concept—we need a solution: eliminate the concept altogether. To some degree this strategy was already present in our diagnosis of the problem. The strategy of elimination was implicit from the beginning. Indeed if one wishes to attack the argument, the place to do it is elsewhere. Having accepted that personhood is a cluster concept, elimination is the obvious strategy to pursue, simply because no account of personhood should prioritize our animalist intuitions over our psychological intuitions or vice versa. So most of the argumentative work was done in the previous chapter and it is there that opponents of my position will rightfully direct their attention. The heart of Galileo's attack on the Aristotelian theory of motion was not his suggestion that speed ought to be eliminated in favor of the more precise notions of instantaneous speed and average speed. The heart of his argument was that our quotidian concept of speed was masking these more precise concepts and that for these reasons our quotidian concept would break down under experimental conditions that highlighted real-life problem spots. Having made this point, the road to elimination was clear, obvious and virtually pre-ordained. And so it is with persons.

This is not to say that this chapter is superfluous. Rather, it is to suggest that the major argumentative work being done, the task is to explain the nature of this elimination,

demonstrate how it is to be carried out and to evaluate potential objections and/or stumbling blocks. Having exposed our conflicting intuitions about personhood in chapter one and having explained the strategy of elimination in chapter two, the remainder of the project will be devoted to defending the viability of the strategy. In short, the viability is limited if conventional philosophical wisdom is true and our concept of the person is indispensable. Chapters three through five will be devoted to debunking this conventional wisdom and rehabilitating the viability of the strategy.

As for this chapter, I will first explain what is meant by "eliminativism" and what it will take to prove its viability. Having done that, I will draw a distinction between two motives for eliminativism—conceptual and ontological. Roughly, the former is motivated by the increased utility to be gained by eliminating the concept of the person from our conceptual apparatus. The latter is motivated by a misconceived attempt to settle ontological questions first and then alter our concepts accordingly. I will reject this second procedure for the simple reason that we have no concept-independent access to ontology. So we will stick with the first methodology and investigate whether eliminating the concept of the person might be a useful move. This chapter will then outline the future of personal identity after elimination, i.e. how the identity question will be answered in a conceptual framework without persons.

Next, three categories of objection will be presented. The first, which I borrow from Parfit, is the Argument from Below. The Argument from Below says that when a set of facts is just composed of some lower level facts, it is the lower level facts which are significant and the higher order facts are significant only insofar as they are composed of the lower facts. But the significance resides at the bottom and migrates up. The Argument from Below is challenged by Mark Johnston, who advocates the Argument from Above. This argument suggests that the importance flows from the bottom up. The Argument from Below lends significant help to my thesis by explaining, on a theoretical level, how a concept as pervasive as personhood could be dispensable. Simply put, the concept is dispensable because it has no

intrinsic importance—its only importance is derivative and flows from its components. And these components remain after elimination, thus explaining—in the abstract—why the strategy is harmless. Johnston's objections (via his Argument from Above) will be evaluated. Although inspired by legitimate Quinean concerns, these concerns are either misplaced or irrelevant to the particular case of the concept of the person.

Second, I will remove two potential obstacles to my proposal. I will show that eliminativism can be consistent with a claim made by Wiggins that 'person' is a natural kind term. While not conceding Wiggins' argument about natural kind terms, I will claim that at least one of the component concepts meets Wiggins' criteria for natural kind status and that this is enough to fulfill his project of sortal concepts. I will also consider Strawson's claim that the concept of the person is primitive. My strategy here will be to endorse his rejection of Cartesianism and qualify my eliminativism in response to this rejection. The result will be a clearer picture of my proposal: I am not claiming that we are hybrids of multiple entities of different kinds. Rather, I maintain that personhood—as a *concept*—should be broken up into its components.

§2.1 CONCEPTUAL VERSUS ONTOLOGICAL ELIMINATIVISM

Before we continue, an important question must be settled: to what degree, if any, is conceptual eliminativism ontological? Or put another way, does eliminating the concept of the person carry ontological commitments? The question is tricky because it involves the general relationship between concepts and reality, a contentious issue which cannot be fully explored here. The best way to answer the question is to review and evaluate our *motives* for eliminating the concept of the person. In so doing we might determine the ontological import of eliminating the concept of the person.

One could imagine two different motives for eliminating the concept of the person. In one case, the ontologist might decide that persons do not exist in the fullest ontological sense of that expression, and in order to bring our conceptual apparatus in line with that

ontological reality, we ought to eliminate the concept of the person. It being the goal of our conceptual apparatus to represent the world, it would be misleading to retain a concept whose corresponding entity did not exist. One might call this variety of eliminativism *ontological* because its motivation stems from a pre-existing determination that persons are not part of the ultimate furniture of the universe.

But consider a second motive for eliminating the concept of the person. Given that we have no concept-independent way of investigating the ultimate furniture of the universe, it would be absurd to claim that we can arrive at some intuitive ontological assessment of reality first and *then* make our conceptual framework fit that reality as closely as possible. How could we investigate our world without the very concepts that make up our conceptual apparatus? It being impossible, the choice of which concepts to use must depend on utility, on the concept's success in investigating the world and pursuing inquiry in all of its forms—scientific, moral, political, etc. This is the motivation for eliminating the concept of the person. The *conceptual* eliminativist engages in precisely this kind of utility calculation. She evaluates the irresolveable deadlock in the debate about personal identity and recognizes that the conflict stems from the concept's status as a cluster concept and its attempt to house diverse and sometimes contradictory components under a single umbrella. Furthermore, the eliminativist suspects that eliminating the concept, far from harming our value theory, might improve it. It might do so by rendering explicit the conflicts among the components of the cluster. Indeed it is possible that these metaphysical conflicts manifest themselves in value theory as well. So eliminating the cluster concept in favor of its components might resolve some of these difficulties. This variety of eliminativism we might call *conceptual* because its motivation stems from the utility of our conceptual apparatus.

So to what degree is conceptual eliminativism ontological? It betrays an ontological commitment in the sense that eliminating the concept of the person commits us to the ontological claim that persons do not exist. This is not to say, of course, that individuals walking around on the street are mere fictions, although this is a trivial point offered mostly

to combat naive skepticism from the non-philosophical public. Indeed, we can still make reference to individuals who are traditionally referred to as persons, just with different concepts. But insofar as the concept of the person has been eliminated, it would be correct to make the ontological claim that persons do not exist.

It is important to understand that claim within the context of the methodology. It is to be distinguished from the claim of the ontological eliminativist, who is committed by a pre-conceptual evaluation of the world to the claim that persons do not exist. The difference between the two views is that the conceptual eliminativist denies that such a pre-conceptual evaluation is possible, so the choice of concepts must be made on alternate grounds—the utlity of the conceptual framework and the success of that framework in carrying out investigation. But this choice then carries with it an ontological commitment, such as it is, because one's conceptual framework is indeed tethered to the world. In this sense conceptual eliminativism is *derivatively* ontological, in contrast to the direct connection in the former methodology. Or put another way, conceptual eliminativism is ontological in the only appropriate way, insofar as the competing methodology proclaims a concept-independent access to the world which strains credulity.

Distinguishing conceptual eliminativism from ontological eliminativism is crucial for a reason: making explicit the appropriate burden of proof. When deciding what kind of concepts we ought to employ, utility is a key barometer. We want concepts that do their job well and allow us to investigate the world, bur there are competing possibilities. So the question is: which concepts work better? Which concepts allow us to construct a successful conceptual fabric? When a concept or a group of concepts fails on that front it is a signal for change.

If it turned out that a given concept was indispensable, this would pose an insuperable obstacle for conceptual elimination. For that very reason the rest of this project will explore whether concepts such as self-concern, responsibility and rights can be analyzed without the concept of the person. So the utility of the elimination, if demonstrated, will be

sufficient to show that the change in our conceptual framework is warranted. We will then be free to use conceptual elimination to resolve the problem of our conflicted intuitions.

§2.2 IDENTITY AFTER ELIMINATIVISM

It might seem as if eliminating the concept of the person makes things more complicated, not less. In one sense this is true. Our conflicted intuitions about the concept of the person were exposed during an inquiry into the nature of personal identity. Eliminating the concept of the person serves to take the 'personal' out of personal identity. How can we find an answer to the question of personal identity without even making reference to persons?

In one sense we cannot. But in another very important respect we can. Elimination will not solve the puzzle of personal identity precisely because it rejects the major premise around which the inquiry is structured. Elimination changes the terms of the debate about personal identity. It urges us to recognize that we've been asking the wrong questions.

After recognizing that 'person' is a cluster concept composed of several discrete component concepts and after having eliminated the cluster concept in accordance with that realization, we are left with a proliferating identity question. Where once there was one identity question, now there are several. Furthermore, there is no reason to believe that the answer will be similar in structure for each component. But what are the components? At the moment I want to leave the question open a bit. The question will be pursued in greater depth in section §2.7. But as a working assumption we can look at our conflicting intuitions to get a sense of what the components will be. Because our conflicting intuitions in the Williams thought experiment are physical and psychological, it is probable that two of the components will line up in some way with these intuitions. That's because the components are responsible for our conflicting intuitions.

An identity theory will be required for every entity that we pull out of the cluster concept. For example, I have identified that one of the components is physical continuity of

the biological body, so an identity theory is required for body counting. Although I have identified psychological continuity as a potential component, I have *not* cashed this out as some kind of pure psychological subject such as a Cartesian ego, as will become clearer later in this chapter. Consequently, psychological continuity is not to be construed as an entity requiring identity criteria, though it may nonetheless require definition.

Finally, it might also be the case that other components—say agency—are identified as we investigate the eliminativist strategy. If agency turns out to be a component concept, its relation to the other components will be considered. For the moment I have left open the possibility that agent identity might diverge from physical or psychological continuity. In the next chapter I will question whether there is any reason to reject on logical grounds alone a definition of agency that does not presuppose phenomenological unity of consciousness. If there is no such reason, it might be *coherent* to claim that agent identity can diverge from animal identity and psychological continuity. Potential cases where agents might not come one to a body include group agents and multiple personalities. If indeed it were shown that agent identity could diverge from the other components we would need a new account of agent identity that we could use for agent-counting. (However, it is important not to misconstrue this as a claim that we are hybrid entities much as a Cartesian claims that we are composed of material bodies and pure mental egos. A better analogy would be Locke's distinction between animal and personal identity—a distinction which is metaphysically similar to the distinction between the possible components I have just discussed.) In any case, though, the point at the moment is simply that the concept of the person is a good candidate for elimination in favor of components that line up roughly with our conflicting intuitions in the Williams thought experiment.

I have tentatively identified the following components stemming from the conflict in our intuitions: biological human beings and physical continuity, as well as psychological continuity. This list is not necessarily exhaustive and there may be more components—perhaps rational agency. I will reconsider this issue more closely in §2.7. But regardless of

the number of components identified, this much is sure: instead of personal identity we now have identity questions to be answered for each component. For those components that represent entities, identity criteria need to be established and there is no reason to think that we can establish a template by giving identity criteria for, say, biological human bodies, and then applying the template to another component like 'agent'. The reason for the difficulty is simple: biological animals and agents may be different kinds of entity that require different identity criteria.

It might seem as if eliminating the concept of the person was a poor bargain. The puzzle of personal identity has expanded and the situation seems more complicated after elimination than it did before. If this is an accurate assessment it might prove to be a fatal argument against elimination—since at least part of the argument appeals to utility.

The assessment is not, however, accurate. Although we are burdened with finding identity criteria for perhaps multiple entities—instead of one—this does not mean that the situation is more complicated. The ease or difficulty of the task is not to be measured by mere addition. Recall the original difficulty: Williams' thought experiment demonstrated that we are unable to provide a coherent and stable account of personal identity because of our conflicting intuitions about what it means to be a person. We have diagnosed the source of the conflict: personhood is a cluster concept composed of components that can diverge. We can resolve the conflict by eliminating the cluster concept and renewing the identity question at the level of the components.

But this second attempt—after elimination—is easier because the source of the conflict has been eliminated. We are free to explore the metaphysics of biological bodies, psychological continuants and agency without the hindrance of our deeply conflicted intuitions. One source of conflict will be eliminated. Although we may be conflicted on other levels—i.e. confusions about psychological continuity, individuation of biological animals, criteria for agency, etc.—we have still removed one source of conflict. And it is not as if we created the conflict on the other levels simply by eliminating the concept of the

person. Those conflicts were there in the first place. We just remove conflict on one level. This is a net gain.

There will be conflicts in the lower level concepts, to be sure. For example, developing a theory of identity for agents will not be simple. Our intuitions on the subject are contaminated by some confusion. Must agency be confined to a single biological entity or might agency be realized across several biological entities? Is there such a thing as a group agent? Could there be multiple agents within a single biological body? Can there be overlapping agents? Furthermore, must agency be realized in a biological body or might there be synthetic substitutes capable of exercising agency? How should we individuate agents and decide where one agent ends and another begins? In real life we take our cues from the biological body because we assume that agents come on to a body. But that assumption is just that—an assumption—and requires justification, for it seems at least logically possible that the identity of agents need not coincide exactly with the identity of biological bodies.

Similar problems might present themselves when we consider biological bodies. Williams takes this to be uncomplicated because he thinks that some kind of suitably unpacked notion of spatio-temporal continuity will do the trick. But age-old questions revealed by the Ship of Theseus thought experiment will need to be answered. When spatio-temporal continuity and composition of matter are put at odds in a well constructed thought experiment our intuitions are conflicted. In real life composition of matter and spatio-temporal continuity usually go together so we aren't put into a position of ranking these criteria. But in marginal cases—and thought experiments—we are called upon to do so and we sometimes offer inconsistent responses. An identity criterion for bodies will have to contend with this problem.

All of the components will face problems associated with branching, as should be obvious by now. Despite Williams' assertion, both psychologist and animalist accounts of personal identity are vulnerable to objections from branching and fission. This means that

each of the *components* in the cluster will be vulnerable to branching. But this should not count as evidence against my strategy of elimination and it is important not to confuse these difficulties with the difficulties associated with our conflicting intuitions about personhood. After elimination the field of personal identity will be more precise and theorists will be better equipped to develop identity criteria for the remaining components. And these efforts will not face the deep conflicts of intuition that reside in our concept of the person. This does not mean that the we will be left with mere mopping-up operations. There will be substantive philosophy ahead. But while before there was no possibility for a resolution, now there is at least the potential for an answer. In theory we are closer to a resolution because we have removed the roadblock. But we must now make the journey.

I will not attempt to offer comprehensive accounts of the component concepts (though I will attempt to give a more definitive defense of my *list* of component concepts in §2.7). Each one requires an in-depth treatment and such treatment is beyond the purview of this project. Indeed, some have already received this treatment in well established fields of philosophical inquiry, including the philosophy of biology, mind, and action, respectively. The point of my eliminativist strategy is that the answer to personal identity—or the only answer that is possible—will come in these fields, not from the field of personal identity itself. Splitting the question up and pursuing it in different contexts is the strategy most likely to yield success. The burden is not on me to provide complete accounts of the component concepts. Rather, my burden is simply to demonstrate that the strategy of elimination will clear the path so that an inquiry into identity can proceed.

§2.3 THE ARGUMENT FROM BELOW

Part of Parfit's argument for Reductionism is the so-called Argument from Below. Parfit has argued that "personal identity just consists in certain other facts" and that "If one fact just consists in certain others, this fact cannot matter in itself. If this fact matters, its importance must be derived from the importance of the facts in which it consists." These

two claims lead to a third: "personal identity cannot be in itself important. Any importance it may have must be entirely derivative."

There are several examples with which to explain the Argument from Below, taken from subject areas other than personal identity. Being married consists in certain lower level facts such as a being in love, having a committed romantic and/or sexual relationship, living together in a common home, the intention to raise a family together, the emotional and financial intertwining of lives, etc. When considering such things as spousal rights (such as inheritance or medical insurance for dependents), many would argue that it is the lower level facts which matter. In other words, because marriage just consists in certain lower level facts like living together in a committed romantic relationship, those who fulfill these facts should receive these rights even in the absence of a marriage certificate. When states recognize common-law marriages in the absence of a marriage ceremony, they are appealing to an Argument from Below. The assumption behind common-law marriages is that the lower level facts are important and the higher level fact of getting married is significant only insofar as it usually consists in these lower level facts.

Parfit argues that personal identity is similar. Just as being married just consists in certain other facts, so too personal identity just consists in certain other facts. In the former case these other facts are that two people live together in a particular domestic arrangement. In the latter case these other facts are physical and psychological continuity. In both cases, it is surely the lower level facts which are morally and rationally significant. Just as it would be absurd to ignore those lower level facts in the absence of a marriage certificate (as one might be tempted to do if one looked solely at the higher level fact) so too it would be absurd to think that personal identity has an intrinsic importance over and above the facts that constitute it (such as physical and psychological continuity).

Parfit uses his Argument from Below to support his reductionist claim that personal identity is not what matters—what matters is the physical and/or psychological continuities. In branching cases where identity is not preserved these physical and psychological

continuities remain and *that is what matters anyway*. And this is true not just in branching but *always*. The Argument from Below is important for my argument as well, although for different reasons. In arguing that the concept of the person ought to be eliminated, one of the potential roadblocks is that the concept is indispensable because it has moral or rational significance for, say, responsibility, self-concern and rights. One way to remove the roadblock is to appeal to the Argument from Below. Being a person just consists in certain other facts and it these facts which carry rational or moral significance for things like responsibility, self-concern and rights. So eliminating the higher level concept is a plausible strategy since it is the remaining lower level facts that are important. So demonstrating the Argument from Below goes a long way towards showing that the strategy is plausible.

Johnston has argued that Parfit's Argument from Below leads to nihilism. The argument runs as follows: If physicalism is true, then all facts about the world just consist in facts about the fundamental particles of the universe. Certainly, the fundamental particles—and their properties—do not matter in themselves (independent of the things that they constitute). ("Particles are particles," one might say). And since physicalism says that everything is reducible to those fundamental particles, then nothing matters at all—hence nihilism. "But this is not a proof of nihilism," Johnston writes. "It is a *reductio ad absurdum* of the argument from below. We should not expect to find the value of things we value divided out among their constituents. That is to say, the argument from below depends upon a fallacious addition of values."[28]

Regardless of the truth of physicalism, the objection is a good one. Parfit points out that there is a distinction between "reductionism about what exists and reductionism about facts" (much as there is a distinction between conceptual eliminativism and ontological eliminativism.) Parfit's point is that the former does not entail the latter; while everything is

[28] Mark Johnston, "Human Concerns without Superlative Selves" in *Reading Parfit*, edited by Jonathan Dancy (Oxford: Blackwell Publishers, 1997), p. 168.

composed of particles it might not be the case that all *facts* can be reduced to facts about particles.

The appeal to the Argument from Below is not universal because Parfit is not claiming that "whenever there are facts at different levels, it must be the lower level facts which matter." However, he is claiming that *merely* "conceptual facts" cannot be intrinsically important. For conceptual facts, what matters are the lower level facts which constitute them. And merely conceptual facts are affairs of the descriptions we choose to apply to the world. To ask a question about a merely conceptual fact is to ask a question about how a concept or term is supposed to be used, not a question about reality itself. Parfit claims that personal identity is just such a conceptual fact because it just is the holding of physical and/or psychological continuity. And furthermore, I would claim that being a person is just such a conceptual fact because all it means to be a person is to be a biological animal/human being and/or conscious subject exhibiting psychological continuity and/or a rational agent. Being a person just is these things. In short, we can adapt Parfit's Argument from Below (which supports his claim about the unimportance of personal identity) to support my eliminativist strategy about the concept of the person.

My Parfitian Argument from Below runs like this: being a person just consists in x (where x is being a human being and/or a conscious subject with psychological continuity and/or a rational agent). Since being a person just consists in x, then being a 'person' is a mere conceptual fact. Consequently, being a person cannot be intrinsically important. Its importance is derived purely from the importance of x. This Argument from Below could be challenged with a Johnstonian Argument from Above. While we argue that being a person is important only because it consists in x, a Johnstonian might argue that x only matters because it constitutes being a person.

Consider an example. This is not a radical thought experiment but a real-life example from the margins. An accident damages my brain leaving me irreversibly unconscious, though my body continues to function without interruption. I continue

breathing, my heart continues to pump and my blood continues to circulate. I am unable to exercise rational agency, nor do I exhibit psychological continuity. However, my status as a biological animal is unchanged. Have I died? Am I still a person?

On my view, being a person is a mere conceptual fact because it consists solely in the lower level facts described above. An answer to the question is a claim about the proper application of the word 'person' and is not a claim about reality. That's because we already have all of the information that we need. And it is these facts which are important and not whether the term 'person' can be applied in this case. What matters, as Parfit says, is reality and not how we describe it. The Argument from Below clearly supports our intuitions about this case.

In marginal cases such as these, we often ask ourselves whether this is a person and we take these discussions to be about something more than just the proper use of a term. We consider these questions crucial and worthy of civic debate. And they are. But that is because we often ask if something qualifies as a person as shorthand for asking whether something is an appropriate object of moral concern. Having antecedently *assumed* that it is persons that are the appropriate objects of moral concern, we frame our questions in terms of the concept 'person' as a way of deciding how we should treat someone. If we think that someone deserves moral treatment, we call him a person. If we do not, then not.

But it is important not to conclude that any intrinsic importance derives from our use of the label 'person'. What is important is reality, not how we describe it. The different answers (yes he is a person or no he is not) are not different possibilities, different ways the world might be. The difference is merely conceptual. And because the difference is conceptual, the Argument from Below applies and the importance flows upward. Personhood is important only because it constitutes being a biological human being, a conscious subject with psychological continuity, and a rational agent. But the importance derives from these lower level facts. If we know these lower level facts, we know everything that matters for morality and rationality. Any further question about 'persons' is a question

about the appropriate use of a term. This supports my contention that we could do away with the concept without doing irreparable harm to our discourse. Since the importance flows from the lower level facts, eliminating the higher level term would pose no problem.

As I stated before, Johnston rejects the Argument from Below in favor of the Argument from Above. For example, one might claim that it is the *official* act of getting married which lends significance to the lower level facts of living together in a committed, romantic relationship. The official act is important because it codifies the relationship publicly and the lower level facts of marriage derive their importance from above. Also, one might argue that the lower facts of being a human being, a conscious subject and a rational agent are only important because they constitute being a person and their only importance derives from this higher level fact. I think this is less plausible in both examples. However, Johnston's inspiration for rejecting the argument from Below is severely Quinean. Given the degree to which I invoked Quine's skepticism in the previous chapter, it is important to evaluate Johnston's use of Quine and determine if we ought to modify—or even reject—the Argument from Below.

Johnston's argumentative strategy is a so-called quarantine maneuver. He suggests that in cases where the "determinacy of personal identity is not guaranteed" the proper response is *not*, as Parfit does, to conclude that personal identity is not what matters; rather the proper response is to extend and appropriately modify our concept for these radical cases which violate the presuppositions upon which the concept is derived. This "quarantines" the marginal cases derived from extreme thought experiments by denying that they pose any problem for our normal concepts. Instead of being proof that our concepts must be thrown out even in regular cases, Johnston's quarantine maneuver urges us to make a *local* modification to our concepts to handle the radical case in question. There is no need to make a universal modification to our current doctrine in response to marginal cases.

The quarantine maneuver is deeply Quinean on several fronts. First, it is inspired by Quine's maxim of minimum mutilation. According to this edict the severity of one's

adjustments ought to be commensurate with the severity of the problem. A minor problem demands an equally minor amendment to one's current doctrine. According to Johnston, the problems exposed by Parfit's thought experiments about division are minor and require only local modifications to our concepts—not wholesale revision. For example, Johnston rejects Parfit's argument that we should move from self-concern to R-variant concern. Instead, Johnston argues that garden-variety self-concern is justified in normal cases—and if division ever happened we would be justified in switching from self-concern to R-variant concern in this bizarre case. But we would not be justified in moving to R-variant concern in *all* cases.

Why should the problems encountered in division be considered minor? Again the inspiration is Quinean. As discussed in the previous chapter, Quine pointed out that our concepts are tailored for our current needs. So our concepts are designed under various assumptions about the way the world is. When these basic facts are violated it is no surprise that our concepts crack under the pressure. But unless these basic facts are violated *permanently*, there is no reason to make permanent amendments to the concept in question. If the basic facts remain standing in all but the extreme cases, the maxim of minimum mutilation calls for a local change. Such a local change might include extending the concept or practice in the appropriate way for our bizarre case. In the case of Parfit's division, a local change is sufficient for a bizarre case. If division were to become as common as sexual reproduction, a wholesale change would be warranted. But because it is not common, personal identity is almost always present. And therefore personal identity is important. By this argument Johnston quarantines the marginal cases derived from thought experiment. He puts the point this way:

> The very case of fission itself undermines essential unity, violates the presupposition that 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. Thus no one is in a position to appeal to this last principle in the fission case. The plausible basis of the principles is undermined in the very description of the fission case.[29]

[29] Johnston, p. 169.

The voice of Quine can be heard clearly in this passage and it poses a problem for my appeal to the Argument from Below. I have already accepted Quine's point that concepts are custom designed for the actual world, so it seems as if Johnston's quarantining maneuver might be used against my eliminativist strategy. Here's how the argument would work:

Our concept of the person works fine in normal circumstances. We usually have no difficulty picking out persons from non-persons. If there are marginal cases—in, say, thought experiments—where the concept does not work well, then we have a warrant for a local modification to our doctrine. In radical cases we might extend or perhaps amend our concept of the person to fit a case which violates the presuppositions upon which the concept was founded. But there is no warrant here for a wholesale revision to our doctrine, nor do we have a warrant to eliminate the concept altogether. This elimination would violate Quine's maxim of minimum mutilation because the response would be out of proportion to the problem. The breakdown of the concept of the person in extreme circumstances is no evidence that the concept does not work well in most situations (indeed almost all) and consequently the extreme cases are no evidence that the concept is inadequate. If the concept is inadequate in extreme circumstances when its presuppositions are violated, it ought to be extended and modified locally, not revised or eliminated universally. In extreme cases where we have difficulty applying the term 'person' we ought to extend our use of the term to fit the bizarre case. We might even amend the concept or invent a new one for the occasion. But throwing out the concept for all cases is an extreme measure in violation of Quine's maxim.

My response to this objection is to appeal to the relevant differences between Parfit's case of division (and his argument about the unimportance of personal identity) and my concern about the coherence of the concept of the person. While the problems associated with division are indeed marginal cases, the conflicted intuitions that I exposed in my previous chapter are symptomatic of our conflicted intuitions *in every instantiation of personhood.* The function of the thought experiment was not to merely invent a marginal

case where the concept broke down but was rather to construct the scenario that would reveal the shortcomings of the concept that were already present in the regular cases. The thought experiment did not invent a scenario which would pose a problem for the concept, the experiment invented a scenario that would *identify and explain* the problems—problems which predated the invention of the thought experiment. Simply put, the problems were there to begin with.

The problems that predated the invention of the thought experiment are our conflicted intuitions about personhood. Although in the main we are certain when the concept 'person' applies, we are not at all certain *why* it applies. While that may not seem to matter much because the component concepts almost always go together, it reveals that the difficulty with the concept is not just confined to the marginal case. Ask anyone whether someone is a person and the answers will be near-uniform. Ask anyone *why* or *by virtue of what fact* someone is a person, and there will be as many responses as there are respondents. And in some cases there will be conflicting responses from one and the same respondent. And that is because our intuitions are deeply conflicted about the concept of the person. Furthermore, it would not do justice to our conflicting intuitions to conclude that the conflict resides solely in the marginal cases. Nothing could be further from the truth. Although the conflict *is most apparent* in the marginal cases, the conflict is ubiquitous and pervasive. The thought experiments not only demonstrate that we aren't clear about what it means to be a person in the marginal cases, we aren't sure what it means in the quotidian cases either. Therein lies the relevant difference between Parfit's case of division and the unimportance of identity and the conflict about personhood that I have identified.

This leads to a second objection. You might ask what the problem is in normal cases where the components of the cluster coincide. My response has been that although in such cases this is uncontroversially a person, it is unclear *why* this is a person, i.e. by virtue of what fact or component it is a person. This anxiety is similar to the anxiety presented by those metaphysicians who cling to the doctrine *no entity without identity*. Although I am not

endorsing this view, the anxiety is the same. You don't know why this thing is a person because you cannot offer a coherent theory of personal identity without violating a major intuition in the process. But the objector might continue: we can offer such a theory about what constitutes a person and that theory is a disjunctive answer which lists the properties common to the class of beings picked out by the term 'person'. But this doesn't say much and Wittgenstein's words apply here: "But if someone wished to say: 'There is something common to all these constructions—namely the disjunction of all their common properties'—I should reply: Now you are only playing with words. One might as well say: 'Something runs through the whole thread—namely the continuous overlapping of those fibres."[30]

Now the objector might respond that she had nothing so vacuous in mind: the disjuncts to which she meant to refer were the components of the cluster. The problem with this line is that while previous accounts of personhood seemed too constrictive (where some marginal cases were concerned), this disjunctive definition of personhood swings to the other side of the spectrum: it is far too permissive. By allowing 'persons' with only one disjunct to qualify as persons, the disjunctive account would inevitably violate our intuitions exposed by the Williams thought experiment. Identifying a 'person' with only the animalist disjunct would violate our psychological intuitions; identifying a 'person' with only the psychological disjunct would violate our animalist intuitions.

What other local modifications to the concept might be appropriate? One might argue that the appropriate local modification would apply the term 'person' when any of the component facts are missing. This local modification would extend the term to cases when, say, agency or psychological continuity is missing from the mix. While this might be an adequate stop-gap measure, it would do nothing to alleviate the more fundamental problem: our conflicting intuitions in the regular cases. These would certainly remain. And these

---

[30] Ludwig Wittgenstein, *Philosophical Investigations*, translated by G.E.M. Anscombe (New York: Macmillan Publishing Co., 1958), p. 32e.

intuitions are the source of the problem with the concept of the person. The conflict explains why the moral and rational discourse emanating from the concept of the person has been fraught with such disagreement and difficulty. It is precisely because there is this basic conflict about what it means to be a person that many issues of deep moral import have remain unresolved. A local amendment to the concept, initiated to help deal with the marginal cases, will do nothing to clean up the concept in the regular cases where the problem yields side-effects in the companion discourses that take their foundation from the concept of the person. So even by the lights of Quine's maxim of minimum mutilation, a wholesale revision of our concept of the person is warranted. Furthermore, given that the importance flows upward from the component concepts and the lower level facts, little will be sacrificed by eliminating the higher level concept.

Of course, I must concede that eliminating the concept of the person would violate an intuition of its own—our belief that persons exist and that personhood is a concept of moral, ethical and legal significance. There is no doubt that a revision of common-sense is required here. So eliminating personhood might be considered a "mutilation" that goes beyond what Quine's maxim would allow. But it is important to remember that more than just intuitions are relevant in the calculation. Although eliminating the concept of the person will require violating a major intuition, it might provide a clearer conceptual landscape within which to pursue our value theory. There is *prima facie* reason to investigate in this direction. If the concept of the person is racked with internal conflict, as we have demonstrated here, then it is reasonable to suspect that personhood has provided an imperfect landscape to pursue our value theory. By eliminating the concept and replacing it with component concepts we might not only eliminate the internal conflict but actually improve the conceptual machinery at our disposal for value theory. And this would be the real justification for eliminativism—showing that the price paid by violating common sense was outweighed by an improved understanding of value theory. The argument would then

come down to utility: eliminating the concept of the person would be better for us. The rest of this project is dedicated to seeing if this is really the case.

To review: I have suggested that the Argument from Below supports my eliminativist strategy because the argument demonstrates that any importance from the concept of the person is derivative and not original. This helps my argument because it supports my position that the concept is dispensable. I borrowed Parfit's claim that the Argument from Below does not apply to every case where some fact is composed of lower level facts. Rather, the Argument from Below applies in cases where a fact is "merely conceptual" because questions about the higher level facts are about our concepts, not about reality. Intuition, I believe, supports my contention that 'person' is one of these merely conceptual facts, although I admit that it would be coherent to claim that 'person' is not merely conceptual and that its importance flows top-down and is an exemplar of the Argument from Above. Although this position is coherent, mine seems more plausible. What matters is the lower level facts about being a biological human being, a conscious subject with psychological continuity, and a rational agent. Once we know these lower level facts, we know everything about reality. Any further question is merely conceptual—i.e. a question not about different ways the world might be but different ways we might describe the world. And as Parfit says, what matters is reality—not how we describe it. Lastly, I have suggested that although Quine's maxim of minimum mutilation is a justified constraint, it does not support a mere local extension to our concept to cover the extreme cases. Simply put, our intuitions about the concept are conflicted even in the quotidian cases, which justifies a wholesale revision—up to and including elimination.

§2.4 WIGGINS AND PERSON AS A NATURAL KIND TERM

We ought to review: I have presented a case for an elimination of the concept of the person. Put another way, I have argued that the best solution to our conflicting intuitions about personal identity is to stop talking about persons and start talking about physical and

psychological continuity and agency. As I have argued in the previous chapter, thought experiments conducted according to proper *Gedankenexperimente* methodology will *expose* our conflicting intuitions but will never *resolve* them. Quine was right to wonder if words have any meaning beyond which our current needs have invested them with. I therefore want to eliminate the concept of the person in favor of its component concepts, in part because in marginal cases our commitments to the various sub-concepts conflict with each other.

It was the Williams thought experiment which revealed that the sub-concepts could be at cross purposes with each other. But we can find cases from real life too. A paradigmatic case where our commitments to these component concepts are at cross-purposes to each other—and where the virtues of eliminativism become clear—is any case where we want to say that an entity is the same animal as before but no longer the same agent and indeed, may no longer exhibit psychological continuity. These cases always make the question "Is he still the same person?" deeply problematic: either the answers don't do justice to one of our intuitions (in the form of a commitment to a sub-concept) or they smack of legislation. Instead of trying to prioritize our conflicting intuitions, why not simply redescribe the situation without making reference to persons? One relevant example could be cribbed from Williams' torture thought experiment. In that case, those with Lockean intuitions who support his distinction between animal and personal identity might say that the patient is the same animal but no longer the same person. (But they might also express their intuition without making reference to persons, i.e. that the procedure preserves bodily identity but violates agent identity and psychological continuity.) There could be other examples taken from the margins of everyday life. One might say of a car crash victim, who is released from the hospital suffering total amnesia, that he is the same animal but not the same person, but why not simply rephrase the description and say that the accident preserved animal identity but violated agent identity and psychological continuity? This description is more accurate than any attempt to determine (or decide) if he is the same *person* after the accident that he

was before the accident. This avenue of response is slightly different from Parfit's claim that these questions do not always have a definitive answer. It is rather to suggest that the question is philosophically irrelevant—i.e. that nothing is lost by changing the terms of the question itself and giving an answer that makes no reference to persons.

One thing that stands in the way of this elimination is Wiggins' claim that 'person' is a natural kind term.[31] If it is a natural kind term then presumably it is not so easy to eliminate. Wiggins claims that natural kind terms are ideally suited to be the sortal concepts over which we quantify. The class of individuals picked out by a natural kind term are grouped together by virtue of nomological regularity and Wiggins claims that it is the concept of person that is the natural kind term. In this section I will explore Wiggins' position, evaluate it as a potential objection to my eliminativist strategy, and demonstrate that it can be rehabilitated even after the concept of the person has been eliminated. In short, while Wiggins goes to great lengths to show that natural kind terms are ideally suited to be sortal concepts, he does little to demonstrate that 'person' is the only natural kind term that fits the bill.

Wiggins' claim can be broken into two parts: his support for the principle of Sortal Dependency or what he calls 'D', and his rejection of the Relativity Thesis of Identity which he labels 'R'. Both principles are inspired by Frege's insight that numbers attach to concepts instead of objects themselves, i.e. that we quantify over concepts under which objects fall. We count three *trees*, four *stones*, even five *things* (referring to small-size objects on the floor), but it is effectively meaningless to look at a given situation and just count "five". Before one can evaluate the veridicality of such a statement, one is in dire need of an answer to the question: "Five what?" When looking at the landscape the count "five" is true if it falls under the concept tree, but false if it falls under the concept copse, for which the proper answer is one.

---

[31] David Wiggins, *Sameness and Substance* (Cambridge: Harvard University Press, 1980), ch. 6.

D is the principle that every identity statement of the form *x* is the same as *y*, or *x=y*, can only be true if there is a sortal concept *f*, under which both *x* and *y* fall, and from which emanates the persistence conditions that explain the identity of the continuant through time. More specifically, Wiggins writes of D that "to fulfill its office and constitute an answer to the *what is it* question, a genuinely sortal predicate must stand for a concept that implicitly or explicitly determines identity, persistence and existence conditions for members of its extension."[32]

By contrast, R is the idea, most frequently associated with Geach, that identity is *relative* to the sortal concept, or in other words that *x* could be the same f as *y* but not the same g (viz. Locke's distinction between animal and personal identity). While D assumes that identity is dependent on the sortal concept, this does not entail that an object could fall under two sortal concepts at one time and then fall under only one of those sortal concepts subsequently. This further entailment would be part of R, which Wiggins rejects because it violates the logic of identity and Leibniz's Law. The rejection is explained most simply when Wiggins writes that "this is a tempting but incoherent decision."[33] The idea here is that if *x=y*, then everything which is true of *x* is true of *y*, assuming the truth of Leibniz's Law that the identity relation is transitive, symmetrical and reflexive. But if *x* is an *f* and *x=y*, then it is impossible for *x* to *fail to be* an f, which is precisely what R denies. Consequently, assuming acceptance of Leibniz's Law, R leads to intolerable conclusions and must be denied on *a priori* grounds. (Wiggins concedes that this argument rests on the assumption that Leibniz's Law is correct, and that someone might, for instance, suggest that the identity relation is not necessarily transitive, symmetric *and* reflexive.)

Having denied R but accepted D, Wiggins argues that if an object falls under multiple sortal concepts there will be an exact correlation between the continuants identified by the concepts. By virtue of this fact, it would be absurd to say of someone, after being

---

[32] Wiggins, p. 62.

[33] Wiggins, "Locke, Butler and the stream of consciousness: And men as a natural kind," p. 163-4.

subjected to Williams' torture operation, that he is no longer the same agent or psychological subject but he is still the same animal, as I am inclined to do in these cases. If Wiggins' analysis and rejection of R were sound, it would seem as if the virtues of my eliminativism would evaporate, for the very description that is evidence of our conflicted intuitions would be guilty of a far greater crime: it would be logically incoherent. If that were the case it could not be evidence that our intuitions were conflicted. It would instead be evidence that the alleged scenario is necessarily impossible. We would need to abandon the description and reexamine my reasons for supporting eliminativism. If the rejection of R is correct, it would seem that we cannot say that the patient is the same animal as before the operation but no longer the same agent and no longer psychologically continuous.

Perhaps we can avoid the quandary. There are three ways to reject R and still retain my description of the patient in Williams' thought experiment. All three avenues involve the claim that my description of the thought experiment is *not* a bona fide case of R. These three avenues are:

(i) The 'is' of constitution
(ii) An ontology of temporal parts
(iii) Taking tensed facts seriously

Avenue (i) is an argument by analogy. The classic case of a supposed 'is' of constitution is a lump of clay and a statue. We often say that this lump of clay *is* a statue, perhaps explaining the situation to an incredulous onlooker at a modern art museum. This leads to philosophical confusion once the lump of clay is reformed, thus suggesting that the statue has been destroyed while the lump of clay persists. This seems like a classic case of the relativity of identity, for the clay *qua* lump of clay continues to exist after the reformation, while the clay *qua* statue does not. The counter-argument for those who are skittish about R is that the 'is' in the sentence "The lump of clay is a statue" is not the 'is' of identity but the 'is' of constitution, thus suggesting that the sentence is not an identity claim at all. Rather it is a claim about what constitutes the statue: i.e. the clay. The sentence is therefore to be analyzed as "the lump of clay constitutes the statue."

Avenue (ii) involves adopting an ontology of temporal parts and switching our description from three-dimensional entities (with spatial parts but no temporal parts) that exist wholly at any moment and persist through time, to four-dimensional entities (with spatial *and* temporal parts) that perdure *across* time. With this switch in our ontology we could say (when we want to express a supposed identity relation) that *x* and *y* are both proper-parts of a four-dimensional mereological sum (entity), and we could then say (when we want to express a supposed lack of identity after the operation) that *a* and *b* are *not* both proper-parts of the same four-dimensional mereological sum.[34]

Avenue (iii)—an appeal to tensed facts—seems like the least ontological of the three avenues—i.e. it seems to have fewer ontological commitments because it is a claim about conceptual facts. If we take tensed facts seriously, we can avoid the conclusion that our description of the paradigmatic case is a bona fide case of R. The confusion stems from thinking that x is the same animal as y but not the same agent, when we should really be saying that x *is now* the same animal that y *was* but x *is not now* the same agent that y *was*. These facts are certainly true. But it would be false to say that x is the same agent now as x used to be (before the accident). And it would be equally false to say that y is the same agent as yesterday.[35]

Officially I am agnostic between the three avenues; there are several reasons to argue that the thought experiment is *not* a bona fide case of R and there is no need at the moment to choose between the avenues. The point is that there are grounds to reject Wiggins' claim that this is a case of R. So the description of the animal who is no longer the same agent does not need to be rejected by virtue of an *a priori* rejection of R.

---

[34] Wiggins claims that accepting an ontology of temporal parts is not a viable alternative and does not offer or consider any arguments against a four-dimensional world view—a somewhat perplexing situation given the serious attention four-dimensionalism has received in the literature.

[35] Harold Noonan makes this point in his review of *Sameness and Substance* (see his "Wiggins' Second Thoughts on Identity," *Philosophical Quarterly* 31 (1981): 260-8). Noonan's example, which is directly analogous, involves a vessel moored on a river. The day before the vessel was moored on the same river but not moored on the same water.

A crucial example for Wiggins' argument is the Story of Lot's wife. He thinks it absurd to invent a concept like woman-salt-pillar in order to explain the continuity of identity through the change. But why invent the concept at all? We could instead stick with two concepts: woman and salt pillar. The first reason is that Wiggins has rejected R on *a priori* grounds. The second is that Wiggins wants a description of the story that has Lot's wife *actually being* a pillar of salt, as opposed to *being replaced* by a pillar of salt. This necessitates the construction of the gerrymandered concept 'woman-salt-pillar'; that being absurd, Wiggins concludes that the story is conceptually incoherent. But there is a crucial disanalogy between the story of Lot's wife and my description of the Williams thought experiment: in the former there is no concept that extends from before the transformation to after, while in the latter the concept animal extends from beginning to end. Before the operation the animal functions as one agent; after the operation the animal functions as another agent.

Wiggins' support for D and his rejection of R make up only half of his argument. The other half is his contention that natural kind terms are ideally suited to be sortal concepts and that 'person' is a natural kind term. But even if that is the case—and I am by no means uncritically accepting a doctrine of natural kind terms—must we abandon the use of natural kind terms if we abandon the concept of the person? Not necessarily—especially since my eliminativism calls for replacing our notion of the person with component concepts such as animal, agency, and psychological continuity. We ought to investigate the status of these replacement concepts to see if they are natural kind terms and will fit the bill as sortal concepts.

One of these component concepts—animal—is clearly a natural kind term by Wiggins' own Aristotelian interpretation of the notion of a natural kind. Following Putnam and Leibniz, Wiggins claims that

> any would-be determination of a natural kind stands or falls with the existence of lawlike principles that will collect together the natural extension of the kind around an arbitrary good specimen of it; and that these lawlike principles will also determine the characteristic

> development and typical history of members of this extension (or at least the limits of any possible development or history of such individuals) (169-70).

According to this criterion it would seem like the animal concept is clearly viable as a natural kind term. Law-like principles collect the extension of "animal" and also suggest a typical development and limits thereof for any particular animal. Consequently, at least one of our replacement concepts is a natural kind term by Wiggins' own account of natural kind terms.

But what of agency? It seems possible that 'agent' will end up not being a natural kind term, at least not by Wiggins' account of what makes a term a natural kind. Wiggins endorses Frege's claim that we can only quantify over a concept which delimits what falls under it in a definite way and does not permit arbitrary divisions. Agency might very well violate the first condition, for Wiggins says that "what is needed for there to be a universally applicable distinction between right and wrong answers to the *special* question 'how many?'" In this sense 'agent' might be different from 'animal'. Wiggins goes on to write that "We can see why identity questions about members of natural kinds may be expected to find the notion of identity at its best, and empirical discovery playing the part it does play here, it is plain why they are the most unsuitable of all candidates for conventionalist treatment." And then later: "It is in no way up to us what to count as persistence through change or through replacement of matter."[36] So agency might not be a good candidate for natural kind status. Although agency *does* depend on empirical discovery and *does not* allow for wholesale conventionalist treatment, it might nonetheless be the case that empirical considerations do not exhaustively determine our ascriptions of agency. This is not to say that they are conventional, but it is to say that we will not find the notion of identity "at its best" here, precisely because there is not a *total* nomological foundation for picking out agents.[37]

Wiggins wants to argue that natural kinds are picked out by reference to a principle of activity "naturally embodied," unlike artifacts which are individuated "with less logical

---

[36] Wiggins, *Sameness and Substance*, p. 88.

[37] I leave open the possibility that competing accounts of agency meet be more amenable to natural kind status, though I claim that the question is moot anyway, because I will soon claim that it is unclear how much is lost if only some of the replacement concepts are natural kind terms.

determinacy and considerably greater arbitrariness, by reference to a parcel of matter so organized as to subserve a certain function."[38] Wiggins notes that it is in artifact concepts where we find the notion of identity at its worst because there is no such thing as the "natural development" of, say, a watch or a clock. Artifact words are not natural kind terms because they are not collected together, according to Wiggins, by scientific law. So what supports the ontology of artifacts? And what makes it possible to treat them as *continuants*? He suggests that it "requires the vestigial continuance of the capacity to subserve whatever roles or ends the artifact was designed as that very artifact to subserve." Wiggins also notes that there are special artifacts in categories all their own. Two such examples are works of art and social artifacts such as an administration or governing body. Artifacts such as clocks can be made of many different kinds of material and with many different mechanisms, and are grouped together by functional descriptions, regardless of composition or constitution. A clock is defined as anything that functions to tell time. While disputes about members of a natural kind can be settled by "getting more scientific facts,"[39] it is still very much up to us to decide whether an artifact will count as persisting or not. Disputes about artifacts are frequently settled by an appeal to convention—although Wiggins claims that the arbitrariness of artifact identity is frequently exaggerated.

This helps explain why Wiggins rejects the possibility that 'person' is a combination of the natural-kind concept 'animal' and a functional/systemic component. The analogy he considers here is the concept of vegetable, which is the combination of the natural kind term 'plant' (stemming from biological law) *plus* a functional component defined by what we consider *edible*. But it is precisely this functional component that Wiggins objects to when it comes to persons. He claims that this functional component, which he presumes to be largely mental, is too difficult to specify; we can't articulate it. In his words we are unable to *fill in the dots* with regard to the mental faculties necessary to meet the functional

---

[38] Wiggins, *Sameness and Substance*, p. 90.
[39] Wiggins, *Sameness and Substance*, p. 88.

component, unlike say, vegetable, where we simply eat the plant in question to see if it meets the standard; he calls the standard transparent. But we are unable to do the same with regard to the functional component of persons. Either we end up being too permissive if we allow any mental faculty to meet the criteria, or we end up being too restrictive and deny personhood to some members of the class of biological human beings.

Furthermore, Wiggins asks—rhetorically—why we don't go all the way and replace the biological-naturalistic view of persons with a purely functional criteria? Wiggins sees no reason to choose any particular mixture of the biological-naturalistic part and the functional part. If we are going to allow a functional part into the definition, we might as well go all the way and make the criteria for personhood totally functional. This would allow any artifact to qualify as a person just so long as it meets the functional definition. The argument is then a *reductio ad absurdum*. Any account of persons that allows an artifact to qualify as a person is unacceptable—simple because too much of artifact individuation is up to us. And that is an unacceptable scenario when it comes to persons.

What then is Wiggins' own view about how we should describe what happens in the Williams thought experiment? His preferred solution is to say that the individual is *both* the same animal *and* the same person throughout the transformation. His rejection of R requires him to go this route (or the route that the individual is *neither* the same animal *nor* the same person after the transformation). Unfortunately his solution ends up violating our powerful psychological intuitions from the Williams thought experiment. The individual is the same person even after suffering complete psychological discontinuity. This is a heavy philosophical price to pay.

I believe that the vegetable idea was dismissed too hastily. I say this because the idea accords nicely with my claim that the best way to understand our conflicting intuitions about personhood is to recognize that the concept is a cluster of different components: biological animal, a good candidate for a natural kind term according to Wiggins himself, and some sort of functional component which corresponds to our intuitions about

psychological continuity and agency. This component (or these components) of the cluster might be artifact(s)—something like a house or a clock whose identity is partly based on a functional definition—not scientific law. But if can quantify over houses successfully (Wiggins never denies this), we certainly can quantify over agents. There is no reason why, in replacing person with its component concepts, that all of the resulting concepts need be natural kind terms. The basic idea, then, is simple: the concept of the person can be eliminated in favor of its sub-concepts. And furthermore, some of those concepts are ready-made natural kind terms, such as 'animal', while others might be more aptly described, in Wiggins' language, as social artifact concepts.

I think this accords quite nicely with the evidence of our conflicted intuitions about personal identity. Where does the conflict come from? In most circumstances, it seems quite simple to individuate, over time, a given person. It is my contention that in many instances we are picking out the sortal concept animal and its identity is usually determinate. (The one exception here being Parfit-style physical spectrum cases, which one is unlikely to encounter in real life). But in other respects our notion of the person does not yield determinate answers to the identity question—either in thought experiments or cases in the margins of real life—and this is precisely because some of the component concepts are social-artifact concepts that may be subject in part to convention and even some legislation, as opposed to being collected around biological law *exclusively*. These components correspond to our intuitions about the psychological element (and agency) in the cluster concept.

This accords nicely with what we find so distressing—and simultaneously obvious—about Parfit's argument about the indeterminacy of identity. It is precisely because personhood is a cluster concept composed of sub-concepts—some of them *ideally* suited for use as natural kind sortal concepts, and others *poorly* suited to the task—that some aspects of the concept seem fraught with indeterminacies, yet other aspects of this concept are more determinate. It is precisely because of the latter that Parfit's revisionary proposal about the indeterminacy of identity seems so shocking to us. But it is precisely because of the former

that Parfit's argument is, upon careful consideration, quite powerful. Both intuitions are at work in our concept of the person: natural kind-terms such as animal and social-artifact/functionally-defined concepts such as agency.

This being the case, doesn't it make sense to keep these sub-concepts separate and deal with them individually? If the component concepts are radically different—some of them well suited for natural-kind status, others best treated as an artifact, then if one really wants a theory of individuation, one had better keep the sub-concepts separate. For as Wiggins readily admits, we use different methods for individuating natural-kind objects than we do for individuating artifacts. There is no hope of getting a clear theory of individuation for persons if we lump these different categories of concepts together. Let us investigate a theory of individuation for animals and agents and an account of psychological continuity. And then let us leave a theory of individuation for persons to the history books of metaphysics.

§2.5 IS THE CONCEPT OF THE PERSON PRIMITIVE?

Besides the essentialism of Wiggins, Strawson's claim that the concept of the person is primitive seems—at first glance—to pose a problem for my eliminativist strategy. If the concept is indeed primitive then perhaps it is logically impossible to break it into components. So it is important to evaluate Strawson's argument from *Individuals* and determine if it prevents us from eliminating the concept of the person. As will become clear, Strawson is using the term 'person' in a very specific context towards a very specific goal: repudiating both the Cartesian and no-ownership theories of the self. Once it becomes clear that my eliminativist position successfully avoids both of these pitfalls, the road to accepting my eliminativist strategy—even after accepting the heart of Strawson's claim that both m- and p-predicates must be ascribed to a single entity—will be clear.

Strawson's first point is that states of consciousness are ascribed to *the very same thing* to which physical states are also ascribed. To defend this conclusion, Strawson considers two

infamous philosophical positions that deny this point. One is Cartesianism, which denies that there is *one* entity to which these different kind of states are ascribed and instead claims that persons are hybrids of two different kinds of entities to each of which *one* kind of state is ascribed. To the body is ascribed the m-predicates, to the pure mental ego is ascribed p-predicates. The other infamous view is the so-called "no-ownership" view, and it holds that states of consciousness are not *ascribed* to anything at all because there is no *subject* to which they can be ascribed. These states just aren't *owned* by anything of any type.

According to Strawson, the no-ownership (or no-self) theorist claims that "only those things whose ownership is logically transferable can be owned at all." But when the theorist tries to articulate his denial of the self he says something like "All of *my* experiences..." are had by a certain body, etc. The problem with the formulation of the denial is his use of the word "my," and any attempt to eliminate its usage makes the experiences necessarily—not contingently—owned, thus violating the constraint that for something to be owned it must logically transferable. The theory falls into incoherence because the no-ownership theorist is unable to eliminate his use of words like "my" and "of," which of course imply the very type of possession and ownership that theorist seeks to deny. Strawson charges the hypothetical no-ownership theorist with incoherence in formulating his denial. Accordingly, Strawson concludes that it is incoherent to hold that the very same state of consciousness which I had could have been someone else's. Experiences like this exhibit what he calls logically non-transferable ownership. And so Strawson concludes that particular states of consciousness owe their identity to the entity to which they are ascribed.

Aside from the problems with the no-ownership view, the second infamous view—Cartesianism—has its parallel difficulties. For example, it is unclear how we ascribe states of consciousness to others on the basis of behavioral evidence (which we clearly do). The difficulty with this procedure is that according to the Cartesian, states of consciousness are supposedly ascribed to a purely mental ego. But if that's the case, how can they be ascribed on the basis of behavioral evidence—which they clearly are—given that such evidence is

clearly more associated with the body, the other entity in the dualism. And there's a corollary problem: it's unclear how we ascribe m-predicates to ourselves in the *absence* of behavioral evidence, which we clearly do not need in the case of self-ascription. But if the Cartesian is right and m-predicates are to be ascribed to the body—and not to the ego—then surely behavioral evidence ought to be required for a valid ascription. But experience tells us that this is not so.

Strawson concludes from this consideration of the no-ownership and Cartesian theories of the self that we need to "acknowledge the primitiveness of the concept of the person." This would seem to propose a possible obstacle to my strategy of eliminating the concept of the person. If the concept of the person is primitive, then surely it cannot be simply eliminated in favor of its component concepts. To do so would be, if Strawson is right, to put our ascription of states of consciousness in jeopardy. But the question here is what Strawson means by "primitive" and what he means by "person". Strawson explains that what "I mean by the concept of the person is the concept of a type of entity such that *both* predicates ascribing states of consciousness *and* predicates ascribing corporeal characteristics, a physical situation etc. are equally applicable to a single individual of that single type." [40] If this is all that is meant by the word person, then it can be accommodated by my eliminativist account because the mere label 'person' is not the important thing here. All Strawson means when he talks about persons is an entity to which *both* p- and m-predicates can be ascribed. And there's nothing hinging on the *name* of the concept that Strawson is here invoking. He might as well call it something else.

What I mean when I advocate the elimination of the concept of the person is that no concept of the person should prioritize biological animalism and psychological continuity in the Williams thought experiment. It is the need to prioritize between these component concepts that I object to.

---

[40] P.F. Strawson, *Individuals: An Essay in Descriptive Metaphysics* (New York: Routledge, 1959), p. 101-2.

When Strawson talks about a person, he might as well be talking about a biological animal *suitably endowed* such that it can self-ascribe p-predicates in the absence of behavioral evidence and others can other-ascribe p-predicates to him on the basis of behavioral evidence. Nothing in my eliminativist strategy prevents this or entails that this is impossible. Remember, Strawson's aim was to attack the Cartesian strategy of solving the conundrum of m- and p-predicate ascription by claiming that we are hybrid entities and that m- and p-predicates are not ascribed to the same entity after all. If Strawson is correct about this, and our intuitions seem to suggest that he is, then the Cartesian strategy is fatally flawed. But in rejecting the Cartesian strategy and maintaining that both types of predicates must be ascribed to the same entity, we are not forced to call such an entity a 'person', and indeed there is nothing philosophically significant in our choice of names for the concept. A more accurate description of such an entity would be a ***biological animal suitably endowed.***

Now what does Strawson mean when he says that the concept of the person (bracketing for a second what that means) is primitive? What he really means, he explains, is that "states of consciousness could not be ascribed at all, *unless* they were ascribed to persons, in the sense I have claimed for this word." So he means that persons must be one entity, not some hybrid, gerrymandered combination as it is in Cartesian metaphysics. When Strawson says "primitive" he means that it cannot be broken up into separate Cartesian parts—a ghost and a machine. This is what he is rejecting when he claims that the concept of the person is primitive. He writes that

> All I have said about the meaning of saying that this concept is primitive is that it is not to be analyzed in a certain way or ways. We are not, for example, to think of it as a secondary kind of entity in relation to two primary kinds, viz. a particular consciousness and a particular human body. I implied also that the Cartesian error is just a special case of the more general error, present in a different form in theories of the no-ownership type, of thinking of the designations, or apparent designations, of persons as *not* denoting precisely the same thing or entity for all kinds of predicates ascribed to the entity designated.[41]

[41] Strawson, p. 104.

This would pose a problem for my eliminativist strategy if it were more ontological—but mine is conceptual. Remember, I am not claiming that we are hybrid entities—a soul and a body—as a Cartesian might claim. My eliminativism is *not* a dualism.

When I eliminate the concept of the person into biological animal, psychological continuant and agent, I am not suggesting that we are somehow a hybrid of three entities of different types—one entity per component concept. On the contrary, I still want to claim that we are essentially single entities, not *dissolvable* in the way in which Strawson discusses with the Cartesian or no-ownership theories. We are biological animals suitably endowed such that we exhibit states of consciousness and exercise rational agency. (Though I do want to leave open the possibility that indexicals such as "we" or "I" can refer ambiguously, depending on the relationship between biological animals and agents, which I will discuss in chapter four. I do not want to presuppose, without argument, that agents must come one-to-a-body.)

That being said, doesn't this make my eliminativism irrelevant? No, it does not, because the fact that we are single entities—not hybrid entities—does nothing to solve the problem of our conflicting intuitions expressed by the Williams thought experiment. The fact that we are single entities does nothing to prevent the fact that our bodily identity and psychological continuity can diverge. And in such a situation it is meaningless (and unimportant) to ask about personal identity. Better to stick with the lower level facts of bodily identity and psychological continuity. These are the facts that matter anyway. These facts are rationally and morally important.

I want to stop using the concept of the person because I think it is unnecessary and because its makes us think that when it comes time to develop identity criteria we need to *choose* between bodily identity and psychological continuity. Indeed it is natural for us to insist upon concrete identity criteria if we are using a concept. As the mantra says, no entity without identity. But instead of devoting ourselves to the difficult task of coming up with

identity criteria for the concept, we should stop using the concept, thus eliminating the need for the identity criteria.

Remember, the Strawsonian point is that both m- and p-predicates must be ascribable to a single entity. There is nothing philosophically significant in calling that entity a person, especially since we might also call it a biological animal suitably endowed. And calling it a person—just by so baptizing it—will not provide us with an answer to our *conflicting intuitions* about personal identity when psychological continuity is transferred, by some causal process, to another body.

But if the Strawsonian point is correct, why not just become an animalist? Why not just follow Williams, Olson, Thomson et. al. and argue that a person is a biological animal and that the criteria for personal identity are the criteria for bodily identity. One might want to argue for this position like this: because states of consciousness must be ascribed to entities such as an animal suitably endowed, then *over time* the identity of such an entity ought to be construed along strictly bodily lines, i.e. through physical spatio-temporal continuity. Unfortunately, though, this will not solve the dilemma arising from the Williams thought experiment. Just because we are talking about an animal suitably endowed does not give us a warrant—despite our first reactions—to choose bodily identity over psychological continuity in the Williams thought experiment.

You cannot extend the Strawsonian point about ascribing m- and p-predicates to the same entity to some sort of account of the entity's identity criteria *over time*. There is a simple reason for this. The Strawsonian account, *as I have reconstructed here*, is consistent with *both* responses (intuitions) to the Williams thought experiment. The account does not entail that one of the responses is wrong. And simply picking bodily continuity over psychological continuity would not do justice to one half of our intuitions.

The Strawsonian account is consistent with both responses because the fact that the entity in question is a biological animal suitably endowed does not by itself entail that bodily identity is the identity criterion of the entity. That is just one possibility. But there are other

coherent possibilities, including the claim that it is the location of the brain—given its position as the causal origin of states of consciousness—that should be the criterion for reidentifying the entity over time. And someone who holds the Narrow Psychological Criterion, where personal identity consists in psychological continuity from its normal cause, i.e. the continued existence of enough of the brain to support consciousness, might very well support this strategy for reidentifying the entity over time. Consequently, the pendulum still swings between animalism and psychologism, between our two conflicting intuitions from the Williams thought experiment. Only a bias would lead one to suggest that the identity criteria of the entity in question (an animal suitably endowed) must be based on the spatio-temporal continuity of the body as a whole. The relationship between these two positions is not a logical entailment. Indeed, it is quite coherent to claim that the entity ought to be identified by the position of the brain, as might be suggested advocates of the Narrow Psychological Criterion of personal identity. So the Strawsonian attack against Cartesianism does not resolve our conflicting intuitions about personhood. Furthermore, my eliminativist strategy (which does resolve our conflicting intuitions) is consistent with his claim that m- and p-predicates must be ascribed to the same entity.

§2.6 THE THEORETICIAN'S DILEMMA

There is a third category of objection to the eliminativist strategy. While the previous group of objections argued that the concept was either a natural kind or primitive, this third objection makes a general claim about the function of concepts and the appropriateness of eliminating them. According to one view, elimination of the concept of the person (or any concept) is not warranted solely by our inability to discern its structure. Although we may not have a complete understanding of the concept of the person, this isn't to suggest that we won't achieve it in the future. For example, there was a time when we did not know that water was H_2O. In fact, there were inconsistencies about water-like substances that met some of the criteria for water but not others. But this was not good reason to eliminate the

concept of water. It was simply a function of our own ignorance and suggested a programme for future research. Eventually the internal structure of water *was* discovered and the appropriate modifications to the concept were made. Perhaps the same goes for the concept of the person. Although we have been unable to offer a coherent theory of persons, this is simply evidence of our own fallibility and the limits of investigation. Just as rigorous scientific experimentation did not immediately yield a correct theory of water, so too rigorous thought experimentation has not yet yielded a correct theory of personal identity.

One way to get a sense of this objection is by looking at the following passage by William James:

> The boundary line of the mental is certainly vague. It is better not to be pedantic, but to let the science be as vague as its subject, and include such phenomena as these if by so doing we can throw any light onto the main business in hand. It will ere long be seen, I trust, that we can; and that we gain much more by a broad than by a narrow conception of our subject. At a certain stage in the development of every science a degree of vagueness is what best consists with fertility.[42]

We can also flush out the anti-eliminativist objection by looking at Hempel on scientific concepts. He writes that

> The scientist does indeed wish to leave open the possibility of adding to his theory further statements involving his theoretical terms—statements which may yield new interpretative connections between theoretical and observational terms; and yet he will regard these as additional assumptions about the same hypothetical entities to which the theoretical terms referred before the expansion. This way of looking at theoretical terms appears to have definite heuristic value. It stimulates the invention and use of powerfully explanatory concepts for which only some links with experience can be indicated at the time, but which are fruitful in suggesting further lines of research that may lead to additional connections with the data of direct observation.[43]

What I want to highlight from both of these passages is the degree to which the concept of the person may indeed "suggest further lines of research" even if, for the moment, the exact fruit of that research may not be visible. Hempel's point is that the theoretician's dilemma emerges once it is realized that theoretical entities are envoked to "establish definite connections among observable phenomena," such that one might as well dispense entirely

---

[42] James, *The Principles of Psychology*, p. 6.

[43] Carl G. Hempel, "The Theoretician's Dilemma: A Study in the Logic of Theory Construction" in *Aspects of Scientific Explanation: And Other Essays in the Philosophy of Science* (New York: The Free Press, 1965), p. 206.

with the theoretical entities in favor of direct relations between what Hempel calls observational antecedents and observational consequents. This would seem to be a good case for conceptual elimination.

Hempel's answer to the theoretician's dilemma provides the meat of the objection to the eliminativist strategy. Hempel's point is that we posit the existence of the theoretical entities not simply for the purpose of current causal connections between observable phenomena—in which case they might indeed be dispensable—but also the *promise* of such connections in the future. The theoretical entities, insofar as they embody a mini-theory about the observable phenomena, offer a theoretical structure for future research that will yield fruit in the future. And this is the point of positing the theoretical entities. To extend the point to the concept of the person, the conceptual apparatus offered by the concept of the person promises a future connection between the observable elements that may not be presently forthcoming. Yet the theoretical concept provides an avenue of research to determine the exact relationship between the physical, psychological and ethical dimensions of personhood. This would seem to speak against the strategy of elimination.

To push the point even further: the positing of the theoretical entity is based at least partly on utility, in this case the promise of future understanding. But the utility of the theoretical concept, whose internal structure remains elusive, extends even further. Even if it were possible to eliminate the theoretical concept in favor of lower level facts, doing so may be damaging if connections among the lower level facts are elusive. Consider an analogy to mental-physical reductions in the philosophy of mind—which was James' subject in the first passage. Although in principle it might be possible to reduce all claims about mental states to claims about physical events in the brain, doing so may not be advisable insofar as the causal connections between physical events in the brain remain beyond our epistemic reach. In short, the explanatory gap is a true gap—it runs in both directions. While in some cases it might help our understanding to reduce mental phenomena to physical facts about brain states, this is usually only true in cases *where we know the physical facts about the brain states.*

In cases where those physical facts remain elusive—which is frankly most of the cases given the elementary state of neuroscience—the mental vocabulary is indispensable for researching our cognitive structure.

The same analysis might hold for the concept of the person. Although in principle we might reduce all facts about persons to facts about biological human beings, psychological continuity, and agency, doing so would only be advisable if we had a complete understanding of those lower level facts. But we don't and the explanatory gap goes in both directions, according to this objection. Pursuing a line of inquiry with the concept of the person allows us to investigate connections which we could not do with our primitive understanding of the lower level facts. One place where this might be true is value theory. We use the concept of the person in value theory because our epistemic reach does not extend as far into the lower level facts. Here as well the explanatory gap goes in both directions. The concept of the person might be useful, not in spite of our ignorance of its internal structure, but precisely *because of* our ignorance of its internal structure.

The preceding paragraphs point not to a single objection to the eliminativist strategy but rather to several possible critical approaches to it. But they all share a common skepticism about the role I have identified for theoretical concepts in general. As such, I will offer a response that is general enough to answer any of these objections.

The question at hand is the appropriateness of the analogy. If water is indeed an appropriate analogy, it would seem as if elimination is unwarranted and unwise. But the analogy might be strained. In the case of water we have good reason to suspect the future discovery of its internal structure. We have good reason to think that the concept, as it were, stands for a single entity with a distinct structure. The question is whether we have similarly good evidence in the case of persons. I argue that we do not.

Consider a different analogy that I have already offered. Medical cluster concepts are often eliminated (or redefined) when it turns out that they do not, in fact, stand for a single illness defined by a single cause. While the appearance of common symptoms at first made

the suggestion of a single cause seem likely, further investigation of the symptoms can lead to the identification of multiple causes, with the result that the syndrome in question is eliminated in favor of two separate diseases (or the syndrome is redefined with the help of the cause and a new classification is offered for the second cause). The question becomes one of evidence. Is the concept of the person more like the concept of water or is it more like a medical syndrome based on symptomology which is now known to arise from two distinct causes?

The evidence of conflicting intuitions exposed by the Williams thought experiment is strong evidence that the latter analogy is more applicable than the former. The Williams thought experiment exposed a deep conflict in our intuitions about personal identity. One might think that the conflict could be resolved, perhaps by more *Gedankenexperimente*, just as further scientific experiments yielded a satisfactory theory of water. But this suggestion ignores the fact that our intuitions are conflicted not just in thought experiments but in real life as well. It is unlikely that thought experiments will offer a coherent theory of personal identity that does not violate one of our major intuitions. This is the case because personhood is a cluster concept and water is not.

Consider medical cluster concepts again. This is a classic case where the existence of the theoretical construct helps guide avenues of future research. A medical syndrome is posited as the cause for observable maladies on the faith that a *single* cause is the culprit. This is legitimate and elimination is not appropriate if the faith is well founded. But what reason do we have for our faith in the concept of the person—faith that a coherent account of personal identity will be forthcoming if we simply look hard enough? My review of the evidence, particularly the Williams thought experiment, suggests that the answers are not forthcoming and the faith misguided. The conflict is too deep are resides not just in thought experiments but in our intuitions about real life cases as well.

Consider the question again in terms introduced by Hempel. We posit the existence of theoretical entities to help establish connections between observables. But in this case the

observables stand in deep tension with each other; indeed, in the Williams thought experiment they outright contradict each other. Consequently, the observables support *different* theoretical constructs—physical human beings, psychological continuity, agency. Far from suggesting further avenues of research, the concept of the person only reveals a deep conflict in our intuitions. And remember that the theoretical constructs are there to facilitate our understanding. And my whole argument has been that this is best accomplished by positing theoretical constructs *other* than the concept of the person. We ought to talk about physical human beings, psychological continuity and agency because doing so is more efficacious than talking about the concept of the person.

This suggests a new, but related, objection to the eliminativist strategy. The anti-eliminativist might come back with the following retort: there *is* a structure to the concept of the person and that structure is entirely evaluative. In other words, the concept of the person is applied to beings who are appropriate objects of moral evaluation—evaluation of the variety exposed by Strawson in "Freedom and Resentment" as being part of the basic fabric of interpersonal relations. The anti-eliminativist would concede that the details of the concept of the person are subject to controversy, but this is just to say that the concept suggests future avenues of research. And that research ought to center around the concept's use to single out beings for moral evaluation. The fact that we are uncertain who qualifies for the label, and under what circumstances they qualify for it, is simply a recognition that we have yet to construct a complete theory of persons. But retaining the concept gives us a structure to solve those problems around our current normative uses for the concept.

This is a serious objection and there are at least two avenues of response. The first option is to throw this evaluative version of the concept of the person into the eliminativist mix. Having already identified biological, psychological and agency components of the concept of the person on the basis of the Williams thought experiment, I might add an evaluative component as well. There might be good reason to separate the evaluative from

the metaphysical in the concept of the person, insofar as they all function differently and using different concepts would keep them separate.

But there is a danger to this hyper-eliminativism. Although it is indeed the case that the concept of the person is used to demarcate a proper object of moral evaluation, this does not exhaust value theory's use of this concept. Persons are also demarcated for the appropriate conferal of moral rights. And there are other normative-axiological uses for the concept as well. If we had good reason to think that there was an exact correspondence between these concerns, a single concept might do. For example, if it turned out that the class of beings who were appropriate subjects of moral evaluation coincided exactly with the class of beings who were appropriate objects of moral rights (as they do in Kant), then one concept would do. But as I will investigate in subsequent chapters, it is not at all certain that this coincidence will obtain. That being the case, we would need a new concept for every axiological use for the concept of the person. And this would seem to be absurd. So the strategy that aimed to simplify ends up, ironically, doing the exact opposite—complicating the conceptual landscape.

Thankfully there is a second option worthy of investigation: recognize that personhood is indeed primarily an axiological concept but decline to separate that element from the other components. I believe there are several good reasons to recommend this alternative. First, this option recognizes the very fact that the anti-eliminativist highlights: that the concept of the person is largely axiological. Consequently, we should leave the axiological considerations within the biological, psychological and agency components. Furthermore, with this option we can clean up the axiological problems with the concept of the person. Part of the problem with the concept of the person has been that a single notion has been used for different axiological phenomena: rights, self-concern, responsibility, etc., and in each case the psychological, biological and agency elements of the concept were appealed to—although often they were invoked latently through the explicit invocation of the concept of the person. We might clear up our understanding of these phenomena if we

appeal directly to the lower level elements just described. This would help us understand if and when we consider psychological factors morally significant, if and when we consider biological elements morally significant, and if and when we consider agency morally significant. This investigation will be undertaken in chapters three through five.

Over the course of two chapters my argument has spanned from our conflicting intuitions exposed by the Williams thought experiment to a modest proposal to respond to this conflict. I have defended my claim that thought experiments, if conducted appropriately, are unlikely to offer a resolution to this conflict and that the concept of the person should be considered a cluster concept; that is the best explanation for the appearance of the conflict. As such, the concept should be eliminated. In addition to explaining this eliminativism, I have shown that this proposal is not impossible, i.e. that the elimination is a logically coherent possibility. I will next show in the rest of this project that we can make do without our concept of the person, particularly in our accounts of traditionally person-centric topics such as self-concern, responsibility, and rights. It is my contention that far from making such accounts impossible, eliminating the concept of the person will help us develop accounts of these topics that bring them closer in line with many of our complex intuitions. But this, of course, must be demonstrated.

§2.7 THE COMPONENTS OF THE CLUSTER

One last issue remains. Until now I have left open the exact list of components from the cluster concept. I have concentrated on the claim that personhood is a cluster concept composed of distinct components, leaving to the side the question of what exactly the components are. Tentatively I identified the components with the conflicting intuitions in the Williams thought experiment. I suggested that one component would line up with our psychological intuitions, one component would line up with our animalist intuitions, and perhaps agency might be a distinct component as well.

But perhaps the list of components could be smaller. We could, for example, eliminate the concept of the person in favor of 'human being'. While this is a sensible proposal it does not change the circumstances that caused the conflict in our intuitions in the Williams thought experiment. Remember, an identity theory for a concept that includes both physical and psychological elements will be problematic because the two elements can be pried apart. And there is nothing special about the word 'person'. The conflict will arise in *any* similar concept that attempts to subsume both physical and psychological aspects under one identity theory. This is the moral one should extract from the Williams thought experiment. Insofar as 'human being' is a concept that attempts to do the same (subsume physical and psychological aspects under one term), it will fall victim to the same fate.

The second question is whether the list of components could be larger. But what other concepts would you include? We could eliminate our components even further and push them one level down; we could, as it were, push them *all the way down*, perhaps even to the level of particles. This might have some advantages and some metaphysicians have supported this kind of eliminativism in a broader context. But that would cast my eliminativism in too ontological a light. I am not trying to make the claim that persons don't exist in the same way that a metaphysician might argue that composite objects don't exist or that nothing really exists except fundamental physical particles—whatever contemporary physics calls them. Remember, the basis for my conceptual eliminativism is largely pragmatic—we can say more and do so without running into problems (say, in the Williams thought experiment) if we eliminate the concept of the person and replace it with biological bodies, psychological continuants and agents.

The third question is why the list can't be different. In other words, why *these* components? The composition of the list stems from the problem that prompted us to go with the eliminativist strategy. That's what is driving this. Because of our competing animalist and psychological intuitions in the Williams thought experiment we are left with two components: human bodies and psychological continuants. These concepts include

their analogs of physical and psychological continuity. It is the elimination of the concept of the person into at least these components that resolves the conflict; elimination of the concept into other components will not resolve the conflict.

I have made special allowances for the concept of agency, which I have suggested might be distinct from both the physical and psychological components. So far my only claim has been that there is no logical reason to suppose that agency could not be a distinct concept from the other two. In the next chapter I will explore this issue at length, considering whether our intuitive notion of agency requires physical continuity, psychological continuity, or unity of consciousness. I will start by appealing to skepticism about the unity of consciousness raised by Nagel. From there I will use as my methodological tool our intuitions about responsibility. I will probe our intuitions to see if we are inclined to attach responsibility in cases that extend beyond psychological continuity. If we are so inclined, this will be powerful evidence that agency does not require physical or psychological continuity; we will then have a warrant for including agency as a third component of the cluster. But the argument for eliminativism does not stand or fall on this point.

One might object that we have departed from the methodology of basing the components on the intuitions from the Williams thought experiment, which was the source of the conflict we are trying so hard to resolve. Until now I have only identified two intuitions that were in conflict. But it is my contention that we have implicit intuitions about agency in the Williams thought experiment that were never explored. Suffice it to say that the body that receives your psychology is not only your psychological continuant but your agent continuant as well. One could conceivably run through the experiment again to highlight these intuitions. And in the next chapter I will consider thought experiments that bring these intuitions to the surface and investigate whether they are indeed distinct from psychological and physical continuity.

The goal of this chapter is to see if a coherent and convincing account of responsibility can be provided without making reference to our concept of the person. I suggested in previous chapters that we could resolve our conflicting intuitions about personhood by eliminating its use. (These conflicting intuitions are no surprise when we recognize that the concept of the person is a cluster concept.) Implicit in this suggestion is a revisionary hypothesis: that the concept of the person is dispensable and that the topics that are traditionally tied to the concept of the person should be tied to the component concepts of the cluster.

To pursue the eliminativist hypothesis I want to explore the roots of a person-centric phenomenon: responsibility. If indeed personhood is a dispensable concept, then it should be possible to provide a coherent account of responsibility without it. Failure to provide such an account would cast doubt on the hypothesis and suggest an alternate course.

The account traces a series of intuitive connections between responsibility and agency. Given our ethical intuition that agents—whatever they might be and independent of their possible connection to concepts such as personhood, bodily continuity and consciousness—are responsible, the relevant question is which account of agency is sufficient to justify our universal intuition of responsibility. As it will turn out, there are reasons to suspect that the account of agency sufficient to ground an account of responsibility will not require phenomenological unity but will be based on some kind of rational unity. If this is correct, then the role of phenomenological unity in the concept of agency has been exaggerated in the past. Consideration of atypical agents such as group and multiple agents (and our ethical intuitions about the responsibility of these agents) will show why one might be skeptical of the phenomenological assumption. (But even if the phenomenological assumption turns out to be warranted and the skepticism misplaced, this would not be a fatal objection to the eliminativist position, since it would still be possible to offer an account of responsibility that makes no reference to persons. It's just that skepticism about the role played by phenomenological unity in agency helps highlight the virtues of elimination.)

More specifically, the mere consideration of these atypical agents—independent of the question of their actual existence—will demonstrate the virtues of my eliminativist position. One can respond to our ethical intuition that these agents are responsible, without going the extra distance and calling them persons—a declaration which would violate our intuitions (exposed in the Williams thought experiment) that persons come one to a body. The eliminativist gets to have his cake and eat it too. If he is inclined to accept the possibility of group and multiple agents, he can affirm his intuition that they are responsible (and hence call them agents) without violating his animalist intuitions (or his psychological intuitions, for that matter) about personal identity. Consequently, my analysis shows that not only is it possible to account for responsibility without persons, it is preferable.

§3.1 WHAT MATTERS FOR RESPONSIBILITY

Since we aim to explore the role of personhood in an account of responsibility, we should work backwards from responsibility and see where it leads us. I begin the inquiry with the following ethical *assumption*: we are responsible. Leaving aside the question of persons for the moment, it seems clear that the ethical assumption of responsibility is a near universal one. I will not offer a proof or defense of this ethical assumption, but it is so universal an intuition that it seems like a legitimate place to begin an inquiry about responsibility. Of course, the ethical assumption, as stated, has little content since both our notions of responsibility and agency remain undefined for the moment. Only once they are defined will the assumption gain much in terms of positive content. But that very lack of positive content is the assumption's greatest asset. The starting point is broad enough that those with different intuitions about responsibility can embrace the assumption as an appropriate starting point.

There is a potential problem with the ethical assumption. Those who are familiar with the literature of freewill and determinism might reject the assumption because they believe that responsibility is an illusion. Traditionally, the freewill paradox has been

construed roughly along the following lines: if determinism is true, then all of our actions are causally determined. If our actions are causally determined they are not free. Without freedom, responsibility is meaningless and vanishes from the stage. *No one* is responsible because freewill is an illusion. If this story is correct it would seem that the ethical assumption is controversial and demands a defense. That defense would seem to require a satisfactory resolution of the problem of freewill—a tall order indeed.

Luckily no such defense is required. The argument presented in the first two chapters concluded with the following proposal: to deal with our conflicting intuitions we ought to eliminate the concept of the person. The burden of this chapter is to explore whether that eliminativism is plausible given that many theorists have put the concept of the person to various uses in moral, political and legal contexts. I advanced the hypothesis that personhood is a cluster concept; consequently in those moral, ethical and legal contexts it is the *components* of the cluster that are doing the real work. So if the hypothesis is correct, then in the case of a moral concept such as responsibility it is one of the components that is doing the argumentative work. Theorists who use the concept of the person to ground their account of responsibility can therefore use its components. And the result is that a potential roadblock to eliminating the concept of the person is removed. The concept of the person is not as indispensable as one might have thought, at least not where responsibility is concerned.

But a theorist who is genuinely skeptical about responsibility—because of determinism—will not feel the relevant anxiety about eliminating the concept of the person. Since such a theorist has no account of responsibility, she is not burdened by the news that the concept of the person is no longer at her disposal, although she may feel the anxiety if she makes use of the concept to ground other axiological concepts (such as moral rights). It is in this sense that the argument in this chapter *does not* depend on a successful resolution of the problem of freewill. Quite simply, a theorist who did not accept the reality of responsibility could not muster the objection to which this chapter responds: that the

concept of the person cannot be eliminated because it is essential to ground our concept of responsibility.

Now that we have a starting point we can define the parameters of our investigation. Given the reasonable assumption that we are responsible beings, the question is: ***what kind of entities are responsible?*** (Or put another way, to which concept does responsibility attach?) In order to get a sense of the alternatives we need to review the eliminativist hypothesis expressed in the previous chapter. I suggested that our conflicting intuitions about the concept of the person were evidence that the concept is a cluster concept. Among the candidates for components of the cluster concept are biological continuity, psychological continuity and agency. So where responsibility is concerned there are three broad possibilities: (1) responsibility attaches directly to the cluster concept itself, thus demonstrating that the concept is indispensable and that elimination comes with a high philosophical price; (2) responsibility attaches to one of the component concepts; or (3) responsibility attaches to more than one of the component concepts. It is the task of this chapter to investigate these three possibilities.

More specifically, as far as the components are concerned, the logical place to begin is the concept of agency. It is clear that if anything is responsible, certainly agents are. However, if option (1) turns out to be correct, then responsibility attaches to the higher level concept and filters downwards, i.e. agents are responsible only because they are persons. This would be another example of an Argument from Above (see chapter two). The other alternative is that responsibility is an example of an Argument from Below. In that case, persons have traditionally been held responsible only because they are agents. As such, the importance for responsibility migrates upwards from the concept of agency. If that is the case, nothing would be lost by eliminating the higher level concept since all of the importance derives from the lower level facts. I hope to show in this chapter a few reasons why responsibility is another example of how the Argument from Below applies to the cluster concept of personhood.

Both of the possibilities discussed in the previous paragraph assume that agency is a distinct concept from the other component concepts. Isn't it possible that agency is not a distinct concept from, say, psychological continuity? And in fact, isn't psychological continuity rather important for responsibility? Indeed, Locke himself argued that it would be absurd to hold a person responsible for criminal actions he no longer remembered, writing that "in the Great Day, wherein the secrets of all hearts shall be laid open, it may be reasonable to think, no one shall be made to answer for what he knows nothing of; but shall receive his doom, his conscience accusing or excusing him."[44] There is then the third possibility mentioned above, i.e. that responsibility attaches to more than one of the cluster concepts that I have identified. Following Rovane, I will explore the possibility that agency is a distinct concept from psychological continuity and phenomenological unity.[45] While I will end up cautiously supporting this possibility, it is important to remember that this claim is not essential to my eliminativist argument. What is necessary is the acceptance of the Argument from Below and the idea that where responsibility is concerned the importance of personhood migrates upwards from its lower level facts. It is my contention that the relevant lower level facts are about agency—not psychological continuity. However, someone might disagree about the relevant lower level facts. For example, they might claim that an account of agency independent of phenomenological unity is untenable. This is a defensible claim, though I think it is incorrect. But just so long as they support the contention that the importance for responsibility derives from the lower level facts—whichever facts they may be—then it will be clear that eliminativism is a viable option.

Turning to the broader question, how will we know if responsibility attaches directly to the concept of the person or whether it is actually grounded by the lower level facts? If we

---

[44] John Locke, *An Essay Concerning Human Understanding*, Book II, Chapter 27, Section 22.

[45] Rovane argues in *The Bounds of Agency: An Essay in Revisionary Metaphysics* (Princeton: Princeton University Press, 1998) that a commitment to overall rational unity is the hallmark of agency and personhood. Part of this claim included the assertion that a common rational point of view does not require a single phenomenological point of view. It is this account of agency which forms the framework within which my inquiry proceeds.

can fill out a convincing and coherent account of both agency and responsibility without appealing to the concept of the person, then this will be good evidence that we should attribute responsibility directly to the component concept of agency. This, in turn, would remove a negative reason for opposing the eliminativist position. Also, if attributing responsibility directly to the component concept of agency instead of to personhood results in a clearer understanding of both agency and responsibility, this would be a *positive* reason for supporting elimination of the cluster concept and replacement with its components.

Why is it that responsibility has so often been associated with the concept of the person in the past? To hazard a guess, I bring the reader back to the Quinean point that concepts—including personhood—are only as precise as needed. They are custom designed according to our current needs. Since the class of agents and persons dovetail so closely, attributing responsibility directly to personhood was good enough for our purposes of establishing responsibility-centered institutions and practices, such as punishment. But a continuation of this practice of attributing responsibility to personhood, if unnecessary, would be unwise given that we have strong metaphysical reasons for elimination.

§3.2 WHAT IS AN AGENT?

If the concept of the person is dispensable we should be able to offer convincing accounts of agency and responsibility without it. I have already put forward the hypothesis that the importance of personhood originates in its lower level facts. Specifically, the hypothesis suggests that 'persons' are responsible only because they also are agents, so what really matters for responsibility is agency—not personhood at all. At the moment this is just a *working* hypothesis—not a conclusion—and forms the basis for further investigation. I suggest making use of the hypothesis with the following methodology: work backwards from the concept of responsibility and from our intuition that agents are responsible. In order for us to affirm our intuition that agents are responsible we need—at minimum—an account of agency that is sufficient to account for the phenomenon of responsibility.

What is that account of agency? My starting definition is the pooling of information, deliberation on the basis of that information and the attempted implementation of a project, plan or response. It seems to me that we can arrive at this starting definition if we consider what counts as an intentional action—as opposed to a mere happening—and fashion from that a minimal definition of agency. When we consider intentional action it is precisely that—action that is produced *intentionally*, not by spasm, not by animal instinct. In other words, the action results—at least in part—by deliberation. Deliberation is only possible when there is information available to form the input of deliberation. And deliberation has no *raison d'être* unless it is directed towards the goal of formulating a plan on the basis of that deliberation. Formulation of that plan is, as it were, its output. So deliberation would seem to be the hallmark of intentional action, of agency. And it glues together the other elements that precede an intentional action: on the one hand information pooling, on the other, implementation of a response. So discussing agency in terms of information pooling, deliberation and implementation, seems like a reasonable place to begin. And it accords closely with Rovane's notion of an agent as having a commitment to overall rational unity.

To understand what this means (information-pooling, deliberation and implementation), let's work by way of analogy from the more intuitive case of agency over time so that we can understand agency at a time. So, for example, an agent is any system that is committed to gathering relevant information, deliberating on the basis of that information and then implementing a response on the basis of that deliberation. Consider my own exercise of agency over time. I am committed to collecting information at different points in time and resolving any apparent conflicts among the information (in the sense that I try to form a coherent picture of the world with that information); I deliberate on the basis of that information at different points in time and attempt to arrive at an all-thing-considered judgment of what I ought to do. In conducting that deliberation I again try to form a coherent picture—like in the information gathering stage—with the exception that this time the picture is of a possible world where I perform a certain action. Because I try to form a

coherent picture I recognize that my competing impulses of what to do, which I experience at different points in time, should be reconciled, though I may not necessarily succeed completely in this reconciliation. I am committed to performing an action based on an all-things-considered judgment—not a spattering of contradictory actions that fulfill transient desires. Furthermore, when an all-things-considered judgment requires a complex action, I am committed to the necessary coordination over time required to implement a long-term project. Component or basic actions are coordinated to form the complex course of action that was dictated by the all-things-considered judgment.

Having drawn a picture of agency over time, consider agency *at a time*. We form the same kind of all-things-considered judgments *at a time* that we do over time. Even on the short term we are committed to, for example, pooling all of the information that we receive in a single moment in our field of vision. We feel compelled to harmonize this information into one coherent picture of the way the world is. However, the same cannot be said of deliberation because *all* deliberation is temporally extended to some degree. This fact has been noted by those—Korsgaard chief among them—who have argued that the deliberative point of view implicit in agency constitutes some kind of identity over time.[46] The very fact that an agent's deliberation takes place over time constitutes the very identity of the agent in question.

Two concepts are noticeably absent in this initial account of agency: psychological connectivity in general and memory in particular. Neither concept has made its way into this theory of agency and this might seem strange given that memory has traditionally been considered essential for responsibility and consequently, for agency. The question of responsibility has been tied to memory at least as far back as Locke, if not further. Locke writes that 'person'

> is a forensic term, appropriating actions and their merit; and so belongs only to intelligent agents, capable of a law, and happiness, and misery. This personality extends itself beyond present existence only by consciousness,—whereby it becomes concerned and accountable;

[46] Christine Korsgaard, "Personal Identity and the Unity of Agency: A Kantian Response to Parfit" in *Philosophy and Public Affairs* 18 (1989): 101-132.

> owns and imputes to itself past actions, just upon the same ground and for the same reason as it does the present.[47]

Locke's intuition here would seem—at first glance—to be evidence of a central role for memory in any good account of agency.

But the assumption that psychological connectivity and memory play a leading role in both responsibility and agency is an assumption that can—and ought to be—subjected to rigorous examination. It seems clear that responsibility applies if and only if you are still the same agent as the agent who committed the action in question. I will call this the self-sameness constraint and I take it to be an essential constraint for any good theory of responsibility.[48] The constraint's sole justification is our ethical intuition that it would be radically unfair to punish one agent for the misdeeds of another and similarly inappropriate to reward one agent for the good deeds of another.[49]

And so the question becomes: are memory and psychological connectivity necessary for agent identity? Or is being committed to rational unity over time—independent of memory—sufficient? We will deal with memory first and save psychological connectivity in general for subsequent sections. Our strategy for investigating if memory is necessary for agency will be to see if the concept is necessary in cases of responsibility. The idea here is

---

[47] John Locke, *An Essay Concerning Human Understanding*, Book II, Chapter 27, Section 26.

[48] For example, F.H. Bradley represents the mainstream view when he writes in "The Vulgar Notion of Responsibility in Connexion with the Theories of Free-Will and Necessity" that "a man must act himself, be now the same man who acted, have been himself at the time of the act, have had sense enough to know what he was doing, and to know good from bad" (*Ethical Studies*, p. 9).

[49] One might object that some have denied this constraint, most notably Parfit, who reaches his conclusion after considering a revision to the Branch Line thought experiment in which Backup is punished for Parfit's crime. He writes in *Ethics*: "Backup protests, 'This is outrageous. These connections are irrelevant. I did not choose to resemble Parfit, or to have these apparent memories. I cannot deserve to be punished for what Parfit did before I even existed'" (838). Parfit admits that most of us would agree with Backup (even though Backup *himself* might very well not agree, because he might *think* that he is Parfit) that he does not deserve to be punished because he is not Parfit. That would show that we believe in some kind of further fact, that we are not reductionists. But it also shows that if we *became* reductionists, we would have to give up our notion of desert. "If it was only this fact which could carry with it desert and guilt, these have also disappeared. No one ever deserves to be punished for anything they did" (839).

However, we do not need to respond to this objection. The *goal* of my chapter was to show that we can offer a satisfactory account of responsibility without making reference to persons, thus showing that the strategy is plausible. But Parfit would not feel this anxiety in the first place since he has little use for a concept of responsibility having given up on desert. Simply put, Parfit would never feel the anxiety that this chapter is meant to alleviate; those who do feel this anxiety do not accept Parfit's position.

that our gut intuitions about responsibility could lead us to our account of agency. If memory turns out to be inessential for responsibility, then it need not be included in an account of agency that is designed to explain our ethical intuition that agents are responsible.

Consider the following thought experiment: as a result of an auto accident resulting in brain damage a patient has no direct, long-term memory. The patient's doctors were successful in repairing all of the brain damage with the exception of the small area of the brain responsible for memory. Consequently the patient receives no information from direct memory sources. However, in response to this shortcoming, the patient has devised an elaborate computer system to compensate for this shortcoming. As he has a thought he immediately records it verbally into the computer system. The computer then catalogues the information in the form of voice recordings and on command plays the recordings back to the patient. So the patient has access to relevant information over time, although instead of having direct access through memory, his access to the information is always from the secondary source of this computer program. With the help of this system the patient is able to execute long-term projects and plans. For instance, the patient forms a long-term intention on Friday to embezzle money from his employer on Monday. He records this intention into the system on Friday, along with a detailed set of instructions for intermediate steps of the plan that he must perform on the weekend so that his criminal plan can be completed on Monday. On Saturday, the computer system reminds him of the long-term intention that he formed on Friday and the steps that he decided on Friday that he ought to perform on Saturday. Finally, with the help of the computer system, the patient is able to complete the criminal conspiracy on Monday with the result that he now has the money—not his employer.

Is the patient responsible for the embezzlement? Our ethical intuition is that indeed he is responsible. He formulated the plan, performed it, reaped its benefits and caused harm to his victim. He should be punished for his crime because he deserves to be punished or perhaps because doing so might prevent him from doing it again. We can take these

powerful ethical intuitions about responsibility as suggestive evidence for an attribution of agency. In other words, our intuitions about responsibility suggest something about our view of agency—and it's a good idea to elicit the former to elucidate the latter. In this case, our intuition that the patient is responsible is our best intuitive evidence that we also consider him an agent. And not only do our ethical intuitions suggest that we consider the patient a responsible agent, they also suggest that we see him as the same agent at the conclusion of the criminal enterprise as he was when he formed his long-term intention to commit the crime. Now the important question is why.

The fact that he had no direct memory connecting the two events in his life seems to be comparatively unimportant when considering the fact that he managed to share information over the stretch of time. It was the sharing of information that was essential to carrying out the project. And the fact that this sharing of information was accomplished through the indirect route of a computer system—as opposed to the more direct route of memory—does nothing to block our ethical intuitions that he is a temporally extended agent who is responsible for his actions. Consequently, it appears as if memory's role in responsibility has been exaggerated (from Locke to the present). Added to the connection that we have traced between responsibility and agency, it appears then that memory's role in agency has been exaggerated too. It seems more likely that it is the pooling of information generally—of which memory is just one method—that is essential to agency.

Before continuing, I must clarify the sense in which I am using the term "psychological continuity." I have assumed that psychological continuity is a phenomenological affair and that it entails some phenomenological unity over time. In other words, a being who has psychological continuity over time has, just by virtue of that continuity, *phenomenological* continuity as well—because that's what it means to have psychological continuity. But this can plausibly be denied, for instance by those in the philosophy of mind who define psychological states by their *functional* role. In that case, psychological states gain their identity by the functional role they play within a cognitive

system. As such, under this use of the term "psychological," the fact that a being has psychological continuity does not entail that it has phenomenological unity. The functional role could be satisfied across several biological bodies, as is the case with a group agent who could then be said to exhibit psychological continuity with the new sense of the term "psychological." Therefore, for this kind of theorist, psychological continuity and phenomenological unity are quite different affairs.

But this debate within the philosophy of mind need not concern us here, although we ought to be careful about terminology. When I use the term "psychological continuity" I do not use it in the way a functionalist about psychological states would use the term. I assume for the moment that psychological states are not functionally defined and that a group agent does not have any psychological continuity. But nothing for our purposes here hinges on this debate in the philosophy of mind, because the functionalist will be able to accept the idea that phenomenological unity is not necessary for agency or responsibility. The mere fact that she has a different understanding of psychological states does not change that. Although the functionalist *would* claim that a group agent has psychological continuity, she would not take this as any evidence that phenomenological unity is a necessary precondition for responsibility and agency. Indeed the functionalist would have *additional* reasons to agree that agents can come larger or smaller than single human beings because the functionalist sees fewer disanalogies between the group agent and the typical agent. Both have psychological states that fulfill certain functions within a cognitive system. The fact that those psychological states span human bodies—and aren't phenomenologically unified—might seem less important for a functionalist.

§3.3 AGENCY AND THE UNITY OF CONSCIOUSNESS

To review: we have questioned the assumption that memory is essential to agency. The argument unfolded like this: the concept wasn't needed to explain our ethical intuition that agents are responsible nor the minimal definition of agency needed to explain that

intuition. By using our ethical intuitions about responsibility as a guide, a thought experiment demonstrated that it is the sharing of information—not memory itself—which is one hallmark of agency. Given this conclusion, it would be instructive to look at psychological continuity in general. Memory is just one facet of our psychological lives and just one piece of the psychological connections that have preoccupied theorists of personal identity. Having questioned the assumption that memory is essential for agency, we ought to question the assumption that psychological continuity is required for agency over time and the assumption that phenomenological unity is required for agency at any given moment.

To question the assumption I suggest that we recall what motivated our account of agency in the first place. We began with the ethical assumption of our own responsibility and asked what account of agency would be sufficient to explain this responsibility. So is phenomenological unity required for agency? It does not appear to be, given what is minimally required for an account of responsibility. Our working definition of agency includes the pooling of information, deliberation on the basis of that information and implementation of a project, plan or response in accordance with that deliberation. But just to be safe, we should consider both possibilities: that agency *does* require phenomenological unity and that agency *does not*. Call the former agency(p) and the latter agency *simpliciter*. Agency(p) exists in human-sized agents who exhibit phenomenological unity by virtue of their cognitive structure. Agency *simpliciter*, in addition to being realized in human-sized agents with phenomenological unity, is also realized in other systems whose sole exclusion from the class of agency(p) is their lack of phenomenological unity. These systems function and act exactly like agents(p) with the sole difference that they do not exhibit behavior that requires us to attribute, naturalistically, phenomenological unity to their mental states.

What is the nature of the difference between these two kinds of agency that is allegedly so important? The difference, of course, is phenomenological unity at one time and psychological continuity over time. The question, though, is how deep this

phenomenological unity really goes. Is it deep enough to have moral significance, i.e. to matter for things like responsibility?

There are grounds for some skepticism here by questioning the very concept of phenomenological unity. Consider, for example, the analysis of split-brain patients offered by Nagel in his essay "Brain Bisection and the Unity of Consciousness."[50] In deciding how many minds these patients have, Nagel considers the following options: (1) they have one mind associated with the left hemisphere (and the responses associated with the right hemisphere come from an automaton); (2) they have one mind and the responses from the right hemisphere are not associated with any mind at all; (3) they have two minds; (4) they have one mind that is dissociated; and (5) they have one mind most of the time but two minds during experimental situations. Nagel argues that neither option is particularly satisfactory and that it is "the idea of a *single* person, a single subject of experience and action, that is in difficulties."[51]

The problem with the first option is that the right hemisphere—by itself—is capable of controlling a body and organizing sophisticated and integrated mental processes. Although the capacities of the right hemisphere may not equal the left hemisphere, they are sophisticated enough that a patient who loses the functionality of her left hemisphere would *not* be a good candidate for the label "automaton". And if a person with only a right hemisphere is not an automaton, there is no reason to attribute to an automaton the responses from the right hemisphere of a split-brain patient. For the same reasons the second option is not plausible, according to Nagel, because we would not be inclined to say that a patient with only a right hemisphere has no mind at all.

---

[50] See Thomas Nagel, "Brain Bisection and the Unity of Consciousness" in *Personal Identity*, ed. John Perry (Berkeley: University of California Press, 1975), p. 227-245. Nagel was the first to bring the case of split-brain patients to the attention of the wider philosophical community. Doctors removed the corpus callosum from these patients in an effort to treat their neurological disorders. The result was that the direct neurological connections between the two hemispheres was cut. By segregating auditory and visual inputs to the respective hemispheres during lab tests, doctors concluded that each hemisphere might be supporting an *independent* stream of consciousness.

[51] Nagel, p. 227.

The argument against possibilities (3)-(5) is the impossibility of deciding among them, according to Nagel, although I will soon offer reasons to support (4). One could say that the patients have two minds, which nicely explains their responses during experimental segregation, but the "two minds" declaration runs counter to our sense of their near-total integration and coordination of activity during all other times. The dissociation that is elicited by experimental situations completely disappears when the experiment is over. Their activities are too normal to be attributed to two minds. Moreover, we understand their behavior by thinking of them as having one mind and we treat them as having one mind (again with the exception of the experimental situations). This integration, which is so total that one might be unable to pick out a split-brain patient in everyday life, is performed by the sharing of information between the hemispheres. Unable to share information via the corpus callosum, the hemispheres make use of the avenues of cooperation still available to them: visual and auditory clues that help explain what the other hemisphere is doing: the result is cooperation whose efficiency is lower than our typical cooperation but is nonetheless efficient *enough* to allow the patient to live an integrated life, perform coordinated activities, and pursue a life plan. In short, the patient is able to exercise unified agency.

The fourth proposal is that the patients have one mind that is severely dissociated. Nagel dislikes this proposal because it has difficulty explaining the severe nature of the disintegration exposed by the experiment. In such circumstances the hemispheres seem to operate at such cross-purposes with each other that it seems implausible to attribute such responses to a single mind. But following Rovane we can offer a slightly different reading of this interpretation. First of all, the dissociation is only achieved through an experimental situation constructed to elicit the very dissociation that it then observes. Secondly, the patient does exhibit "a poignant attempt to achieve unity" by *striving* for *rational* unity when faced with the phenomenological disunity created by the experiment. This is one good reason to adopt the fourth proposal that the patient has a single, dissociated mind. But the

reason has nothing to do with phenomenology, but rather the patient's commitment to rational unity even in the face of segregated streams of consciousness.

Nagel considers a fifth and final option: the patient has one mind most of the time but two minds during the experiment. The option codifies nicely both the integration witnessed during normal periods and the dissociation witnessed during experimental situations, but the drawback is that it commits us to saying that minds can be brought in and out of existence by merely segregating sensory inputs to the hemispheres. Indeed, if one were to alternate rapidly between segregation and desegregation we would be committed to saying that a mind was blipping in and out of existence like a light switch being flipped on and off. This seems absurd. So the fourth option still seems best: the patient has a single dissociated mind because he *attempts* to achieve rational unity. The point isn't so much whether the attempt is successful, the point is the attempt itself. After all, even "normal" human beings do not achieve total rational unity.

Nagel wants to take away a particular conclusion from the previous discussion and it is a conclusion that is helpful for our purposes. The phenomenology of the split-brain patients falls somewhere in between the case of normal patients with intact brains and the case of multiple patients engaged in cooperative activity. But it seems impossible to say anything more precise because the phenomenology of split-brain patients cannot be reduced to either side of the spectrum. The facts just don't support it.

Now here's the rub: not only should we be skeptical about the ascription of phenomenological unity to these patients, according to Nagel we should question the depth of such ascriptions to ourselves (i.e. normal patients with intact brains). He writes that

> The fundamental problem in trying to understand these cases in mentalistic terms is that we take ourselves as paradigms of psychological unity, and are then unable to project ourselves into their mental lives, either once or twice. But in thus using ourselves as the touchstone of whether another organism can be said to house an individual subject of experience or not, we are subtly ignoring the possibility that our own unity may be nothing absolute, but merely another case of integration, more or less effective, in the control system of a complex organism.[52]

---

[52] Nagel, p. 242.

Nagel is pressing the idea that our own so-called phenomenological "unity" is the result of nothing more than the cooperation and information sharing achieved by the corpus callosum and the "functional integration" which results. However, such functional integration is achievable by other means than the corpus callosum (as the case of the split-brain patients demonstrates); and furthermore, even normal patients with intact brains do not achieve total functional integration as the well-explored phenomenon of self-deception demonstrates. And so our own phenomenological unity is not as deep as one might have thought.

Nagel's argument can be adapted for our present investigation. We can draw an analogy between the cooperation of the two hemispheres of a single biological brain (who have made what Rovane would call a commitment to rational unity) and the cooperation of multiple human beings engaged in a coordinated activity (who have made a similar commitment). In the former case the hemispheres cooperate using a direct neurological connection that allows for the sharing of information virtually instantaneously. In the latter case the cooperation takes the form of overt communication and is mediated by the biological actions necessary to carry out the communication. Furthermore, the cooperation is not instantaneous and takes time to accomplish. These two factors mean that such cooperation is more difficult to achieve.

Besides the differences just noted, both are cases of *cooperation.* Both are examples of a commitment to rational unity. Of course, what *is* different is the *medium* of cooperation through which that rational unity is achieved. In the former case the medium is the corpus callosum and in the latter it is—usually—verbal communication. (Although it is possible that there might be other methods of cooperation). It would be wrong to deny that this is a difference, but so too would it be wrong to exaggerate this difference as a deep difference in kind. And so although it is important to concede that there is a different avenue of cooperation, it seems legitimate to question whether this difference is deep enough to carry moral significance. After all, our investigation has centered around an account of agency that

is sufficient to explain our ethical intuition about our own responsibility. And we decided that what was important for agency was the capacity to pool information, deliberate and exercise intentional action. And this process can certainly continue without so-called phenomenological unity. Although the hemispheric cooperation in patients with intact brains may make this process easier, it is by no means necessary. Cooperation sufficient to pool information, deliberate and implement a project, plan or response is possible without so-called phenomenological unity because a commitment to rational unity is possible without it. So, it is unclear how so-called phenomenological unity itself can be morally significant for responsibility. What *is* important is the hemispheric cooperation that is achieved with phenomenological unity. Since such cooperation can be completed with methods other than an intact corpus callosum, it is clearly the cooperation itself which is the source of the moral significance. And in both cases the cooperation is evidence of a commitment to rational unity.

§3.4 GROUP AGENTS AND MULTIPLE PERSONALITIES

To review: I have claimed that not only should the concept of the person be eliminated—it *can* be eliminated because it is dispensable. Everything that we need to say about persons can be said better with the components of the cluster concept. The burden falls on me to offer a convincing and coherent account of responsibility that makes no reference to persons.

This account took as its starting point our ethical intuition that we are responsible. With the assumption that we are responsible for our actions, we decided that whatever else we may be, we are responsible *qua* agents. And so we set upon the task of finding an account of agency that was sufficient to explain our responsibility and justify our ascriptions of responsibility. At the very least, such an account must include the sharing of information, deliberation and implementation of a project, plan or response. The assumption that memory is essential for agency was questioned. Furthermore, the significance of

psychological continuity and phenomenological unity was questioned by showing that these concepts are unnecessary to ground responsibility and consequently unnecessary for the minimal definition of agency needed to explain our ethical intuition that agents are responsible.

Given this portrait of agency it is essential that we consider the case of group and multiple agents.[53] The reason is simple: group agents differ from human-sized agents precisely because—among other things—they do not exhibit phenomenological unity. (The other major reason is that they are realized in multiple human bodies, not just one.) But if the significance of phenomenological unity is questioned, and what is really important is the sharing of information and cooperation that is *achieved* by this unity, then group and multiple agents ought to be included in our account of agency. And then here's the rub: the virtues of the eliminativist position become clearest in the case of group and multiple agents. If we don't eliminate the concept of the person, the existence of group and multiple agency forces us to prioritize between agency and bodily continuity—since the two diverge in group agency. Regardless of which concept is prioritized an intuition will be violated. But if the concept of the person is eliminated, no trade-off is required. For these reasons an in-depth analysis of group and multiple agency is called for.

We will start with group agency. Since the importance of phenomenological unity has been exaggerated, there is one less reason to deny the existence of group agents on a priori grounds. An example would be illustrative: the paradigmatic case of group agency is the corporation. Consider the board of directors of a large corporation. The board of directors is committed to pooling information, conducting deliberation on the basis of that information and implementing a plan. The goal of the plan is usually to make as much money as possible, with certain restrictions, in a given field of commerce. The board also will have a set of procedures according to which the deliberation will proceed. For example, each

---

[53] See chs. 4-5 of Carol Rovane's *The Bounds of Agency: An Essay in Revisionary Metaphysics* for an in-depth analysis of both group and multiple agents. My discussion of these issues is framed entirely by the conceptual landscape she offers in these chapters.

member of the board of directors might have the opportunity to state their case for their preferred course of action and the entire board would then vote, with a simple majority deciding the course of action to be pursued. There might also be rules for election to the board of the directors and the number of weighted votes to be given to each member. Furthermore, there might be provisions for an annual shareholders meeting by which the deliberation might include shareholders with voting rights in the corporation. It would not be necessary for the board to deliberate about each corporate action. Rather, the board might concern itself mostly with macro-policies and the appointment of executive officers whose duty is to execute the micro-decisions necessary to fulfill the macro-policies.

It is important not to misconstrue the claim: I am not suggesting that all, most or even some of the corporations in existence in the United States are group agents as I have defined them. Rather, the claim is that a corporation *would* be a group agent if it reached the level of integration described above. Whether any corporations actually do meet this standard is a purely *empirical* question on which I remain agnostic. But the empirical question is beside the point for my argument. The important claim is simply that if a group of individuals were to engage in a coordinated activity with a sufficient level of integration they would be agents. And it is unclear why their lack of phenomenological unity would be evidence against their status as agents. Furthermore, as agents these group agents would be responsible.

This accords nicely with our ethical intuitions about corporations. We consider them responsible when their corporate actions have disastrous consequences for consumers, for innocent bystanders, or for the environment. Insofar as these corporations are engaged in coordinated, goal-directed activity (usually for profit) they are responsible for the consequences. And our legal system has encoded this ethical intuition in several statues under which corporations can be held criminally or civilly liable. In most of these situations case law and legislation have defined corporations as legal *persons*. Indeed, in American

jurisprudence if a piece of legislation makes reference to 'persons' it is usually assumed to include corporations if not otherwise specified.

It is unfortunate that the law has assumed that in order for corporations to be responsible they must be legal *persons*. Although this label may seem commonplace to those in the legal profession, lay persons have trouble swallowing the idea that corporations are persons—precisely because our intuitions about persons include bodily and psychological continuity. Calling corporations 'persons' certainly violates these intuitions. What the law often fails to recognize is what this chapter aims to elucidate—that there is no need to call an entity a 'person' in order for it to be responsible. We could stop short and say that corporations are responsible insofar as they are *agents*. And the result is that our biological and psychological intuitions need not be violated.

There are several objections to the possibility of group agency. I will consider three of them here. The first objection is that a group agent cannot have an intention in the same way that a biological human being can have an intention. That's because an intention is a mental state that is realized in the physical brain of the individual biological organism to whom the intention is attributed. And having an intention is essential to performing intentional action; it is the hallmark of rational agency.

There are two potential responses. First, we attribute intentionality naturalistically in the same way regardless of the size of the agent. We attribute it based on behavior in the case of individual biological agents, so why shouldn't we do the same in the case of group agents? Both are cases of attributing intentions to further the goal of interpretation (in the Davidsonian sense). And why assume from the outset that the sole object of interpretation is a biological organism and *only* a biological organism?

The second response comes from Rovane: although group agents do not have direct intentional control over their behavior, neither do human-sized agents when executing temporally extended actions.[54] For example, when a human-sized agent decides to do

[54] See Rovane, p. 146.

something tomorrow, he must form a long-term intention that the action be completed tomorrow. But such long-term plans do not constitute direct intentional control because there is no guarantee that the future agent will follow the directive. In that sense, the lack of intentional control faced by the agent over time is analogous to the lack of direct intentional control of the group agent. Since no one would want to deny the existence of temporally extended agents (that being a *reductio ad absurdum*), no one could sensibly deny the existence of group agents on the same grounds. So the criteria of direct intentional control is obviously too strict then. It is not a necessary condition of agency.

The second objection claims that our working definition of agency requires them to come one to a body. That's because it is only with a functioning body that an agent can fulfill the third criteria of agency: the implementation of a project, plan or response. We are embodied agents not just because we are biological animals but because our very status as agents requires us to be physically realized in order to implement a project, plan or response. However, although an agent must be physically realized in *some* material substance in order to execute plans, there is nothing to suggest that the physical medium in question must correspond one-to-one with a single biological animal. It must correspond to some physical medium—most likely biological animals. But this also is consistent with the physical medium being a collection of biological animals or even a segment of a biological animal (as would be the case with a multiple agent). With the definition of agency that is necessary to account for our ethical intuition about responsibility, there is nothing essentially tied to the existence of a *single* biological body.

Here's the third objection: responsibility can be attributed to the individual agents who make up the corporation, so there's no need to posit the existence of a group agent just so you can justify your ethical intuition about responsibility. Attributing the collective responsibility to an aggregate collection of individual agents would be sufficient to account for our ethical intuition. But the psychological structure of the group agent doesn't equal the simple sum of the individual psychological states. You can't get the latter just by adding up

the former. There is one sense in which every 'action' performed by the group agent is performed by an individual biological human being and could be redescribed simply by listing the individual 'actions'. But here we are talking about physical action and bodily movements. Of course it falls to the biological entities to perform the component 'actions' that make up the actions of the group agent. But this is a very impoverished notion of action.

What is missing here is *deliberation*. The group agent exhibits a completely different deliberative structure than the sum of the individual agents and absent this deliberative structure it is impossible to make sense of the group agent's actions. Only this structure can explain how the agent arrived at all-things-considered judgment, in accordance with its deliberative policies, about what course of action was best. In short, only this deliberative structure can explain the rationality of the group agent.

Peter French makes the same point by contrasting crowds with corporations, where the former exhibit no common deliberation in contrast to the latter. This fact flies in the face of the Legal Aggregate Theory of the Corporation, which "allows that biological status has legal priority and that a corporation is but a contrivance, the name of which is best used for summary reference... [T]o treat a corporation as an aggregate for any purpose is to fail to recognize crucial logical and metaphysical differences between corporations and crowds."[55] Unfortunately, French makes the same mistake as many others in thinking that conferring agency upon corporations requires conferring personhood as well. Indeed, the title of his chapter is "The Corporation as a Moral Person."

One might retort that this seems like a perfect place for the Argument from Below. The significance derives from the lower level facts, the individual *components* of the group agent, so responsibility should attach at this level. But this would be a mistaken application of the Argument from Below. The significance (for responsibility) does not rise from the

---

[55] Peter A. French, *Collective and Corporate Responsibility* (New York: Columbia University Press, 1984), p. 35.

lower level facts for the simply reason that the lower level facts do not contain all of the relevant information; they are not mere conceptual facts, as Parfit might say. Something essential resides at the upper level: the deliberative structure of the group agent. Failure to recognize this fact is, in French's words, to confuse crowds with corporations. Agency is about more than just physical actions, it includes the information-pooling and deliberation necessary for an all-things-considered judgment. Once we realize this, it becomes clear that the *organization* of the group agent is essential and that the Argument from Below does not apply here. That's why certain situations compel us to ascribe responsibility to a group agent instead of individual human beings.

The link between *individual* human beings and agency having been broken (although the link between human beings in *general* and agency having been strengthened), the possibility opens up for the existence of atypical agents. So in addition to group agents we must allow for the existence of multiple agents within a single human body. Having claimed that agents must be physically realized (i.e. not disembodied) in order to meet their identity criterion as agents, there is nothing in their identity criterion which requires them to match up one-to-one with biological human beings. So in addition to one agent spanning several human bodies, there might be several agents cohabiting within a *single* human being. What is referred to as multiple personalities in the discipline of psychology are good candidates for multiple agents within a body.

A multiple personality is committed to sharing information, deliberating and implementing a plan of action. Furthermore, each multiple personality is committed to this cooperation within itself but not with the other personalities. Although some personalities may receive the same information (by virtue of their cohabitation in the same body), there is no *commitment* to sharing information (but they may be burdened with it). And they certainly do not deliberate together. Each multiple personality acts on its own, with its own talents, with its own interests in mind as a goal. Indeed, although multiple personalities

sometimes act in concert, they also act at cross-purposes with each other, as one would expect with two agents with different deliberative schemes and potentially different attitudes.

Once again, recognizing the possibility of multiple agents in a single body accords nicely with our ethical intuitions about responsibility (just as it does for group agents). If multiple personalities were to exist (more on this below), then we could consider it inappropriate to hold one multiple personality responsible for the actions of another. Conversely, we consider a multiple personality responsible for its own actions and the existence of a second multiple agent does nothing to lessen the responsibility of the first agent. In cases where we think a multiple personality is not responsible for their actions, it is usually because we think the case it is not one of bona fide multiple personality but rather of a single agent who is insane (and thus not responsible). But that's a separate issue. Our ethical intuitions tell us that *if* multiple personalities exist then they are responsible.

A skeptic might object here that in cases of diagnosed multiple personality disorder the patient is either engaging in outright deception, being influenced by an overly zealous psychiatrist, or simply engaging in a form of self-deception carried to a psychotic level. Many of us are inclined to choose one of these options in almost every case of alleged multiple personality disorder and there has been a substantial amount of scholarly work questioning the diagnosis of multiple personality disorder. But the objection misses the point—which is that *if* such entities existed, we would be strongly inclined to say that they are ***responsible*** for their actions and only their actions. In turn, we ought to say that they are agents. This does not require us to posit their actual existence. The claim that all cases of multiple personality disorder are fake is not an objection to the view that agency is what matters for responsibility. Quite simply, there is no empirical element to the position being asserted, so an empirical objection is not going to go very far.

Whether or not group agents and multiple personalities actually exist is an open question. It is an empirical question that is beyond the scope of this investigation. The relevant issue here is simply that *if* an entity meets the criteria—and it is a decidedly

empirical question whether they do—*then* they are an agent. (The claim is a conditional and the antecedent may or may not obtain.) Furthermore, as agents they are *responsible* for their actions. And this accords exactly with our intuitions about the connection between responsibility and intentional action. We believe that a corporation should be held morally and legally liable for its corporate endeavors (because such responsibility often can't be attributed to the individuals who constitute the corporation) and that it would be a mistake to hold one multiple personality responsible for the crimes of another.

Another objection to the view we are considering is that the existence of group agents is impossible to envision because it requires the disappearance of the individual agent—something that seems improbable. Does the human-size agent cease to exist when he decides to join a group agent? If this were the case, how could he decide to leave the group agent (since he doesn't exist anymore)? This objection can be dealt with by recognizing the existence of overlapping agents.[56] Conferring agency upon a group of human beings does not mean that the individual human beings *must* cease to be agents themselves. Recall that our definition of agency centers around a commitment to the pooling of information, deliberation and implementation of a plan. Since we've already denied that there is a one-to-one correspondence between human bodies and agents, since we've established that phenomenological unity is unnecessary for agency, and since we've already established that there can be more than one agent in a single body, then the basic materials for an account of overlapping agency are already at our disposal. In these cases an individual human being contains two agents: an individual agent and one piece of a larger, corporate agent to which the individual belongs or participates. So there is an overlapping of agents at the site of this biological human being. This explains how the individual agent persists and how the individual agent could choose to withdraw from the group agent. He would do this by withdrawing his commitment to pool information and deliberate with the other parts of the group agent.

---

[56] See Rovane pp. 203-208 for a discussion of overlapping persons.

The following example might be illustrative: a businessman decides to join the tightly-knit board of directors of his corporation. Part of the businessman's life is dedicated to his personal and familial obligations. But a second part of the businessman's life is dedicated to the goals of the corporation. Insofar as he participates in information pooling and deliberation with the board of directors, he is one piece of a larger group agent. An individual agent still exists that has not been subsumed by the group agent and is not committed to the goals of the corporation. This would be an example of overlapping agency.

Recall that I am not claiming that group agents, multiple agents or overlapping agents exist. Nor am I suggesting that corporations or multiple personalities are bona fide group and multiple agents. Nor am I even suggesting that the only coherent account of agency is one that admits of their possibility. One does not need to accept even the possibility of group and multiple agency—although I argue that one should—in order to embrace the eliminativist proposal. (All that one really needs to accept is that a convincing account of responsibility requires no reference to persons. Without that, the eliminativist option is unpalatable.) Rather, what I am suggesting is that the virtues of the eliminativist proposal come into focus if we consider an account of agency that allows for group or multiple agents. If one eliminates the concept of the person, then there is no need to prioritize the component concepts and identify one with personhood. In a case where agency and bodily identity diverge (as in the case of group or multiple agents), this prioritization will inevitably lead to the violation of one or more of our intuitions about personhood.

Furthermore, if one accepts this account of agency, as I think one should, it becomes clear that agency is a separate component concept distinct from the other sub-concepts of psychological and bodily continuity. Having questioned the link between phenomenological unity and agency (based on the requirements of fulfilling our ethical intuition about responsibility), we now have a clearer picture of the cluster concept of the person. In addition to the sub-concepts of bodily and psychological continuity, agency itself is a

separate and distinct component concept of the cluster that is independent of the other two components. Consequently, elimination of the cluster concept must be followed by replacement with *at least* the three component concepts. Doing so will produce a better conceptual apparatus without sacrificing our ability to explain our own responsibility. And indeed, doing so will yield a better account of responsibility because it explains (and justifies) our desire to hold corporations and multiple personalities responsible for their actions.

§3.5 AGENCY AND INTERPRETATION

The account of agency that I have presented is indeed a revisionary one. Remember, we started this investigation of agency to see if we could provide an account of responsibility that made reference to agents instead of persons. (Failure to provide such an account would require us to second-guess our eliminativist strategy.) Our investigation started with our ethical intuitions about responsibility and worked backwards to find an account of agency sufficient to explain those intuitions. According to the account, phenomenological unity and psychological continuity are not essential to agency. Consequently, multiple and group agents are at least conceptually possible. This revelation departs from the common understanding of agency in the philosophical literature but accords with a revisionary account of agency offered by Rovane and others.

When presenting a revisionary metaphysical proposal we usually want to know whether we can live in accordance with it. Failure to act in accordance with a revisionary metaphysical picture is, while not a fatal objection, certainly cause for some skepticism about the new proposal. So is it possible to treat group and multiple agents in the same way that we treat human-sized agents? To answer that question we need to say how we identify and interact with agents. As it turns out, the process of this identification and interaction—Davidsonian interpretation—not only allows for this revisionary metaphysical picture, it

supports it.[57] Consequently this appeal to Davidsonian interpretation does more than show that the account can be put into practice, it also shows that if we accept the Davidsonian programme then we should also accept at least the possibility of multiple and group agents.

In preceding sections I defended a view of agency that I think clearly shows the benefits of eliminating the concept of the person—although the account is not needed to make the case for eliminativism (it merely brings it into focus and shows how eliminativism can give us an improved understanding of responsibility). I explored the question of agency by considering two controversial examples: multiple personalities and corporations. These cases bring to mind this question: By what process do we discover what is an agent? Answering this question would go a long way to explaining how it is possible for us to treat entities such as multiple personalities and corporations as agents, just as we treat most human beings as agents. In short, it is necessary to consider agency from both ends: ourselves as agents and our *attribution* of agency to others. What connects these two things is Davidsonian interpretation and the need to make rational sense of the behavior we are confronted with and to attribute mental states on that basis. And looking at the process by which we attribute agency in controversial cases (such as multiple personalities and corporations) might help us decode the process by which we attribute agency in more garden-variety cases.

We attribute agency in these cases to further the goal of interpretation. Consider the case of multiple personalities. In such a case one is presented with behavior that, when attributed to a single agent (as is customary) seems blatantly irrational; one can't make sense of it except to say that the agent is acting irrationally or suffering from a massive case of self-deception or weakness of will or some such phenomena. But if the assumption that one is dealing with a *single* agent is thrown out, new possibilities emerge to explain the behavior and make rational sense of it. In this case we might see this human being as genuinely having

---

[57] The connection between interpretation and agency, and its relationship to Davidson's work, was advanced by Carol Rovane in "Rationality and Identity" in *The Philosophy of Donald Davidson*, edited by Lewis Edwin Hahn (Chicago: Open Court, 1999).

multiple personalities that are acting as independent agents. The point here is that not only is the behavior of the agent subject to interpretation in the Davidsonian sense, but the number of agents itself is subject to interpretation.

This claim needs to be explained in depth.

According to the Davidsonian model, reasons provide a causal explanation for an agent's action. In the light of observed behavior we attribute reasons in order to arrive at a complete theory of an agent's thought, talk and action. In short, we try to *understand* her, to interpret her utterances and her actions. And we attribute the beliefs and desires needed to explain this behavior. This process of Davidsonian interpretation, usually an unproblematic and unnoticed affair, becomes slightly more complicated when an agent performs an action that violates her own judgment about what rationality demands. These are the well known cases of weakness of the will and self-deception, where the assumption of rationality (i.e. the principle of charity) does not yield a coherent theory of an agent's thought and action. But according to Davidson, weakness of will cannot simply be a result of the agent's incontinence; rather it is a failure of rationality in the sense that the agent fails to act in accordance with an all-things-considered judgment giving her a reason to do the contrary. Davidson asks,

> Why would anyone ever perform an action when he thought that, everything considered, another action would be better? If this is a request for a psychological explanation, then the answers will no doubt refer to the interesting phenomena familiar with most discussions of incontinence: self-deception, overpowering desires, lack of imagination, and the rest. But if the question is read, what is the agent's reason for doing *a* when he believes it would be better, all things considered, to do another thing, then the answer must be: for this, the agent has no reason. We perceive a creature as rational in so far as we are able to view his movements as part of a rational pattern comprising also thoughts, desires, emotions and volitions. (In this way we are much aided by the actions we conceive to be utterances.) Through faulty inference, incomplete evidence, lack of diligence, or flagging sympathy, we often enough fail to detect a pattern that is there. But in the case of incontinence, the attempt to read reason into behavior is necessarily subject to a degree of frustration.
>
> What is special in incontinence is that the actor cannot understand himself: he recognizes, in his own intentional behavior, something essentially surd.[58]

---

[58] Donald Davidson, "How is Weakness of the Will Possible?" in *Essays on Actions and Events* (New York: Oxford University Press, 1980), p. 42.

It is clear from Davidson's analysis that both of these special psychological phenomena—weakness of the will and self-deception—are posited in the face of evidence of an agent's irrationality. I am claiming that the principle of charity warrants an ascription of weakness of will or self-deception only for a *range* of irrational behavior. The observed behavior might seem irrational if we cannot detect a pattern in intentional movements which we *assume* to be part of the same agent simply because it issues forth from the same body. But if one were able to detect a pattern by dropping this assumption we would be justified in doing so. We would then be able to detect a pattern under a different assumption—one which did not assume that the boundaries of rational agents and biological human beings always and necessarily coincide. Behavior that is so irrational that even weakness of will or self-deception will not explain it might be interpretable as the rational behavior of an atypical agent, such as a group agent composed of several human beings (a corporation), or an agent sharing, with others, a single human body (a multiple personality).

The point here is that the model of making sense of behavior—of arriving at a unified theory of an agent's beliefs, utterances and actions—is more than something we apply to agents that we have antecedently identified. It is the model we use to determine *how many agents there are.* It is also the model we use to determine *which* agents there are. We do not simply use the criteria of rationality to explain behavior of an agent that has been defined for us, *deus ex machina.* Rather we use the criteria to explain the observed behavior, which we presume to be more or less rational, and go on to attribute the behavior to the number of agents, with the appropriate duration, that maximizes the rationality of the observed behavior. The goal of finding meaningful patterns is not just the manner by which we explain an agent's behavior, *it is the manner by which we "find" agents themselves.*[59] It is

[59] Davidson himself does not take this position, despite Rovane's claim in "Rationality and Identity" that this refusal puts him in some tension with his own overall views about rationality and interpretation. In responding to Rovane's claim, Davidson has written that "our mental vocabulary is entirely concerned with describing and explaining what animal bodies do and are," perhaps explaining why he thinks that biological human beings are the proper objects of interpretation.

part and parcel with the project of agent counting and is required by nothing less than the Principle of Charity itself.

If the empirical situation should ever present itself, we would have a good warrant for giving up the assumption that agents come one to a body. The assumption falls victim to the demands of charity, which compels us to explain the behavior in question as rational. A tenacious reluctance to give up the assumption violates charity for the sake of a metaphysical commitment (or a metaphysical prejudice, perhaps). If recognizing the "death" of an agent and positing the existence of a new agent serves the purpose of detecting a rational pattern, then only the most metaphysical of metaphysical commitments would serve as a prohibition against doing so. The point here is simply that charity would provide no basis for the prohibition. The prohibition would have to be external to the game that is the hallmark of agent counting: arriving at a unified theory of an agent's thought, talk and action.

Consider a case of radical self-deception in which a human being claims that she is a "new" agent that has replaced an "old" agent. Let us presume that there is an appropriate causal story to explain this, such as a massive psychological or physical trauma, coma, brain surgery, or a progressive neurological disease that has unexpectedly reversed itself. Let us stipulate that these causal stories are legitimate and can be taken as evidence that the present agent is not being insincere. The agent insists that she is not the same agent who was previously identified with her human body. She is motivated to correct what she perceives to be the public's misconception of her (as being the same agent as the previous one) in order to be relinquished from the commitments of the previous agent. She might, for example, argue that she does not need to get divorced because she is not married. It is the previous agent who contracted to be married and there is no basis to compel *another* agent to fulfill that contract, and so no divorce can be granted in a situation where no marriage exists. Or perhaps the agent argues, along similar lines, that she need not repay her student loans because it is the previous agent who used the money to attend college.

If this empirical situation were ever to present itself we would be justified in eschewing self-deception in favor of a more radical ascription—the existence of two successive agents in a single human body—assuming that we need to make that decision to make rational sense of the observed behavior. This justification rests on the fact that we are required to make the attribution to fulfill our other commitments—one of which might be, for example, our goal of successful communication. When we communicate we must arrive at a unified theory of our interlocutor's utterances and actions, and this might only be possible by rewriting our assumption that agents come one to a body.

That being said, it may be the case that positing weakness of the will or self-deception (and not ascribing the existence of atypical agents) might better serve the purposes of proper agent counting. It might be rational to approach a human being and in attempting to reason with it posit the rather banal existence of a single, temporally unified agent. It is at least theoretically possible that we could be presented with a third, more problematic state of affairs, where it makes just as much sense to posit the existence of an atypical agent as it is to attribute weakness of the will or radical self-deception to a single agent (because the attributions equally maximize the rationality of the observed behavior). In such a case it is unclear what evidence could possibly require us to go one way over the other. And it does not seem plausible to suggest that metaphysical prejudices should be wheeled in for tie-breaking purposes. Rather, it seems clear that one is left with competing (and equally compelling) theories of an agent's thought and talk. One would be left with a legitimate *decision*. But this is a rather theoretical anxiety. In nearly all cases (if not all), rationality will demand that we go one way over the other.

Consider a hypothetical corporation that meets my criteria for agency. Because human beings who form a corporation pool information relevant to their corporate endeavors and because these human beings then deliberate on the basis of that information and implement a project, plan or response, there is a legitimate temptation to treat these human beings as a collective agent. Or one can attribute reasons and explain the observed

behavior by attributing it to many (perhaps hundreds or thousands) of individual agents with their own reasons and discrete causal chains for their actions. What is the criterion for making this attribution? In the case of my hypothetical corporation, the way to maximize rationality—as demanded by Charity—is not by attributing reasons to all of the individual agents. The attribution of reasons to a corporate group agent maximizes the rationality of the observed behavior.

The distinctive stance that we adopt towards an agent during radical interpretation goes deeper than just understanding what an agent means. It also penetrates into all rational interactions with an agent, from persuasion and lobbying to seduction and deception. In all such rational interactions we assume that the agent will be responsive to reasons. Many of our interactions take the form of giving an agent a new reason to do something—perhaps by helping him learn a new piece of information that he did not know, perhaps by lying to him. Regardless, the agent must be responsive to reasons if such interactions are to yield success. This is what separates inanimate objects from rational agents.

This is all part of radical interpretation because the goal of interpretation is to construct a unified theory of an agent's beliefs, desires, utterances and action. And reasoning (broadly construed) with an agent requires that we have such a unified theory. It is required not just because reasoning so often takes the form of verbal communication. Even in such cases where no communication is involved, reasoning is only possible once we've interpreted the reasons that form the causal chain to the agent's actions. Reasoning with an agent without such a theory would be impossible because there would be no direction, no road map. So all reasoning with an agent is part of the larger process of interpretation, of constructing this unified theory.

How might we reason with a group agent? (This task is more demanding than simply explaining its behavior.) The most rational course of action is to treat the group as one agent (if interpreting it in this way maximizes the rationality of the observed behavior) and then provide it with reasons sufficient to alter its conduct. In the case of a corporation,

we provide reasons to change its behavior on the basis of its economic self-interest. This is certainly the rationale at work in most legal jurisdictions where corporations meet the legal definition of personhood and can be held civilly (and sometimes criminally) liable for their conduct (and are consequently subjected to compensatory and punitive fines). This decision is warranted by the fact that charity demands that we assume the behavior to be more, rather than less, rational. By attributing agency to the group—as opposed to just the individuals—we are able to construct a unified theory that explains the observed behavior and forms the basis of our interactions with the group. And as we saw in the previous section, to attribute responsibility solely to the individuals would be to ignore the deliberative structure of the group, to confuse crowds with corporations.

To review: we have established that agency consists in the pooling of information, deliberation and implementation of a project, plan or response on the basis of that deliberation. This does not require phenomenological unity or psychological continuity, so group and multiple agents are at least a conceptual possibility. And we pick out these agents as part of the process which Davidson called radical interpretation. So it's no wonder that we have to admit that group and multiple agents are at least possible. In radical interpretation we attribute reasons as a causal explanation for an agent's actions. Since what produces that action is information-pooling and deliberation—not phenomenological unity itself—the process of radical interpretation will allow attributions of agency that do not exhibit phenomenological unity. Indeed it must.

But this account of agency can answer more than the question: "What counts as an agent?" We can now extend this model to deal with agent identity over time. We have explored how agency attributions are part of the larger game of behavior interpretation and constructing a unified theory of an agent's thought and action. We can see now that the same attitude is appropriate for agent identity over time, for we attribute actions to a single, unified agent insofar as charity requires it. We ascribe actions to a single agent until it becomes too difficult to explain the behavior in question or to influence future behavior. At

that point we can either posit a new agent—as a successor to the old agent—or posit the existence of multiple agents at a time, although it seems improbable that such an empirical situation would often present itself that would make such an attribution rational. It is more likely the case that, when confronted with a functioning human being, one attributes to him or her the existence of a single, temporally unified agent that is responsive to reasons. We treat human beings in this way because we need to, for it allows us to construct intentional explanations, predict behavior, and then provide reasons that will influence their behavior.

What I am therefore suggesting is that we broaden the general line that our agency attributions are dependent on the game of interpretation (and consequently the demands of rationality) to the further idea that agent identity over time is just as much entailed by Davidsonian interpretation. That's because just as we are not entitled to assume that agents come one to a body, neither are we entitled to assume that an agent's identity over time corresponds to the identity over time of a biological human being. Both assumptions are up for grabs during the process of interpretation. The input of interpretation is not behavior that issues from an agent whose identity is fixed or given by some other procedure; the input is behavior *simpliciter*. And the very boundaries of the agent in question are to be determined within the context of interpretation.

With this connection between agency and radical interpretation, we are now in a position to explain how and why we hold agents responsible (in the form of punishment). It should be obvious at this point that holding an agent responsible can only take place within the larger context of interpretation. In interpretation we are attributing reasons as a *causal* explanation of an agent's behavior. When we want to influence that behavior, as is often the case when we punish someone, we do so by giving the agent reasons to alter their behavior. In a sense, we identify the reasons that caused the agent's behavior in the past so that we can insert new reasons into the causal chain to influence the agent's behavior in the future. For example, we give an agent the following reason not to commit a crime: by threatening him with punishment, we try to make it rational—by his own lights—to comply with the law.

This is the rational machinery of deterrence. (The degree to which deterrence is successful is another story.) We try to engineer the situation such that the cost of the risk of punishment is higher than the gains of the criminal act, therefore producing an all-things-considered judgment by the agent that the cost is too high. It does not always work out that way—but that's the goal.

So even the process of punishing and holding agents responsible is part of the larger game of interpretation. Altering the causal chain of an agent's behavior by inserting reasons is only possible if one can make rational sense of the behavior in question (which is the goal of interpretation). If one is unable to make any rational sense of the behavior one must give up one's conception of the object as a rational agent at all, indeed as any kind of agent. And that conception is a necessary condition for punishing someone in the hopes of altering their future conduct.

Our need to hold agents responsible for their actions brings up an obvious objection to the possibility of group and multiple agency. According to this objection, the fact that group agents cannot be punished is evidence for the absurdity of claiming that group or multiple agents exist. Because punishment is meted out to particular human bodies (i.e. it is human beings that are put in jail), it is impossible to punish agents who do not correspond one-to-one with biological human beings. The fact that we cannot hold them responsible for their actions is good cause for skepticism about their metaphysical possibility.

But the objection is misplaced. First of all, it is overstating the case to suggest that punishing atypical agents is impossible. It *is* possible, although our practices of legal punishment would need to be revised if this possibility were more than a rare occurrence. Physical imprisonment is not the only form of punishment available; there are many alternatives ranging from monetary fines to loss of social status and privilege as well as public condemnation. And even physical imprisonment is *possible*, it's just a bit impractical. In the case of group agents, the human beings involved could be imprisoned—a viable alternative for the board of directors of a corporation that needs to be punished. And although it is

technically unfeasible, it is at least *logically* possible to imprison a multiple agent when he is in control of a biological body and then release the biological body when another agent emerges. All of these options are available for the punishment of the atypical agent. And the fact that our current system of punishment is ill-equipped to handle such cases is just evidence that our system was designed with our current needs in mind: the punishment of *typical* agents. Were those needs to change, no doubt the system would change in response. None of this can be taken as evidence that atypical agents are logically or metaphysically impossible.

There is an important lesson to be learned from Rovane's point about the connection between Davidsonian interpretation and agency. We have shown in previous sections that psychological continuity and phenomenological unity are not required for an account of agency sufficient to explain our ethical intuitions are responsibility. With the help of Davidson and Rovane we have now *modeled* how this account of agency actually works. If we are right, then agency is a *distinct* concept from both physical continuity and psychological continuity—the two component concepts that we had already identified.

If the virtues of eliminativism were clear when dealing with just two concepts (as we were in the Williams thought experiment), the virtues of eliminativism ought to be doubly clear now that we have identified *three* discrete components of the cluster concept of the person. Just as eliminating the concept of the person allows us to identify a psychological or physical continuant without ascribing *personal* identity, so too does it allow us to identify a continuing agent without ascribing personal identity. Since interpretation requires us to make agency attributions that go beyond the confines of individual biological bodies, it is a good thing that we can do so without ascribing personhood, for doing so would violate our physical and psychological intuitions about persons. The final section is devoted to this issue.

§3.6 ARE AGENTS PERSONS?

This issue brings us head-to-head with the accounts of personal identity offered by Rovane—who *does* ascribe personhood to agents.[60] And the conflict with Rovane is especially threatening to the eliminativist strategy, since my motivations for suggesting elimination are the same as her motivations for offering a revisionary account of persons: our conflicting intuitions exposed in the Williams thought experiment. If agency is a distinct concept independent of psychological or physical continuity, as we both agree, then perhaps Rovane's strategy to break the deadlock between psychologism and physicalism is preferable. Indeed, Quine's maxim of minimum mutilation might *seem* to support Rovane here: a revisionary proposal is better than an eliminativist proposal because the former is less radical. The claim here would be that no proposal should be more radical than is needed, and elimination of the concept of the person is a more radical remedy than needed to resolve our conflicting intuitions. Equating persons with agents, as Rovane does, might do less collateral damage. And we could wheel in ethical intuitions to support this revisionary proposal. Our ethical intuition that agents are responsible could be evidence that our account of personhood—and hence personal identity—should be agent-centered. So we could follow Rovane and say that persons are agents and still get an alternate account of personal identity that breaks the deadlock between the animalist and the psychologist.

But, as we shall see, this is not the case. Quine's maxim of minimum mutilation will end up supporting elimination for the simple reason that the seemingly less radical strategy of equating persons with agents actually requires us to mutilate the very intuitions we were trying to resolve. We end up ascribing personhood to group and multiple agents and in the process do damage to our intuitions that elimination avoids. To see this, we should turn to Rovane's analysis of the problem of personal identity.

Rovane summarizes her argument this way: "(1) we cannot resolve the philosophical dispute about personal identity, between the proponents and the opponents of Locke's

[60] My argument in this section also applies to the agent-centered account of personal identity offered by Korsgaard as a response to Parfit's work.

distinction between personal and animal identity, without revising some aspect of our commonsense outlook; (2) since both sides of the dispute are coherent and well supported by common sense, we cannot strictly prove that one side or the other must be correct; (3) we must seek positive reasons to embrace one side or the other anyway; (4) we must seek these positive reasons in a substantive account of the kind 'person'." Rovane's substantive account will be largely ethical, i.e. it will highlight the degree to which persons have a unique capacity to engage in agency-regarding relations. Points (1)-(3) roughly correspond to the agreement between Rovane and myself. But (4) is where the disagreement begins. The resolution need not come from a substantive account of the kind person. Although I have no problem with the substantive account of *agency* that she offers, a satisfactory resolution to the problem of our conflicting intuitions about personhood might come from simply abandoning the concept. And since it is a cluster concept it is ripe for elimination.

There are at least two reasons to recommend elimination over the substantive, agent-centered account offered by Rovane.

First, once you have accepted that agency is a distinct concept independent of psychological and physical continuity, and once you have accepted at least the theoretical possibility of group and multiple persons, equating persons with agents violates a firm intuition: our deep conviction that persons come one to a body. Put simply, we assume that being a person is tied to the existence of a single biological body. And calling a group or multiple agent a 'person' violates this assumption. My hypothesis requires violating no such intuition. And so my position is consistent with ***both*** intuitions. I need not give up one for the other.

Eliminativism allows us to say everything we want to about group and multiple agents while saving us from calling them persons and violating a major intuition, as other theorists would. These theorists either construct a hasty and ill-advised distinction between "legal" and "metaphysical" persons (and then claim that group agents qualify for the former label but not the latter) or they bite the bullet (and violate the intuition) and just call them

persons *simpliciter*. But we can eschew both of those unpleasant options. This chapter has demonstrated that what is needed for a convincing account of responsibility is a suitably sophisticated account of agency and intentional action. Since the appeal to persons is unnecessary, it is unclear why a substantive account of persons is needed at all. Given the metaphysical confusion surrounding this cluster concept and given its dispensability in matters axiological, elimination is an attractive proposal. No mention of persons is needed to make a case for responsibility, and I will argue in subsequent chapters that no mention of persons is needed for coherent and convincing accounts of self-concern and rights either. Given all of these factors, it seems ill-advised to violate our animalist intuitions about personhood (in the form of an agent-centered account of personal identity) just to resolve our conflicting intuitions in the Williams thought experiment. Although the resolution is coherent, its cost is too high.

One might object that we have the very same animalist intuitions about agency that we do about persons. Not only do we assume that persons come one to a body, we also assume that agents come one to a body. That is, the argument for eliminativism flows from the realization that an agent-centered approach to personal identity requires us to violate our animalist intuitions about persons. But if we eliminate the cluster concept and replace it with its components—one of which is agency—then we face the same problem with agency. The burden of violating the animalist intuition is not relieved—it is just pushed one level down. Instead of violating our intuition that *persons* come one to a body, we violate our intuition that *agents* come one to a body. According to this objection, then, the supposed benefits of eliminativism fail to materialize.

The objection can be met with the following point: our animalist intuitions about personhood are much stronger than our animalist intuitions about agency. Although we have intuitions about both, the objection only goes through if it can be demonstrated that our animalist intuitions in both domains are equal in strength. But this is almost certainly not the case. Indeed, our animalist intuitions about agency are much weaker and can be

revised upon careful reflection. The difference can be summarized as follows: the Williams thought experiment demonstrated the degree to which our animalist intuitions about personhood are persistent and pervasive. They are not easily dismissed. Indeed, they may be impossible to dismiss. In contrast, careful reflection about agency allows one to reject the assumption that agents come one to a body. While in one case the intuition is resilient, in the other the intuition is revisable. Although we might assume, to begin with, that agents come one to a body, that assumption is easily replaced after careful reflection. Reflection reveals that the hallmark of agency is intentional action and the information-pooling and deliberation that precedes it. This definition establishes agency as a distinct concept independent from physical and psychological continuity. So agents need not come one to a body. It is at least theoretically possible that an agent could span several biological bodies or occupy only one part of a biological body. The result of this realization is that our animalist intuitions about agency disappear. But our animalist intuitions about personhood, exposed by the Williams thought experiment, are not so easily resolved. Consequently the objection fails because the violation of intuition is *not* just pushed one level down. Although perhaps it does not disappear entirely, it is nonetheless reduced.

There is a second reason why elimination is preferable to equating agents with persons. Doing so would violate our intuitions about human beings that do not meet the criteria for rational agency, such as young infants or the irreversibly unconscious. These human beings do not meet the definition of agency and the point at which they *do* become agents (in the case of newborns) is gray. Those committed to an agent-centered account of personal identity will in turn have to deny that these human beings are persons. This might violate some of our intuitions about personhood.

Many want to confer personhood upon these human beings based on one of two intuitions: they have biological bodies which some consider to be constitutive of personhood and they are appropriate objects of moral respect. Equating agency with personhood entails that these individuals may not be persons—a conclusion which violates some of our

intuitions. My eliminativist position would not run afoul of this intuition. Simply put, the violation doesn't come from the claim that these human beings aren't *agents*. The violation comes when the agent-centered theorist is forced into the following entailment: they aren't *persons* either. This causes an uncomfortable dilemma. It violates not only the animalist intuition (since these human beings *are* functioning biological animals) but also some ethical intuitions as well (that these individuals are appropriate objects of moral respect).

Eliminativism responds to this anxiety by denying the importance of the concept of the person. There is no longer any violation of the animalist intuition because we do not deny that these human beings are functioning human animals. Furthermore, eliminativism insists that the moral significance travels upward from the lower level components of the cluster concept (agency, biological continuity and psychological continuity); it does not migrate downwards from the cluster concept. Consequently, with the issue of personhood out of the way, we can recognize the startling fact that we decide to treat these human beings as objects of moral respect for reasons external to personhood, and sometimes even agency. In short, we can recognize that certain *biological* facts may be the source of our moral concern for these human beings.

It is important, however, not to exaggerate the differences between this eliminativist approach and an agent-centered account such as Rovane's. The difference is partly one of terminology, of notation. In the end, the difference between calling an entity a person or an agent *simpliciter* may not amount to much. Both positions recognize that the conflict in our intuitions will never be resolved in favor of either the animalist and psychologist positions. The only hope is a third way. One account seeks a revisionary proposal about agency, the other calls for elimination of the concept altogether.

But it is important not to underestimate the difference between the two positions either. There is an innovation of real value in the eliminativist strategy. That innovation amounts to a challenge against the prevailing assumption that the concept of the personhood is indispensable because of its singular role in our value theory. This sets the eliminativist

strategy apart from not just Rovane's account, but also the vast field of personal identity literature which has tackled its research under this assumption ever since Locke. As we have seen in this chapter, the assumption proved very weak in the case of responsibility. As for other elements of our value theory, we will have to see.

In this chapter we have pursued the eliminativist strategy by searching for a coherent account of responsibility that makes no reference to persons—a necessary step if the strategy is to be considered viable. Our investigation unfolded by examining our ethical intuitions about responsibility as a guide. Those intuitions demonstrated that information-sharing and deliberation—not psychological continuity or phenomenological unity—were important here. This formed the outlines of an account of agency which permits both multiple and group agents—which means that the proper objects of interpretation aren't just biological human beings. This accords nicely with our desire to hold group agents such as corporations legally and morally responsible. And it is here that the eliminativist strategy demonstrated not just its viability but its virtue. Only eliminativism allows us to confer agency upon these entities without also conferring personhood.

Having traced a line of argument that starts with responsibility, goes on to agency and then ends up back with responsibility again, we might be accused of engaging in a circular argument. But the circle here is anything but vicious. We have done precisely what we aimed to do: paint a picture of responsibility, agency and interpretation that makes no reference to persons. We have appealed to our ethical intuitions about responsibility not as a foundation for constructing a metaphysical theory. This would be a severe misreading of this chapter. Rather, our ethical intuitions about responsibility have served as a baseline against which to judge the success of our endeavor. Remember, the goal of this inquiry was to relieve our anxiety that eliminating the concept of the person would irreparably harm our value theory. We wanted to know if there was something in our ethical intuitions about responsibility that *required* us to posit the existence of persons. So we started with our ethical

intuitions about responsibility to see if they could yield an account of responsibility that made reference only to agents—not persons. As it turned out, they can.

I argued in the first two chapters that the Williams thought experiment exposes our deeply held conflicting intuitions about personhood. I took this conflict as evidence that personhood is a cluster concept and argued that a satisfactory resolution was best achieved by eliminating the cluster concept in favor of its components. I later identified those components as psychological continuity, physical continuity and agency. In order to demonstrate that the strategy is plausible, it is necessary to show that the sub-concepts will be sufficient for the uses to which we put the concept of the person. If the sub-concepts are not sufficient the strategy is in peril. Personhood would be indispensable after all and not a good candidate for elimination. To this end, the previous chapter was dedicated to showing that a coherent and convincing account of responsibility can be given without making reference to persons. This account was not only coherent, it illuminated that the relationship between personhood and responsibility actually tracks the component concept of agency. This realization helped provide a more nuanced account of responsibility that, among other things, codifies our ethical intuition that group and multiple agents are responsible entities. Furthermore, the eliminativist account of responsibility allows us to codify this ethical intuition without violating any of our animalist intuitions about persons—a cost that a traditional person-centric account of responsibility must bear.

In addition to responsibility, self-concern is another phenomenon whose analysis makes heavy reference to the concept of the person. It is rational for us to demonstrate prudential concern for ourselves *qua* persons. Furthermore, the Williams thought experiment, which exposed our conflicting intuitions, was, first and foremost, a thought experiment about self-concern. Indeed, it was the very link—the *assumed* link—between self-concern and personhood which helped expose the conflict in our intuitions. And it is precisely this conflict which I think can not only be resolved—it can be avoided before it arises. This is possible if we embrace eliminativism. Furthermore, if we reject the assumed link between personhood and self-concern, we can not only avoid the dilemma but we can still offer a convincing account of self-concern. Indeed, it is my hope that eliminativism will

liberate us from taking our expressions of self-concern as evidence of one theory of personal identity over another. Let me explain this point by revisiting the torture thought experiment.

The concept of the person, we are told, is essential because what matters in survival is the continued existence of a person. What matters is that *we* continue to exist, and we are *persons*, whatever that is. So, the story goes, if we figure out what we value in our survival, we will decode our implicit beliefs about the content of the concept of the person. It is in that spirit that theorists of personal identity have used self-concern as an investigative tool for figuring out what we mean when we call ourselves persons.

Williams recasted the old Locke-Shoemaker thought experiment in terms more explicitly related to self-concern.[61] This new version of the experiment exposed our conflicting intuitions about the primacy of bodily and psychological continuity. While Shoemaker used the thought experiment to advocate for a psychological criterion of personal identity, Williams performed the thought experiment twice, the first time eliciting a powerful argument for the psychological account, but the second time eliciting a convincing case for a bodily criterion of personal identity. On the first try, the subject (A) is asked who should be tortured: the person resulting from your (A's) transferred psychology programmed into the brain of B's body, or the person resulting from your (A's) body receiving B's psychology programmed into your brain. The answer, based on self-interest, is that the pain should go to the A body programmed with B's mental states, for it seems plausible to describe the situation as a case of "switching bodies".

During the second trial of the thought experiment, Williams asks the reader to imagine the following scenario: I am told that tomorrow I will be tortured. I'm also told that

---

[61] It is important to remember its origins in Locke's experiment—and Shoemaker's neo-Lockean version of it—of the prince and the cobbler. The whole point of asking where personal identity goes in the case of switching consciousnesses (or switching bodies, depending on which questions you choose to beg) is to put pressure on our quotidian notion of what it means to be a person. In our everyday lives, bodily and psychological continuity rarely diverge, so the vague concepts of 'person' and 'personal identity' are sufficient for our uses. Under tightly controlled experimental situations one can put pressure on the concept and force the subject of the *Gedankenexperiment* to choose (or prioritize) between competing accounts of personal identity.

by the time the torture is administered I will not remember being warned about the torture; a procedure, before the torture, will make me forget what I was told. Not only will I forget the warning, though, I will forget most of the things which I now remember. Williams points out that this would be little consolation since I can *fear* getting into some kind of accident which results in amnesia and a lot of pain. In neither case does the prospect of the amnesia lessen the fear about the upcoming pain. Furthermore, I am told that I will lose far more than just memories—dispositions and impressions will be eliminated and replaced by new ones before the torture is to begin. I also am told that the dispositions and impressions that I will receive are the result of being 'copied' from the brain of another individual and placed in mine. Once again, this news does little to lessen the fear, since Williams points out that I can imagine going crazy and thinking that I am George IV before undergoing the torture; it seems as if 'going crazy' though does nothing to ease the prospect of being tortured. If anything, it exacerbates it.

The problem, of course, is that the second case is the same as the first one, except that the story about what happens to B's body (the next closest candidate for being *you)* is left out of the description. But Williams quite rightly asks why that should make any different to your self-concern; he concludes with some puzzlement:

> Thus, to sum up, it looks as though there are two presentations of the imagined experiment and the choice associated with it, each of which carries conviction, and which lead to contrary conclusions. The idea, moreover, that the situation after the experiment is conceptually undecidable in the relevant respect seems not to assist, but rather to increase, the puzzlement; while the idea (so often appealed to in these matters) that it is conventionally decidable is even worse. Following from all that, I am not in the least clear which option it would be wise to take it one were presented with them before the experiment. I find that rather disturbing.[62]

What seems right about the torture thought experiment is that it does justice to the common-sense fact that we can (and do) demonstrate self-concern for our bodies, which we often take to be ourselves (as opposed to things we merely own). Williams rightly notes that depending on how one describes the experiment, i.e. the degree to which one describes to the subject a new body with a psychological profile that is closely related to the subject's pre-

---

[62] Williams, "The Self and the Future," pp. 61-2.

operative psychology, one can elicit different and contradictory impulses about self-concern. What seems unfortunate is that Williams later goes on to equate persons with bodies.[63] The result of the experiment, as demonstrated by the quote above, is a deadlock. What the torture thought experiments really show is how complex our self-concern can really be: we often have attitudes of self-concern that focus on our psychological continuity—indeed this seems to be the most prevalent case—but it is often the case that we feel self-concern towards bodily continuity, even in the absence of close psychological continuity. The evidence for this is the second trial of the torture thought experiment.

The result of this example of self-concern is that it becomes powerful evidence for adopting a physicalist approach to personal identity. The thinking goes something like this: even in the face of extreme psychological discontinuity, my self-concern tracks physical continuity. So we have good reason for adopting an animalist approach to personal identity. But we have now arrived at an extreme dilemma. Not only does this animalist approach violate our psychological intuitions about personal identity, it even violates the results of the first trial of the Williams thought experiment. One way to avoid this dilemma is by rejecting the assumption that an expression of self-concern in the Williams thought experiment is evidence of any theory of personal identity. We accomplish this by rejecting the assumed link between self-concern and personhood. That, in turn, is accomplished by eliminating the concept of the person and then re-analyzing the phenomenon of self-concern with the component concepts which remain after elimination. It is this analysis which follows.

Here is the general structure of the argument to follow: §4.1 recapitulates Parfit's argument about self-concern and his conclusion that Relation-R—not identity—matters most in survival. (He reasons that the question of identity in division is empty and what really matters is the lower-level facts of Relation-R, which *is* preserved in division.) The result of his argument is a suggestion that it would be rational for us to eschew normal self-

[63] His positive account does not rely so much on the thought experiment above, which is offered solely for the sake of showing that this style of thought experiment could not lend overwhelming support to either the psychological or the bodily criteria of personal identity.

concern in exchange for some kind of R-variant concern for our R-related descendants. We call this concern quasi-concern because it does not presume the identity relation between the subject and object of the concern.

Section 4.2 of the chapter borrows the *structure* of Parfit's argument in an effort to make an analogous point. By way of analogy, the debate between animalism and psychologism is similarly empty since neither offers a different outcome, just a different *description* of that outcome. Eliminating the concept of the person in favor of its lower level facts is plausible since it is the lower level facts which matter. These lower level facts are the components: physical and physical continuity, and agency. Just as it is rational to shift from self-concern to quasi-concern so too it is rational to switch from self-concern to component-concern. Component-concern does not presume that the object and subject of the concern are both the same *person*. All that it does presume is that the object of the concern is related by some component to the subject of the concern; he is either a physical continuant, a psychological continuant, or an agent continuant.

Section 4.3 offers the conceptual machinery needed to explain both of these phenomena (quasi-concern and component-concern) by borrowing the vocabulary of identificatory surrogates. Martin introduces this terminology to express the prudential concern we show for R-related descendants with whom we do not share the identity relation. I will make use of the terminology in an analogous way to express the prudential concern we show for our psychological, physical and agent continuants once the concept of the person has been eliminated.

Section 4.4 takes us back to the Williams thought experiment armed with this new conceptual machinery. With it we can offer a nuanced analysis of the seemingly contradictory instances of self-concern in the Williams thought experiment. In addition to the often recognized prudential concern for our psychological continuants, our account of component concern explains how it is possible to have self-concern for our physical continuants (which we saw in the second trial of the thought experiment). In a similar

fashion we can explain how it is possible to have self-concern for our agent continuants. This chapter will then explore how our values can affect which of these components we demonstrate the most prudential concern towards.

The final section of the chapter explains how all of this is evidence for not just the plausibility of the strategy of elimination but also evidence for its preference. We no longer need to take our concern for our bodies in the Williams thought experiment as evidence for the physical criterion of personal identity—which consequently violates our psychological intuitions about personhood. Plus we can recognize that we show prudential concern for our physical and agent continuers. This is all made possible by elimination of the concept of the person, resulting in fewer intuitions violated and a more satisfying explanation of self-concern that recognizes the multiple domains in which this phenomenon appears.

§4.1 PARFIT'S ARGUMENT ABOUT SELF-CONCERN

In addition to Parfit, Perry and Nozick are among those who accept some element of the claim that identity is not what matters most in survival. A second camp—Unger, Sosa, Lewis and others—claims that identity *is* what matters, to the exclusion of everything else. Let us recall the disagreement. The first camp is revolutionary and attempts to overthrow the common-sense position that identity is what matters. Their argument relies on thought experiments such as division, which we explored in the first chapter. These experiments run along the following lines: an operation manages to split your consciousness and produce 'two' (question-begging) individuals who bear the same psychological relation to you. On one popular version of this thought experiment, your brain is split and your left cerebral hemisphere is transplanted into a brainless body that is qualitatively similar to your old body. At the same time your right hemisphere is transplanted into another brainless body that also is qualitatively similar to your old body. Both 'Lefty' and 'Righty' bear the same psychological relation to you, both 'think' that they are you, both are committed to the same projects, both have, at least initially, the same beliefs, desires and attitudes as you before the

operation. Since the logic of numerical identity is no longer preserved in this situation, it would be erroneous to claim that you survive as both of these individuals, since transitivity would entail that the two individuals are the same as each other, which seems absurd. Not having a reason to choose one individual over the other (by virtue of symmetry), it seems that the best *description* is that you do not survive the procedure—in the strict sense of the term 'survive'.

Parfit argued in *Reasons and Persons* that the problems associated with division disappear if one adopts reductionism. Remember, the case of division poses a problem for personal identity because there are four possible outcomes to the case: (1) I survive as neither Lefty nor Righty (i.e. I do not survive the operation); (2) I survive as Lefty; (3) I survive as Righty; (4) I survive as both Lefty and Righty, which is usually cashed out to mean that I am one person with two bodies. If you believe in a further fact to personal identity, there must be a determinate answer to the question of which outcome obtains in the case of division. And the determinate answer is that I go wherever the further fact goes, although that may be extremely difficult, if not impossible, to figure out. If one does not believe in a further fact and adopts Parfitian reductionism, however, the choice between the four different outcomes evaporates. Parfit notes that

> On this view, the claims that I have discussed do not describe different possibilities, any of which might be true, and one of which must be true. These claims are merely different descriptions of the same outcome. We know what this outcome is. There will be two future people, each of whom will have the body of one of my brothers, and will be psychologically continuous with me, because he has half of my brain. Knowing this, we know everything. I may ask, 'But shall I be one of these two people, or the other, or neither?' ***But I should regard this as an empty question*** (my emphasis).[64]

Parfit explains his position with an analogy to social organizations such as political parties, clubs and nations. When *they* split, it may be an empty question whether the original organization continues to exist as one of its descendants, the other, both or neither. In this case, the question may very well be empty because these are not four different outcomes but four different *descriptions* of the same outcome. If the question can be empty for a social

---

[64] Parfit, *Reasons and Persons*, pp. 259-60.

organization such as a political party, why can't the question be empty in the same way for personal identity?

As we saw in chapter one, Parfit distinguishes two ways in which a question can be empty. The first is when a question has no answer but we could *give* the question an answer. However, such an answer would be arbitrary and we would have no basis for rationally deciding between one answer over an another. The second kind of emptiness, according to Parfit, is when the competing answers do not describe different outcomes, but rather are different descriptions of the same outcome. However, our decision about which answer to *give* to the question is not wholly arbitrary. Some descriptions may be better descriptions than others, and we have reasons to prefer one answer over another. Not all descriptions are created equal. But it is important to remember that even better and worse descriptions are still just *descriptions* of the same facts.

The case of division is an instance of Parfit's Argument from Below which I defended in chapter two. Personal identity just consists in certain other facts (physical and psychological continuity, etc.) and since personal identity just consists in certain other facts, it must be those other facts which are important and not personal identity itself. Therefore, there is nothing that really hangs on determining personal identity in problem cases such as division. (If it did we would equate double survival with no survival. This is clearly absurd, for while double survival may not be as good as normal survival, it certainly is much better than regular death.) Because we already know the underlying facts, and because whatever importance personal identity was thought to have actually stems from its underlying facts, then determining the extra question of personal identity is rather irrelevant. It is the facts which matter; these facts we already know.[65] In that case, choosing between the four descriptions is an empty question because we already know the results of division: two similar bodies, each with half of my brain, each psychologically continuous with me from

---

[65] Mark Johnston has argued against Parfit's claim and has advocated his own *Argument from Above*, which states that the underlying facts only gain their significance from the fact that together they constitute the higher fact—not the other way around. For a discussion of this objection see chapter two.

before the operation. From the Argument from Below we conclude that because personal identity just consists in these underlying facts, it is these facts which are important, not the empty question about identity.

If the above argument is correct, then it is rational to exhibit the same kind of prudential concern for both Lefty and Righty as one would in a 'normal' situation where numerical identity is preserved. It would be rational because everything that matters in normal survival is preserved in this case—the only difference being that there is more of the same. And in Parfit's phrase, double survival cannot be the same as no survival. Indeed, there might be advantages to double survival. Multiple yet mutually exclusive life projects can be pursued. Also, risky life plans can be entertained without sacrificing our values and commitments: Lefty can fight in the Spanish Civil War while Righty stays home and raises the children!

In order to express this prudential concern when the logic of identity is not preserved, we should use the term 'quasi-concern'; it is like normal self-concern except that it does not *presuppose* that the future individual to whom one demonstrates prudential concern bears the identity relation to you. Bona fide self-concern is therefore a limiting case of quasi-concern where the identity relation *does* hold. The only reason we think that identity is what matters in survival is because identity tends, in real life, to track the things which do matter, such as the relations that are preserved in quasi-concern. Ergo, identity is not what matters in survival. But here comes the rub: since identity does not matter in survival perhaps we ought to switch from self-concern to quasi-concern even in normal cases. As Parfit points out this has some practical results. We should exhibit prudential concern to anyone to whom we are suitably related even in the absence of identity.

As I said before, Unger, Sosa and Lewis (among others) claim that identity *does* matter in survival. They reject the conclusions from the division thought experiments. I will not rehearse their objections here but will consider in §4.3 whether any of them prove fatal to the argument I advance in §4.2. The objections are not relevant here because I do not

want to borrow the conclusion of the revolutionary camp that identity is not what matters. Rather, I want to borrow the conceptual machinery left over from their argument.

My aim in this chapter is to show that we can eliminate our concept of the person without doing violence to our understanding of self-concern. A careful analysis of the debate over whether *identity* matters in survival has provided the theoretical machinery needed to offer a new account of self-concern in the next section. (This theoretical machinery includes: [1] the idea that the question at hand is fundamentally empty, and [2] the Argument from Below which states that it is the lower level facts that matter.) It also has provided the necessary tools that will allow us to *express* how self-concern can track different components of the cluster concept of the person which may, under experimental conditions, diverge. (These tools include the use of a quasi-relation). Again, I will use the conceptual machinery introduced to make an argument about what matters in survival for my own purpose of showing that self-concern can attach to more than one component of the cluster concept of the person. As we will see, the structure of the two arguments will be analogous. And in the end this new analysis of self-concern will more closely track our intuitions on the subject.

There is an objection that needs to be considered here. Some have argued that there must always be a definitive answer to the question of identity—even in cases such as division. This cannot be an empty question. Chisholm has used this reasoning as evidence for a further fact view. During an exchange with Strawson, Chisholm gave his best effort at explaining why the question of our identity must always be determinate, saying:

> What I want to insist upon—*I concede I cannot give you an argument*—is that this will be the case even if all our normal criteria for personal identity should break down. Thus even if Lefty has half my fundamental personality traits and Righty the other half, even if Lefty has half the cells of my present nervous system and Righty the other half, even if Lefty thinks he remembers having done just the things I did on the even-numbered days and Righty thinks he remembers having done just the things I did on the odd-numbered days, and so even if there is no procedure or criterion whatever by means of which anyone could reasonably decide whether one or the other will be I, none the less, the questions "Will I be Lefty?" and "Will I be Righty?" do have definite answers. In the case of each of us, the answer will be simply "Yes" or it will be simply "No." And what I also see clearly and distinctly to be true is that, even under the perplexing conditions I have just described, there is no possibility whatever that the answer to both questions would be "Yes." *But, as I say, I cannot argue these*

> *points*. The most I can do is ask you sometime to really contemplate such questions [emphases added].[66]

It is difficult to evaluate Chisholm's point that we could never give up our belief that our identity is determinate. The evaluation is difficult partly because at least some of the argumentative power seems to come from Cartesian-style "clearly and distinctly" talk. But even if Chisholm is correct in his point, all he has shown is that we are unused to thinking in this way. This is no surprise since our concept of the person evolved in a world where division (for human beings) is not a regular possibility. One might be able to construct new concepts and a lexicon better suited to the challenge. In the following sections I hope to provide just such a lexicon so it will be possible to think and act in terms of component concern.

§4.2 AN ANALOGOUS ARGUMENT FOR COMPONENT CONCERN

Parfit has now argued that there are empty questions in problem cases other than Division. In *Reasons and Persons* Parfit supported a version of the psychological account of personal identity that he called the Wide Psychological Criterion. It was "wide" because the cause of non-branching psychological continuity did not need to have its "normal" cause for personal identity to be preserved. Parfit has since backed away from his support of the wide psychological criterion, arguing that it is a mistake to choose between the competing criteria of personal identity. These competing criteria include the different psychological criteria, Nagel's brain criterion, and the animalist criteria invoked by Williams, Thomson and Olson. Parfit now claims that thought experiments designed to force a decision between these criteria are an attempt to get an answer to a question which is essentially empty. In each problem case, all that remains is different descriptions of the same outcome, not different outcomes. If we know the underlying physical and psychological facts, then we already know all of the underlying facts that need to be known. As we see, then, Parfit has extended the

---

[66] Roderick Chisholm, "Identity Through Time" in *Language, Belief and Metaphysics*, edited by Howard Kiefer and Milton Munitz (Albany: State University of New York Press, 1970), pp. 188-9.

claim about emptiness from the case of division to the debate between the psychological and physical accounts of personal identity. Both are empty questions. So the claim about emptiness applies not just to the case of division and the argument about quasi-concern from the last section. It also applies to the Williams thought experiment.

Parfit also made use of the Argument from Below in making his argument about quasi-concern. He claimed that personal identity just consists in certain other facts and that it is these facts which are morally and rationally important—not personal identity itself. I hope to make clear that the Argument from Below applies just as much to the Williams thought experiment as it does to the division thought experiment. Consequently self-concern is not a function of personal identity but is rather a function of the certain other facts which constitute "personal identity" so-called. These "other facts" are the component concepts that I have identified in previous chapters. And once we stop talking about persons and start talking about their components we will get a more nuanced understanding of self-concern. Furthermore, it is this account of component-concern which nicely explains the evidence of bodily concern exposed by Williams in the torture thought experiment.

The conflicting intuitions exposed by the Williams thought experiment were only a problem when we were trying to determine what it means to be a person. This investigation was motivated, in part, by the assumption that something hinges on finding out what it means to be a person. The whole field of personal identity assumes this question to be important, otherwise we would be investigating identity criteria for something else—not persons. The fact that we are interested in *personal* identity shows that we take it to be important to figure out what it means to be a person. We should abandon this assumption and with it the goal of determining what it means to be a person. As suggested in previous chapters, our concept of the person is a cluster concept composed of distinct (and sometimes competing) sub-concepts. So let us also abandon—for the moment—the concept of the person and replace it with the component concepts that we are dealing with here: bodily continuity and psychological continuity. We ought to be broadly Quinean here and

recognize that our concepts have little meaning beyond which our current needs have invested them with. Our concept of the person is only as precise as our current needs require, i.e. where psychological and bodily continuity coincide. Under experimental situations our current concept does not work as well. To ask, under these extreme circumstances, about our concept of the person is not to mine our intuitions for new information; rather, it is to encourage the subject of the thought experiment to legislate greater precision in the concept than previously existed.

If we recognize our concept of the person as a cluster concept and eliminate it in favor of its components, the problem exposed by the torture thought experiment virtually disappears. It no longer becomes inconsistent that we demonstrate self-concern for psychological continuity in one version of the thought experiment and self-concern for bodily continuity in a second version of the thought experiment. With our use of the concept of the person temporarily suspended, might it just be the case that we exhibit self-concern for both of these things? In that case, the final question in Locke and Shoemaker's thought experiment ("Which one is me?") is an empty one. The possibilities do not represent different outcomes but represent different descriptions of the same outcome. Finding out which person is me is an empty question, a fact that is more obvious once we recognize 'person' as the cluster-concept that it is and replace it with its component concepts. "Where will I be?" is a question that can be answered with the following question: "Why do I have to choose?" A simple description is no longer appropriate given the extreme nature of the circumstances. A more complex description can be fashioned out of the remaining component concepts. We can instead describe a physical continuant and a psychological continuant.

The emptiness claim is consequently extended to the debate between the animal and psychological criteria of personal identity, the debate between neo-Lockeans on one hand (Parfit, Perry, Shoemaker, etc....) and neo-Aristotelians on the other (Williams, Thomson, Olson, etc....). We ought not choose between the criteria because they do not represent

different outcomes of the operation but rather different descriptions. So the question in the torture thought experiment: "Is the person with my body, despite changes in psychology, me?" is an empty question. The important point is that one's prudential concern for that future individual is rational, though not in the same way as in the case of Division. In the case of Division, the quasi-concern is rational because everything that exists in normal survival (and hence regular self-concern) exists in Division, except that there is more of it. This is not the case in the torture thought experiment. Not everything that exists in normal survival exists here, in particular there are massive psychological changes. So a non-trivial difference is exposed. However, at least *some* of what matters in normal survival is preserved, and what is preserved is *significant* because the results of the experiment show that we fear the procedure. We fear the procedure because the body is a locus of self-concern, independent of the psychological facts of the case, and independent of the empty question of whether that future individual will be me.

What becomes of self-concern given the proposal to eliminate our concept of the person? The assumption has always been that our concept of the person was indispensable, i.e. that it is indispensable because of our metaphysical or ethical commitments. It has traditionally been assumed that we have a special concern for ourselves as *persons*, and that this special category is only to be understood as rational in light of our being persons. Can we still explain self-concern, explain its power, its pervasiveness, its reasonableness, without making reference to persons? Not only can we explain the phenomenon of self-concern without making reference to persons, we can do so with greater nuance than before. Self-concern is one domain where the advantages of eliminativism are clear.

Return to the torture thought experiment. In that case we are given a choice between showing self-concern for two future persons who stand in different relationships to you: one who is bodily continuous and psychologically discontinuous and another who is bodily discontinuous but psychologically continuous. The subject of the thought experiment is forced, by invention of the experiment, to *choose* between these two future 'persons'. But it is

important to remember that the extremity of this forced choice is a function of the design of the experiment, and the legitimacy of forcing this hard choice on the subject is warranted solely because it was assumed that self-concern is a phenomenon that attaches to persons. With that assumption pushed to the side—and with it the goal of determining *personal* identity—we are free to answer with common sense: we exhibit self-concern for our psychological continuants *and* we exhibit self-concern for our physical continuants.

This terminology is important because it allows us to express how you can be *prudentially* concerned for a future individual independent of whether that future individual is you. In the case of division, numerical identity is called into question because of the presence of a competing candidate for numerical identity, hence the need for quasi-concern. Within the context of my argument about eliminativism and self-concern, the competing candidate is slightly different. In the torture thought experiment you are envisioning someone to whom you are bodily continuous but psychologically discontinuous. How can you express the kind of self-concern that was found in Williams' thought experiment without begging the question of whether that future person is you? You can adopt the use of component-concern, which does not presuppose that the future person to whom one demonstrates prudential concern is you. It does not presuppose either the psychological or animalist criterion of personal identity.

It should now be clear that we have borrowed the structure of the argument in §4.1 to make an analogous argument. In division the question of personal identity is empty. The underlying facts (i.e. psychological continuity) are preserved in division, and because it is these facts from which identity derives what was alleged to be its importance, then the underlying facts are important even in the absence of identity. Hence the conclusion that it is rational to show some R-variant of self-concern (quasi-concern) for one's post-fission descendants. Now extend the structure of the argument to our conflicting intuitions in the Williams thought experiment. The different answers given by the competing criteria (physical and psychological) are just different descriptions of the same outcome, not

different outcomes, so self-concern must get its importance not from personal identity but from its underlying facts: in this case the physical and psychological facts. The Argument from Below applies in this case as well. Because personal identity just consists in other facts, any rational or moral importance that we thought derived from personal identity actually derives from below, i.e. from the facts that compose it, from its *components*. So personal identity is not what matters for self-concern. What matters are its underlying facts—its *components*.

This presentation shows that anyone sympathetic to both the claim about emptiness and the Argument from Below ought to accept that self-concern derives not from personal identity but from its components. We should therefore stop talking about self-concern and start talking about component-concern, because if it is the underlying facts of the components that really matter, then concern will derive from *them*, not personal identity. Component-concern is the prudential concern that one demonstrates for the components of your existence: psychological continuity, bodily existence, and agency. It is rational to be selfishly committed to the continued existence of these components and one need not answer the question of survival in a particular component combination in order for one to demonstrate rational concern. For example, it is rational for you to selfishly exhibit component-concern for a future individual who is bodily continuous with you but has undergone psychological disruptions. And the further question: "Is this person me?" is an empty one. We know all of the facts of the situation because they are included in the description.

Similarly, if your body is destroyed but a new body—qualitatively *different* from your old body—is reprogrammed with your old psychology, it would be rational to demonstrate component-concern for the new mind-body combination. And one need not determine the identity of the new mind-body combination. One might call this combination me, or one might call this combination a new person. But these would not be different conclusions. These would merely be different descriptions of the same outcome.

Now here's the rub: in our ordinary lives we exhibit normal self-concern. We naturally assume that this self-concern attaches to the concept of the person because we take ourselves to be persons. However, I have claimed that 'person' is a cluster concept and we ought to consider eliminating the concept in favor of its components. It is the components which are important, and everything that we want to say about persons we can say better with the sub-concepts. And so it is with self-concern. In everyday existence when we demonstrate garden-variety self-concern for our future selves, we are not required to choose between the components. The difference here—as opposed to the thought experiment—is that all of the components are present. In everyday life we need not choose between the body as a locus of concern and psychological continuity as a locus of concern because the two usually go together.

Parfit and most other philosophers sympathetic to neo-Lockean psychological accounts have argued that psychological continuity (or in Parfit's case the R-Relation) is the component that matters for self-concern. It is precisely this relation which is preserved in the case of Division and consequently it is rational to be prudentially concerned about one's post-fission descendants. While this is certainly true, I think it is wrong to proceed confidently on the basis of this evidence to the conclusion that self-concern always attaches to psychological continuity.

Consider, again, the results of Williams' torture thought experiment. In a situation where one will be forced to endure severe psychological disruptions, followed by torture, it is not irrational to display self-concern for this future individual. Before we accepted the claim about emptiness, this data was unfortunately taken as suggestive evidence *against* the psychological criteria of personal identity and as evidence *for* the physical criteria. This was certainly a mistake. But after accepting the claim about emptiness, we are free to properly interpret this piece of data. Although component-concern almost always attaches to the component of psychological continuity, there are cases where it manifests itself in the component of physical continuity as well. Although component-concern is much rarer in

the case of physical continuity than it is in the case of psychological continuity, nothing prevents us from realizing that Williams' thought experiment demonstrates that the body *can be at least theoretically* a locus of self-concern. And this is just the logical extension of accepting both the Argument from Below and the claim about emptiness. If both of these claims are true, it makes sense to start talking about component-concern and stop talking about self-concern. More importantly, there are no strange ontological costs involved in recognizing atypical cases of component-concern (such as in the torture) case, because once we have accepted the emptiness claim and the Argument from Below we no longer need to admit these atypical cases as evidence for a particular criteria of personal identity.

Under what conditions do we demonstrate self-concern for bodily continuity? In the torture case, we clearly exhibit self-concern for the future person who is bodily continuous with us, and perhaps it is rational to do so even in the absence of any of the logical machinery of personhood. Regardless of whether this future body is the same person as me or not (and to me the question here appears empty), the sight of our body being tortured, *regardless of its psychology,* would be distressing to us in the extreme. As biological entities we are embodied. Not only do we 'identify' others with bodily criteria, our own existence is frequently focused around our physical nature. Furthermore, that concern for our physical nature does more than merely 'piggy-back' unto our concern for our psychological continuity. Our concern for our bodily continuity is not merely a consequence of the contingent fact that our psychology must be housed in some physical realizer. Even in the absence of this emergent psychology, we identify ourselves as biological animals and often exhibit self-concern for ourselves in this way. How else to explain the very real anxiety that many suffer when contemplating organ donation, autopsies or chronic vegetative states? It would be premature to chalk up these anxieties to irrational or philosophically naive concerns. Rather, these anxieties stem from the fact that we consider ourselves psychological beings who are embodied and who exercise agency, and we show self-concern in all of these domains because we take ourselves to be all of these things. In the absence of an umbrella

concept of the person, we are at liberty to admit that it is at least conceptually possible that we show self-concern in *all* of these domains. (This will be explored in greater detail in §4.4.)

To suggest that we show self-concern in all of these domains is not to suggest that our self-concern is equal in each domain. We may very well exhibit self-concern in greater and lesser degrees in each of these domains. I have already identified self-concern in the biological domain because it is, in part, surprising. More obvious is the concern we adopt for our psychological continuants. Our character traits, memories, beliefs, desires, our phenomenological unity are all important to us—indeed they are usually more important to us than the features of our biological existence. It seems clear that while our self-concern tracks *both* psychological and bodily continuity, it is by no means correct to suggest that they track both equally. Indeed, if forced to prioritize, it might be appropriate to rank our psychological self-concern over our bodily self-concern, although as Williams suggests, there would be an element of risk in this. The risk stems from the fact that we have no previous experience from which to draw a rational conclusion. I will also explore in §4.4 the degree to which such rankings might be influenced by our values.

We should now consider the corollary of the Chisholm objection updated for the new argument about component concern. This corollary objection claims that I must *believe* that there is a definite answer to the question of which continuant in the Williams thought experiment is me. Put another way, it was assumed that in order to make sense of our prudential concern for an individual *that we needed to conclude that this person was me.* But nothing is further from the truth. We do not need to think of that future person as me in order to demonstrate prudential concern here. *A forced choose between continuants in a thought experiment only makes sense if you have accepted, antecedently, the concept of the person.* We have suggested instead that the question of whether a future person is me is an empty one and has no rational or moral importance.

The fact that we demonstrate prudential concern in the Williams case tells us nothing about personal identity but ends up telling us a lot about self-concern: it is a much more complex relation than we once thought when we believed personal identity to be both determinate and important. Since it is neither, self-concern need not manifest itself in *just one* component. Perhaps it is *rational* for us to demonstrate prudential concern for a future individual because our prudential concern extends beyond the one component of psychological continuity. We have prudential concerns that, surprisingly, extend beyond this one component.

§4.3 IDENTIFICATORY SURROGATES

In the last section I laid the groundwork for an account of self-concern that makes no reference to persons. Instead of feeling prudential concern for a future person to whom one is identical, it is rational to exhibit prudential concern for one's continuers: psychological, physical and agent. Each of these continuers tracks a component of the cluster concept. I defended this account by appealing to an argument structurally similar to the Parfitian argument for quasi-concern. In my case, the argument states that the debate between psychologism and animalism exposed in the Williams thought experiment is fundamentally empty and that the Argument from Below applies in this case because the importance of personhood derives from its lower level facts, its components. So we need new conceptual machinery to help us talk about self-concern in the absence of persons and in the presence of components such as psychological and physical continuity as well as agency.

Recent work in the literature on self-concern lends support to this project. For example, Martin argues that we have "identificatory surrogates." He says that we "identify ourselves in the *future* primarily by anticipating *having* experiences and *performing* actions. Ordinarily we remember or anticipate having only *our own* experiences and performing only *our own* actions. That's because ordinarily our options include only those we have in real

life."[67] But in hypothetical situations we have identificatory surrogates to whom it is rational to show prudential concern. These surrogates are individuals who are *continuers*, i.e. they continue our lives by falling in certain psychological relation with us. Although continuers need not be numerically identical to us, they nonetheless continue our psychological lives. And, according to Martin, we can imagine our continuers having experiences and performing actions in the future in the very same way that we imagine ourselves having experiences and ourselves performing actions. The rub, according to Martin, is that even normal self-concern may be 'continuistic' because we are concerned about our future selves insofar as they 'continue' our lives, but not because they stand in a relation of numerical identity. He writes that surrogate self-identification

> requires, at the minimum, that a person relate psychologically to the other's experiences and actions almost as if they were the identifier's own. In the case of future experiences, say, it requires that you would adopt pretty much the same constellation of attitudes and behavioral responses toward the person who will have the future experiences, and toward the experiences themselves, that ordinarily you would adopt only toward yourself in the future, and only toward your own experiences.[68]

This is the precisely the kind of relation that might hold between a pre-fission individual and her post-fission continuers. Similarly, this is also the kind of relation that would hold between an individual and her future physical or psychological continuer. This last example demonstrates how important it is to realize that continuer relations need not be exclusively psychological. The torture thought experiment shows that we can imagine a *physical* continuer having experiences—indeed having quite painful ones—and express self-concern for those reasons. And the structure of Martin's identificatory surrogates shows that we need not think of that future person 'as me' in order for that concern to be prudential.

Consider also how Adams defines the self-interest relation. He writes that

> What we are attached to in ourselves, in a reasonable self-concern, is not just our bare metaphysical identity, but also projects, friendships, and at least some of the most important features of our personal history and character... If a possible life contains so little of the concrete content that I care about in my actual life that it should not matter to me that it

---

[67] Raymond Martin, *Self-concern: An experiential approach to what matters in survival* (New York: Cambridge University Press, 1998), p. 93.

[68] Martin, pp. 98-9.

> could, metaphysically, have been mine, let us say that it bears no *self-interest relation* to my actual life.[69]

Although Adams is expressing his point in terms of possible worlds, the point here is pretty much the same. If a future individual, say a physical or psychological continuer described in the Williams thought experiment, has enough of what Adams calls the "concrete content that I care about in my actual life," then it makes sense to bear some kind of self-interest relation to that individual.

Furthermore, since the continued existence of one's body may be part of the 'concrete content' that one values in one's actual life, then it makes sense to exhibit self-concern for a future individual who retains that concrete content, even if it does not include total psychological continuity. This is the essence of the component-concern that I advocate. Because we exhibit component-concern for the continued existence of our bodies in our actual lives, it should not be surprising that we exhibit the same component-concern for our bodies even in extreme experimental situations that involve massive psychological disruptions.

My point in drawing these parallels with identificatory surrogates has been twofold. The first has been to show that my new notion of component-concern is not so radical if you have already adopted something like quasi-concern. It is an easy pill to swallow, I think, especially if one already accepts these recent revolutionary proposals to widen or amend the self-interest relation to future individuals to whom one is not numerically identical. Indeed, the arguments leading to these two conclusions offer similar structures and share common concerns. The second motivation for borrowing Martin's notion of identificatory surrogates is that we need some way of expressing our relationship to our physical, psychological and agent continuers that captures the idea that they somehow 'continue' some significant aspect of our lives without implying that the future continuer is the same *person* as us.

---

[69] Robert M. Adams, "Existence, Self-Interest, and the Problem of Evil," *Nous* 13 (1989): 60.

Two points of caution about my use of Martin's logical machinery. The first is that I am using his notion of identificatory surrogates for a completely different purpose from the one for which he introduced the term. Martin's project is to model our prudential concern for a continuer with whom we are psychologically continuous, regardless of whether the identity relation holds. Paradigmatic cases include branching thought experiments. It is important to note, however, that Martin's project is restricted to an explicitly *psychological* arena, mostly because he conceives of self-concern as an experiential element of our psychology. In contrast, I have broadened the term's use to include our prudential concern for *physical* continuers as well, in order to provide my own model for our prudential concern without the concept of the person. This is an extension that I suspect Martin would neither approve of nor endorse.

A second note of caution: Martin introduces identificatory surrogates to cover cases of prudential concern when the identity relation does not hold. Similarly, Parfit's notion of showing concern for R-related individuals is meant to cover the same ground: cases of prudential concern in the absence of identity. (In Parfit's case the structure of the argument is that while identity is not preserved, everything that is important is preserved, so identity can't be all that important.) But now a crucial difference is exposed between the two cases of identificatory surrogacy. Unlike Martin, I want to make use of the terminology even though I don't want to concede the point that a continuer (say, for example, a physical continuer) is not identical to you. But I also do not want to claim the reverse either, i.e. that the identity relation definitively holds between you and the continuer. Rather, it is important to remember that the question of whether *personal* identity is preserved in your continuer (which ever kind of continuer it is) is an empty question.

The rationale behind this position is important to remember. The question is empty not because of anything having to do with the concept of *identity* per se. (This is the case with division, where the reason identity isn't preserved is because the duplication means that the logical structure of identity, the one-to-one relation, is no longer preserved). In my case

the question is empty because of the concept of the *person*. Because personhood is a vague cluster concept composed of distinct components, any question which pits these components against each other will be an empty question. The cluster concept isn't precise enough to be extended to extreme cases where one component is present but another is absent. So in the case of the Williams thought experiment, it is neither correct nor incorrect to say that the subject's identity—qua personhood—is preserved in the procedure. And our use of the language of identificatory surrogacy should not suggest that somehow this identity *qua* personhood is either preserved or violated. Rather, the question is empty because of the limitations of the concept of the person. And this is a departure from both Martin and Parfit on this subject. What we can say in a situation such as Williams' thought experiment is that a *component* is preserved and that a patient is rationally justified in showing prudential concern for a physical or psychological continuer. The patient is justified because the continuer is an identificatory surrogate. This neither affirms nor denies that the surrogate is the same 'person' as you. What I hope to have shown here is that this question is somewhat beside the point and that an answer to it is not necessary for a coherent picture of self-concern.

At this point it is important to evaluate some potential objections. There are various objections to the argument advanced in §4.1 and we ought to see if any of them prove fatal to my argument about component-concern presented in §4.2 In previous chapters I defended Parfit's Argument from Below from an objection advanced by Mark Johnston and I shall not repeat the defense here. However, there are objections against the argument in §4.1 For example, Unger uses a series of thought experiments about pain to conclude that normal survival is always preferable. "*Any* case that lacks strict survival will be worse than *every* case in which the person himself does survive," he writes.[70] Unfortunately, most of Unger's cases involve comparing the anticipation of pain in a case with strict survival to the anticipation of a comparable pain without strict survival. But this is consistent with strict

[70] Peter Unger, *Identity, Consciousness and Value* (New York: Oxford University Press, 1990), p. 211.

survival merely being a tie-breaking consideration. As Martin quite rightly points out, what would be most instructive would be our attitudes about different levels of pain. In other words, would we choose strict survival with lots of pain over continuistic survival with a much smaller pain? The answer is no longer so obvious.

Sosa objects to the argument in §4.1 because he thinks that it leads to a slippery slope towards a view where continuation matters over survival even in cases without causal relatedness. Martin nicely replies that this burden falls on either side. A defense of why only identity matters in survival would be vulnerable to the same slippery slope, and as such, the problem cuts both ways. Finally, Lewis has a well-known solution to the case of division: a four-dimensional view of persons, but it seems possible to envision thought experiments that put even the four-dimensional view under enough pressure to call into question whether identity is what matters in survival.

In any case none of these objections are fatal to the argument in §4.2, since none of them *directly* go after the Argument from Below (with the exception of Johnston) or the *structure* of the argument in §4.1. Rather, they provide additional reasons why the conclusion of §4.1 should be rejected. If any of these objections are successful then the §4.1 conclusion is in peril. But none of them directly refute the conclusion offered in §4.2 about component concern. None of the objections touch §4.2 because you can't help yourself to the idea that any of the continuants in a Williams-style thought experiment are not you. The question is simply empty—so any objection appealing to the necessity of identity (which all of them do) in survival will be irrelevant to the argument at hand. I have not denied that *personal* identity is preserved in a Williams-style case. I have claimed that the question is *empty* because personhood is a cluster concept that lacks the precision needed to answer the question adequately. We ought to replace personhood with its components—psychological continuity, physical continuity and agency—which we can easily track and for which identity questions are a lot easier to answer.

§4.4 RETURN TO THE WILLIAMS THOUGHT EXPERIMENT

4.4.1. This account explains why we have self-concern for the body

Consider the example of something I might value. Previously I introduced the relation of fatherhood. I might be committed to being a father to my son. The concept of fatherhood is, *pace* adoption, a biological notion. I am not only my son's father in the sense that I fulfill a certain social role in the household (which is preserved in adoption), but also I am my son's father in the biological sense that he is my offspring. In some situations the latter may be important and may be something I value independent of the former. I value the fact that we have common ancestors and consequently a common cultural origin, that we share at least some of the same genetic predispositions, that we look sufficiently alike that when my son looks at me he sees a picture of what he might become, and conversely when I look at my son I see a portrait of what I once was. These concerns need not be dismissed as metaphysically naive nostalgia for the body. There is something crucial here and it stems from the fact that we are, among other things, biological animals. For someone who values fatherhood in this way, it is rational to have self-concern for biological continuity independent of its status as a physical realizer for psychological continuity. The continuation of the body is not just an added "extra". It is, instead, absolutely essential to the project of being a father. If one's psychology were transferred to another body and your old body were destroyed, one would no longer stand in the father relation to one's child. In a very non-trivial sense you would cease to be a father.

You might also value the life of the mind. You might value the contemplation of esoteric mathematical theorems, obscure theological questions or the latest advances in particle physics. In that case, the continued existence of one's psychological continuity would matter very much—possibly more than the continued existence of one's body. In fact, one might not care at all which body realized your psychological continuity. Moving to a new body would not stop you from manually calculating • to yet another decimal point. One

might not even care if one were transplanted into a completely paralyzed body, just as long as there was a functioning brain (or a comparable physical realizer) and some capacity for communication. This would be an identificatory surrogate. In this case one would demonstrate extreme self-concern for psychological continuity and less concern for the continued existence of a *particular* body. In everyday life, of course, we are not forced to so radically prioritize the domains of our self-concern; bodily continuity, psychological continuity and agency usually go together. It is only when they do not that the different varieties of component concern become apparent. This becomes obvious when people are asked about their wishes if they become brain dead but their bodies still function. Some people do not hesitate to sign a card for organ donation, essentially authorizing the proactive ending of their biological lives. Others make arrangements to have their hair curled and fingernails manicured.[71]

I suggested that the move from self-concern to component-concern is not as radical a move as it might have first seemed. Indeed, I have tried to show that it is the logical extension of two claims advanced by Parfit which, although certainly revisionary, have attracted a wide degree of support in the philosophical literature on personal identity. However, some might find my point moot, because while they accept that the body could theoretically be a locus of self-concern (because of the move from self-concern to component concern), they do not consider it a practical possibility. Although it is not necessary, in the deep metaphysical case, that self-concern will always go with psychological continuity, they see it as being always contingently true. So let me offer an example that might illuminate the significance of this new account of prudential concern.

Earlier we discussed an individual who greatly values being a father to his son. Indeed, this relation is a biological relationship, because the individual would cease to be the

---

[71] The bio-ethical rationale behind organ donation rests upon the medical communities notion of "brain death." The idea here is that the organ donor is dead—even though his body continues to function—because his brain is no longer supporting high cognitive functions. It has been noted recently that this concept of "brain death" was in fact introduced into the medical literature to facilitate organ donation.

boy's father if the father's consciousness was transplanted into another body. In this case the biological relation of fatherhood is predicated on the existence of a *particular* body. Expressing component-concern for the continued existence of one's body, even in the absence of psychological continuity, is merely a case of valuing one's existence as your son's father and expressing prudential concern for an identificatory surrogate. This does not seem absurd—especially once we remember that questions about *personal* identity arising form the Williams thought experiment are empty.

In *Reasons and Persons* Parfit argued that the existence of a new body would not necessarily affect psychological continuity. Although conceding that a change in body limits one's options "and might thus indirectly lead to changes in my character," he sees this as insignificant to the fate of psychological continuity. He approvingly cites Quinton's argument that

> It would be odd for a six-year old girl to display the character of Winston Churchill, odd indeed to the point of outrageousness, but it is not utterly inconceivable. At first, no doubt, the girl's display of dogged endurance, a world-historical comprehensiveness of outlook, and so forth, would strike one as distasteful and pretentious in so young a child. But if she kept it up the impression would wear off.[72]

Parfit's argument here is not universally accepted. Indeed many, including Olson and Wilkes, have questioned the degree to which psychological continuity can be preserved in a body transfer. But even assuming that Parfit's argument is convincing, it would be wrong to take it as evidence that the body cannot be a locus of self-concern or that bodies are irrelevant.

Consider another case, this one of my own: I am married to a fantastic woman whom I dearly love. She needs me in her life. I have no job and no children and no hobbies. My sole reason to live is to be my wife's husband, a function and vocation I consider noble. Also, one of the ways I function as my wife's husband is by loving her sexually. It is the avenue through which I can love her and take care of her. This sexual relationship is the most important thing in my life. A non-trivial aspect of this sexual relationship involves the

---

[72] Parfit, p. 254.

potential creation of a baby. But the relationship is not exclusively about procreation. My wife finds my body comforting in much the same way that a baby finds a parent's body comforting.

In this thought experiment I am forced to undergo an operation. The result of the operation will be two individuals. One individual will be psychologically continuous with me from before the operation but will have a new body. This body will be the body of an old woman. The second individual will have my original body but will be forced to endure some—though not total—psychological disruptions. He will therefore not be perfectly psychologically continuous with me before the operation. He will seem like a slightly different person—in the non-philosophical sense of the expression—when he returns home.[73]

Would it be rational to exhibit component-concern for one's physical continuer, to choose him as your identificatory surrogate? I find it hard to argue that this would be wholly irrational, that it would be irrational for me—on prudential grounds—to identify with this physical continuer over the psychological continuer, that it would irrational for me to want this continuer to survive so that he could return to my wife. Now remember, the question whether this 'person' is me is an empty question because we already know the facts. And the facts are: this continuer has more of the components that I consider important. It matters to me that I have a male body and that I have a *particular* male body, because it is this body that my wife is used to. On prudential grounds I might identify with this continuer because it contains more of the elements which I consider important. I concede, however, that if the psychological disruptions were *major*, a patient might be more likely to choose the psychological continuer as an identificatory surrogate. But even this is not necessarily the case, as we saw in the second trial of the Williams thought experiment.

---

[73] Notice that, unlike Unger, I have compared scenario's with different psychological disruptions. This is where our responses will be interesting.

An objector might claim that it would make no difference if my body were replaced with a replicated body, say by teletransportation. This shows that self-concern for a particular body is not rational. But at this point, I can concede the point without hurting my argument. Perhaps it would not matter if a body were replaced with an exact replica. However, this replicated body would be related to my original body in the following way: qualitative similarity resulting from an identifiable causal connection. But this does nothing to show that it is impossible to identify the body as a locus of self-concern. Although a replicated body might be a *different* body, it is causally related to one's original body—it has the same design.

In my case of the devoted husband, I helped myself to the fact that the post-operative individual with my body was not completely psychologically discontinuous with me. I think I was warranted in helping myself to this arrangement. One might object: the results would be different if it were a straight case of switching bodies, where the consciousness of one was implanted in the consciousness of the other. But, as the torture thought experiment in Williams demonstrates, why should it matter how we describe what is happening in the other 'person'? Why should it matter if we point out that the psychological disruptions in my body coincide with the emergence of psychological similarities in another body? As Williams pointed out, this shouldn't affect the results in the torture thought experiment and it shouldn't affect the results in my experiment. Although surprising, and although rare, it is possible for the body to be a locus of self-concern.

At this point an objection can be raised against our account of component-concern: it is rational to demonstrate prudential concern for our bodies only when they are ours, i.e. when we own them. It isn't rational to care what happens to a body that is no longer yours. What does it matter if a body that was once the physical realizer of your psychology—but which has been replaced by a new physical realizer—is dismembered or tortured? Because the answer is that it does *not* matter, the objector concludes that psychological continuity is the hallmark of personhood and that we are not bodies but *have* or *own* bodies. The reply to

this objection is that it assumes precisely what it aims to prove, i.e. what allows one to say that the body that is being tortured is not *yours*? This is precisely what is at issue and what the torture thought experiment calls into question. Under one description it may be plausible to answer that the tortured body *is* you, in which case prudential concern would be appropriate. But the point to emphasize is that there isn't a real disagreement about reality here. We know all of the facts from the thought experiment. The further question: "Which one will be me?" is an empty question that seeks more precision from our concept of the person than is possible. So we should end our talk about persons. And we could then admit that our prudential concern applies in multiple domains that in our normal lives happen to coincide. We demonstrate prudential concern for our psychological continuers more often than other continuers, though it is neither unheard of—nor irrational—for a physical continuer to be an identificatory surrogate.

The objector helps himself to the conclusion that my body *some day belongs to someone else*. One is no longer allowed to help oneself to this piece of information if one has admitted that the debate between the physical and psychological criteria is empty. It is not determinately true in a deep sense that this future person is *not* you. Furthermore, this future person has one of your components, and one might be justified in expressing prudential concern for the continuer of this component.

4.4.2. This also explains how we can have self-concern for our agency-continuer

In the previous chapter we concluded that the hallmark of agency was the information-pooling and deliberation that precedes intentional action. This view supports Rovane's claim that links agency with a common rational point of view—in contrast to the more traditional phenomenological point of view usually associated with agency. So what is required for agency is the pooling of information, deliberation on the basis of that information, and the implementation of a project, plan or response based on that deliberation. Given this definition, an agent need not coincide with biological human

beings. Theoretically, groups of human beings could coalesce into a collective agent, just so long as the group pooled information and deliberated collectively. There also could be multiple agents inside a single human being, although this is an even more remote empirical possibility. In the case of the collective agent, we need not assume that the individual agents who compose the group agent fall out of existence when the collectivity emerges. They might, for example, segment themselves in such a way as to form overlapping agents.[74] Several agents might form a collective agent amongst themselves for professional reasons, yet might segregate their personal lives from the group agent. The biological human body would therefore include overlapping agents: one agent dealing with personal affairs and a piece of the larger group agent dedicated to some corporate endeavor.

It is precisely in this kind of complex metaphysical situation that an eliminativist account of self-concern could help. First of all, given my eliminativist position, we are rescued from claiming that the group agent is a group person—a violation of common-sense that I discussed in the previous chapter. More importantly, we can give a straight forward explanation of the prudential concern in this situation. Suppose the collective agent in question is a military unit with a common objective. Perhaps they are rebels fighting for a common political cause. The members of the unit are deeply committed to achieving this political goal through their military efforts. In fact, each member of the military task force is willing to die in order to protect the other members of the force and in order to further its military mission, and consequently, its political goals. Indeed, one of the members is called upon to perform a risky maneuver which will, invariably, lead to his biological death. As a consequence of his commitment, though, he completes the maneuver without hesitation. He is killed in the effort.

One could analyze this scenario as a case of altruism. This has been a traditional strategy, although it is not convincing in every case. But we now have the conceptual machinery to analyze the scenario as a case of prudential concern. And here's how that

---

[74] Rovane discusses overlapping agents in chapter 5 of *The Bounds of Agency*.

works: The individual has freely joined the collective agent. In turn, the collective agent has calculated how to employ its resources in order to achieve its objectives—including what costs it is willing to bear in pursuit of that objective. If the collective agent is then faced with the possibility of, say, outright destruction, the individual might exhibit prudential concern for the continued existence of the agent (of which it is a part) by sacrificing its biological life to preserve the agent. Here is a case of prudential concern for one's agent continuer, where the concern does not track the continuity of a particular body or a psychological continuer, but rather the continued existence of a particular agent. The agent is the identificatory surrogate, not the biological or psychological continuer. This shows how this kind of sacrifice, in the face of an important objective, can be *rationally* chosen (a potential problem when analyzing altruism). And this analysis was made possible by dropping our talk of persons and shifting our attention to its components.

In real life agents come one to a body so our prudential concern for our physical, psychological and agency continuers will coincide. All three are mutually reinforcing. In normal situations, demonstrating prudential concern for one is demonstrating prudential concern for the other, because only with the continued existence of your body do you continue to exist as an agent. It is easy to change this synchronicity by bringing in the case of group agents. However, nothing in my argument requires the acceptance of collective agents. The idea is metaphysically controversial. The conclusion one ought to draw from this demonstration is that *if* one believes in the *conceptual* possibility of collective agents, this eliminativist account of self-concern can be useful in explaining the rationality of the decisions involved.

To review our argument: if we eliminate our concept of the person, not only do we not violate our common-sense notion of self-concern, but the field of possibilities opens up nicely. What if 'person' is not a natural kind term but rather a cluster concept composed of distinct sub-concepts? Then perhaps the concept of self-concern attaches not to persons but to its components. The results of the torture thought experiment are no longer an apparent

paradox in need of resolution. Rather, the results are simply a consequence of the fact that self-concern is multifaceted, although we don't appreciate this in everyday life. Because our concept of the person was designed for use in everyday life, it is usually adequate to the task. However, putting pressure on the concept through extreme thought experiments demonstrates that it will crack under this pressure. We should not assume that it has more precision than our needs have invested it with. So we should eliminate our concept of the person in favor of its components and recognize that self-concern attaches to those components. In the process we get a more plausible description of what happens in Williams' thought experiment. Not only can we identify psychological continuity as a locus of self-concern (this is rather unsurprising), but we can also identify physical continuity as a locus of self-concern as well. Furthermore, it is at least conceptually possible to identify agent identity as a locus of self-concern, although I make no *empirical* claims as to the actual existence of this phenomenon.

4.4.3. Which components we care about depends on our values

In the torture thought experiment, we are forced to choose between two future persons, one who is related to us with more physical continuity and less psychological continuity, and another with no physical continuity but with more psychological continuity. Suspending our talk about persons for the purpose of our discussion about self-concern allows us to explore the fact that our choice of identificatory surrogates might be informed—at least in part—by what we value. In that sense, if our embodied nature is a fundamental part of an account of human flourishing—and how can it not?—is it any surprise that our self-concern would track, at least in part, physical continuity? I exhibit prudential concern for my physical continuers and not just because bodily continuity may be the physical source for my psychological continuity. I might be committed to bodily continuity on additional, independent grounds, like in the torture experiment when I shutter at the news that my physical continuer will be tortured. I am unsure if that person will be me, but that is not the

point. The relevant point is the underlying facts of the case. I am *certain* that the future person will bear a certain relation to me, and on those grounds alone my feelings of prudential concern are justified.

The picture of self-concern that I am pushing is independent of any theory of what constitutes a person. This account of self-concern coincides with my proposal to treat 'person' as a cluster-concept composed of components such as psychological and physical continuity as well as agency. Because all of these components are part of our lives, we exhibit self-concern in all of these domains. We care about the continued existence of our psychology, it is possible for us to care about our physical continuers, and it might even be possible for us to choose an agency continuer as an identificatory surrogate. During everyday life these components coincide, but in thought experiments we are asked to choose between these components in an effort to determine which component is more central to personhood. But if this question is an empty one, then we can safely admit that it is at least theoretically possible to express self-concern in each of these domains. And the way that we prioritize our self-concern in these different domains may be informed, at least in part, by what we value. Here's why: different people are committed to different values. I say this not in the sense in which a relativist about ethical values might say this. Rather, it is to suggest that people live different lives and so they value different things. Some people value intellectual achievement and mental curiosity. Others value physical perfection. Still others value the beauty of nature or the safety of their neighbors. There is no limit to the things that we value.

It is possible to value a body, it is possible to care about physical attributes. This is not just because one might have a superficial attachment to beauty. One might value physical strength because it is necessary for one's (quite possibly noble) life plan. Perhaps, if I am a father, I value being a father to my son. And being a father is, at least in part, a biological relation between a body and its offspring, an attachment which is not preserved entirely by continued psychological continuity in a *new* body. But if the continued existence

of my body is important because of what I value, it would not be irrational to demonstrate component-concern for my physical continuer, regardless of whether he is me or not—*precisely because this last question is empty*. That is why we need not dictate that concern for one component is rational while concern for another is irrational. Not everyone has—or even *ought* to have—the same values. That is why it has been so difficult to offer a normative account of what matters in survival. Part of it depends on what you *value*. Some of these values require psychological continuity, others require physical continuity.

A philosophical account of self-concern that ranks the different components must make at least some value judgments. I am not suggesting that these judgments can't—or shouldn't—be made. My account has so far advanced no substantial thesis about which variety of component-concern is more important. All I have shown is that prudential concern for one's physical or agency continuers is at least theoretically possible. And I have suggested the following explanation for the plurality of responses in cases of self-concern: which continuer you care about may be partly influenced by your values. My account makes no normative claim about which domain of self-concern has primacy, according to either the demands of rationality or the demands of ethics. What I *am* suggesting is that any theory of self-concern should operate within the conceptual landscape that I have sketched. And what is important to note is that this incredibly rich conceptual landscape about self-concern makes no reference at all to persons.

At this point in the argument a door opens for an objection. If self-concern attaches to our values, commitments and projects—what we want for ourselves and for the world—doesn't this imply that persons are agents and that my eliminativism is unwarranted? The objection goes like this: in talking about "what we value" we have lapsed into a nomenclature typically associated with agents. Agents pursue plans and projects to achieve the things that they value. In deciding what you value you must first deliberate—another hallmark of agency. The objection continues: this lapse of language is no accident. It is an indication that you have really been talking about agency. This might be good suggestive

evidence for an agent-centered account of personal identity. Korsgaard has offered a Kantian analysis of persons as unified agents. And in an effort to jump-start the deadlock between the animalist and psychological criteria for personal identity, Rovane offered the revisionary proposal that persons are agents with a unified rational point of view and a commitment to overall rational unity. The objection suggests that we have reason to support one of these agent-centered accounts of personal identity over my eliminativism.

But this would be an unwarranted conclusion; the facts do not require it. Nothing here requires us to embrace an agent-centered approach to personal identity. Rather, we could stop short and admit that agency plays a more *central* role in self-concern than was first apparent. There is no reason, however, to go further and take this as evidence for a particular account of personhood. If the objection is correct (and I am not conceding that it is), one might admit that prudential concern for both psychological and physical continuers is fundamentally derived from *agency*-concern, which is the more primary phenomenon. Under this view, we value psychological and physical continuity in part because these things help us, as agents, achieve our projects. So when we express prudential concern in these domains we are really expressing self-concern for our agency. But the objection gives us no specific reason to take this alleged fact about self-concern as evidence for a particular account of personal identity. The evidence can be explained by changing slightly our account of self-concern, so why go the more extreme route of taking it as evidence for a revisionary, agent-centered account of personal identity?

It is a question of utility. As explained in the second chapter, the eliminativist strategy ought to be evaluated on utilitarian grounds: the revision is warranted if it helps us clean up our conceptual landscape. This standard goes for other revisionary strategies as well, including the strategy of equating persons with agents. And it seems unlikely that the alleged benefit here (for self-concern) would outweigh the disadvantages. The first disadvantage, already explored in the previous chapter, was the violation of our animalist and psychological intuitions about persons. Another disadvantage, to be considered in the next chapter, is that

an agent-centered approach might yield counter-intuitive results about moral rights. In various cases we are committed to the moral rights of beings even when they are no longer agents; conversely we are not committed to the moral rights of some atypical agents. The totality of these consequences must be taken into consideration when evaluating the relative merits of different revisionary proposals. And it seems to me that an agent-centered approach performs no better than its eliminativist alternative and may, in some cases, perform a little worse. All of this for the supposed benefit of an improved understanding of self-concern. But this understanding is already available with the eliminativist strategy because it too allows for agency-concern.

§4.5 THE FIRST-PERSON POINT OF VIEW

I must address an obvious anxiety about eliminating the concept of the person: what becomes of the first-person point of view? I suggested that self-concern be replaced with component-concern; this recognizes that we demonstrate prudential concern for our physical, psychological and agency continuers. Can we reconcile this multiplicity with the first-person point of view over time? We must be able to think of ourselves as not just one entity in the future but several. Can we project the first-person point of view into the future in this way? Furthermore, doesn't eliminativism require us to give up the unity of the first-person point of view in the present as well? Surely this is impossible. The eliminativist position has not just eliminated the conception of the person, it seems to have eliminated the subjective self, the phenomenological point of view. After all, we think of ourselves—in the present—*as one thing*. Although we can suspend judgment about which future continuer in the Williams thought experiment we would be, it is impossible to suspend such judgment about ourselves in the present. I *think* of myself as a single entity, I *refer* to myself as a single entity ('I'), and this notion seems hardwired into our cognitive structure. We can't think of ourselves as being three *separate* things without losing our sense of self. We can't give up our

conception of ourselves as unified beings in the present, and the concept of the person is the best concept we have to express that self-conception.

Luckily there are theoretical precedents from which we can draw a response. For example, in *The Principles of Psychology*, James posits the existence of multiple selves: the material self, the social self, the spiritual self and the pure ego. But these are not meant to be taken as multiple entities, as different metaphysical beings like souls, of which there are several floating inside your head. Rather, James calls them "constituents of the self" and even refers to language that evokes the account of component concern that I put forward in this chapter: "A tolerably unanimous opinion ranges the different selves of which a man may be 'seized and possessed,' and the consequent different orders of his self-regard, in an hierarchical scale, with the bodily Self at the bottom, the spiritual Self at top, and the extracorporeal material selves and the various social selves between." And then later, James continues:

> So it comes to pass that, as aforesaid, men have arranged the various selves which they may seek in an hierarchical scale according to their worth. A certain amount of bodily selfishness is required as a basis for all the other selves. But too much sensuality is despised, or at best condoned on account of the other qualities of the individual. The wider material selves are regarded as higher than the immediate body. He is esteemed a poor creature who is unable to forego a little meat and drink and warmth and sleep for the sake of getting on in the world. The social self as a whole, again, ranks higher than the material self as a whole. We must care more for our honor, our friends, our human ties, them for a sound skin or wealth. And the spiritual self is so supremely precious that, rather than lose it, a man ought to be willing to give up friends and good fame, and property, and life itself.[75]

The interesting idea expressed in these passages—and it is an idea which rings true—is the degree to which we are capable of looking at ourselves in different modes, through different lenses, with different concepts. James is *not* suggesting in this passage that we are multiple entities magically conjoined in a single being, as a Cartesian might think. Rather, he is suggesting that there are multiples ways that we can think of ourselves.

Examining the analog of this problem in another thinker might prove instructive. Kant has two versions of the self: the phenomenal and the noumenal. The two selves straddle the traditional Kantian distinction between what is subject to scientific law,

---

[75] William James, *The Principles of Psychology* (New York: Dover Publications, 1950), pp. 314-5.

observable and hence knowable (the phenomenal) and the world of the thing-in-itself of which we must necessarily remain ignorant (the noumenal). The noumenal self is active, free and a suitable object of moral evaluation. The phenomenal self is *acted upon*; consequently it is neither free nor an appropriate object of moral respect. This seems to deepen our anxiety, not improve it any. Kitcher notes that "it is not clear how the I of apperception can be fitted into this scheme. Kant is staggeringly ambivalent on this fundamental point, and in the end he refuses to categorize it as either phenomenal or noumenal."[76]

But it is important to remember that in addition to referring to two selves, Kant also talks about a *single* self as considered from multiple vantage points—one phenomenal, the other noumenal. And the overall incoherence with the rest of Kant's system—an issue addressed by Kant scholars like Kitcher—is not our burden here. What is relevant to our problem is the notion that we can view ourselves from multiple perspectives.

I want to suggest that my account of self-concern yields the same fate for the first-person perspective after elimination of the concept of the person. Remember: eliminativism is conceptually motivated. I have suggested that we stop using the concept of the person and use different concepts instead. And just as it is coherent to think of viewing ourselves from two Kantian perspectives (and I think this is what James was getting at), so too it is coherent to use multiple concepts—biological human being, rational agent, etc.—when engaging in the first-person point of view. If this seems difficult, just picture the same phenomenon when considered over time. I can express self-concern for my biological continuer and for my psychological continuer. There is nothing incoherent there. In the same fashion I can think of myself *in the present* using the concept of a biological human being *or* another component concept. None of this requires giving up the first-person point of view.

It seems as if the anxiety about losing a grip on the first-person point of view originated in thinking of eliminativism in ontological terms. It is worrisome to think of

---

[76] Patricia Kitcher, *Kant's Transcendental Psychology* (New York: Oxford University Press, 1990), p. 139.

ourselves as multiple entities or multiple selves in the ontological sense of that expression. It makes us seem too much like Cartesian monsters, an unlikely combination of disparate elements that should never have been combined. But eliminativism is a claim about the *concepts* that we should use. Once we remember that, it becomes clear that the first-person point of view is consistent with thinking of ourselves under multiple concepts.

§4.6 THE VIRTUES OF ELIMINATIVISM

If the analysis of this chapter is correct, we have demonstrated that not only is it possible to offer an account of self-concern that makes no reference to persons, it is in many ways preferable to the traditional way of analyzing the phenomenon. Remember, we undertook an analysis of self-concern in the first place because it was a potential stumbling block to the eliminativist strategy. I suggested elimination of the concept of the person because our conflicting intuitions are irresolvable by any other means. But this strategy is only plausible if we can offer personless accounts of phenomena traditionally analyzed with the concept of the person. In this chapter I have responded with an account of self-concern. As it turns out, not only is a personless account possible, it offers many advantages to a traditional, person-centered analysis of self-concern.

The most promising virtue of the eliminativist account of self-concern is that we no longer need to take our prudential concern for our physical continuer during the second trial of the Williams thought experiment as evidence for an animalist account of personal identity. This would be problematic because it would violate our deep psychological intuitions about personal identity. It would be better if we could make sense of this example of prudential concern without requiring us to be animalists about personal identity. The corollary also is problematic. If we take our prudential concern for our psychological continuer as evidence of a psychological account of personal identity, we run the risk of violating our animalist intuitions about personhood. One way to avoid this dilemma is to

split the assumed link between personhood and self-concern—which is exactly what eliminativism aims to do. In the process, fewer intuitions are violated.

A second virtue of this eliminativist account of self-concern is that we can recognize the diversity of our prudential concern. Previously we were unable to do this. The assumed link between personhood and self-concern prevented recognition of this diversity. We could only make sense of one kind of prudential concern because anything more would prove problematic for an account of personal identity. Any prudential concern that did not track the alleged "essential" element of personhood—either the physical element for an animalist or the psychological element for a Lockean—would have to be declared either irrational or not a bona fide case of prudential concern. Otherwise the theorist would be faced with a contradiction. How could you have *prudential* concern for a physical continuer if personal identity is largely psychological? Conversely, how could you have prudential concern for a psychological continuer if personal identity is largely animalist? Certainly neither of these two phenomena are explainable if you maintain the assumed link between personhood and self-concern. But with eliminativism the dilemma disappears. Having eliminated the cluster concept in favor of its components, we can recognize that it is at least conceptually possible to have feelings of prudential concern for our psychological, physical *and* agency continuers. This demonstrates not just the plausibility of eliminativism but its virtues as well.

In previous chapters I have suggested eliminating the concept of the person for metaphysical reasons. I outlined our conflicting intuitions about personhood as exposed by the Williams thought experiment and I suggested that any account of personal identity that prioritizes animalism over psychologism, or vice versa, will inevitably violate one of our intuitions. This was due, in part, to the concept's status as a cluster concept, where diverse (and in fact competing) components have been housed under a single term. Moreover, I demonstrated that the methodology of thought experiments was unlikely to yield a resolution of the conflict by elevating one set of intuitions over the other. A better resolution to the conflict, I argued, would be to eliminate the cluster concept altogether in favor of its components: biological human beings, psychological continuity, and rational agency.

It is no surprise that the eliminativist strategy involves violating a major intuition of its own: that persons exist, that they are metaphysically significant and axiologically indispensable. Person-talk has so dominated both our value theory and our metaphysics that elimination seems like a drastic step. But one must remember, first of all, that eliminativism is not ontologically motivated. Second, though, it is also important to question the assumption of personhood's supposed indispensability for value theory, something which I think has rarely been done. With this re-evaluation in mind, I have undertaken an investigation in the previous two chapters into responsibility and self-concern. As it turned out, not only was it possible to offer accounts of these phenomena without making reference to persons, doing so provided certain explanatory advantages unavailable with the concept of the person. Quite simply, the hallmark of the cluster concept—its housing of diverse components under a single term—also proved to be its Achilles Heel in matters axiological. Separating out the components individually allowed a fuller understanding of responsibility and self-concern and allowed us to make sense of some marginal phenomena that the traditional, person-centric strategy dealt with poorly.

All of this leads us to the final and most important axiological subject: moral, legal and political *rights*. The most pressing objection facing the eliminativist position is the status of these rights. It is the *person* who is traditionally considered to be the locus of moral concern and it is the *person* who is assumed to have certain *rights*. Consider the following sampling of recent moral philosophy:

> [T]he statement that some person or group of persons has a certain right entails a correlative 'ought'-judgment that all other persons ought at least to refrain from interfering with that to which the first person or group has the right.[77]

> Offhand it hardly seems likely that persons who view themselves as equals, entitled to press their claims upon one another, would agree to a principle which may require lesser life prospects for some simply for the sake of a greater sum of advantages enjoyed by others.[78]

> Persons' rights are rights to particular performances and are rights over particular items of property.[79]

> Morals by agreement offer a contractarian rationale for distinguishing what one may or may not do. Moral principles are introduced as the objects of fully voluntary *ex ante* agreement among rational persons.[80]

In all of these cases—and many more that I trust I need not share—it is the concept of the person which is the center of the moral discourse. It is *persons* who are the proper objects of moral respect, it is *persons* who are parties to the social contract, it is *persons* who are subject to the categorical imperative, it is *persons* who are behind the veil of ignorance, it is *persons* who have rights. If we eliminate the concept of the person, what happens to these moral theories?

This is a pressing issue because of the near-hegemony of rights-based discourse. But the objection does not hinge on this near-hegemony; some moral theories that are not so concerned with rights still have a notion of the person as a center of moral respect. And elimination of the concept of the person would prove deeply problematic if these moral discourses—including rights-talk—were collateral damage. Elimination looses its plausibility if it forces us to give up our ethical discourse that requires the use of the concept of the

---

[77] Alan Gewirth, *Reason and Morality* (Chicago: University of Chicago Press, 1978), p. 133.

[78] John Rawls, *A Theory of Justice* (Cambridge: Harvard University Press, 1971), p. 14.

[79] Loren E. Lomasky, *Persons, Rights, and the Moral Community* (New York: Oxford University Press, 1987), p. 142.

[80] David Gauthier, *Morals by Agreement* (New York: Oxford University Press, 1986), p. 9.

person. Part of this traditionally person-centric discourse is any moral framework that attributes rights. There are countless examples where rights are attributed to moral or legal *persons*. And if we can't make use of our concept of the person, what becomes the locus of moral rights?

My goal in this chapter is not to defend one foundation for moral rights or one moral theory or outlook. Rather, my point is to defend the strategy of eliminativism by demonstrating its plausibility—i.e. that moral theories in general that derive rights can still do so even without the concept of the person. In other words, I want to preserve a space in which this moral discourse can continue even in the absence of the concept of the person. I take this to be a necessary task in defending the plausibility of the eliminativist position. If eliminating the concept of the person required foregoing all rights-based discourse, this would very much count against the strategy (although it is possible to engage in moral theory without rights-based discourse.) So the burden is on me to show that this discourse can continue without the concept of the person. Furthermore, it would count as a *virtue* of eliminativism if it could be shown that eliminating the concept of the person in favor of its components offered some conceptual advantages for analyzing rights-talk. Remember, it turned out that elimination of the concept of the person offered some distinct advantages for our accounts of responsibility and self-concern. Perhaps the same will be shown to be the case about rights. If it is, it will give us another reason to pursue the eliminativist position. I am not claiming that we *must* accept a rights-based discourse, that any particular theory of rights is correct, or that we have or don't have certain rights. All that I aim to show is that any significant moral, political or legal *claims* about rights can be made without the concept of the person. Moreover, they might be made *better* without the concept of the person. One way to carry out this investigation is to look at cases of rights where the ascription of rights to a being is controversial—controversial because there is no consensus if the being is indeed a 'person'. Perhaps eliminating the concept of the person will help resolve this difficulty.

How might our rights-based discourse be improved by eliminating the concept of the person? As a general rule, theories of rights offer some definition of personhood and for various theoretical reasons (often appealing back to features of the definition) they then attribute rights to persons. However, the general structure of this argument proves problematic because it requires us to deny rights to entities that do not meet the definition of personhood being used. (These beings don't meet the definition of personhood usually because they lack one of the components). Who are these entities? These entities who do not qualify for personhood do not come from the strange recesses of *Gedankenexperimente*. Rather, they come from the margins of everyday life; they are the sources of controversy that plague ethics and moral theory, the law, politics and religion. They are the controversial cases that we debate in the courtroom, the classroom, the hospital and the church.

A patient in an irreversible coma may be entitled to some rights because he is a functioning human animal but not entitled to other rights because he is no longer a rational agent. This case (and cases similar in structure to this one) demonstrate not just the plausibility of eliminativism but its virtues as well. Theorists who associate personhood with just one of its components (say agency in the case of Gewirth) and who believe that *persons* have rights, are forced to deny rights to entities who fail to qualify for personhood. This is troubling because it violates our ethical intuitions. We sometimes believe that it might be wrong to kill a patient in an irreversible coma. And we also sometimes believe that group agents—even though they are agents—do not have the same rights as biological human beings. Eliminating the concept of the person and dealing instead with the component concepts gives us a conceptual framework within which we can do justice to these intuitions. I call this the exclusion problem and I will offer an expanded analysis of it in §5.2.

Eliminativism has the potential to solve the exclusion problem. Although we will have to wait to see the results, the potential is justified for the following reasons: without the concept of the person we can look to the components to justify our ascriptions of rights. After having eliminated the concept of the person, I identified the following component

concepts: biological human beings, psychological continuity and rational agency. These facts are morally significant for the attribution of rights. We confer rights based on which components are present. We can decide which rights stem from which components and ascribe rights on that basis. Then the controversial cases of ethics (where a component is missing) will be easier to deal with. We won't feel compelled to deny rights simply because the being does not meet the criteria for moral or metaphysical personhood. We will be able to look at what really matters—which components *are* present—and attribute the appropriate rights for the appropriate reasons.

So not only is eliminativism plausible—it might be preferable. Elimination of the concept of the person—with its myriad components sheltered under a single umbrella—has the capacity to acknowledge that different rights might be attributable to different concepts. In other words, all rights don't have to flow from the same source. This might nicely explain or resolve our ethical intuitions in cases where one of the components is missing. For example, if agency were the source of all rights we would be forced to admit that a human being who is no longer an agent—because he, say, languishes in an irreversible coma—has no rights whatsoever, including the right to life. But with the concept of the person eliminated, we could admit that rights stem from a different source: the components of the cluster. And if the components don't all go together all the time, that's fine too. While some rights derive from our status as agents and the necessary features of generic action (as Gewirth might say), other rights might stem from our capacity to feel pleasure or pain, and still other rights might stem from our status as functioning biological animals. Eliminating our concept of the person might shed light on the fact that the conceptual foundation of rights-based discourse might not be as unitary as we had thought. So as a final benefit, we might recognize that *different* rights have *different* sources—that moral theory is not as unified as we might have hoped. Later in this chapter I will refer to this shift as moving from a univalent moral theory to a multivalent moral theory.[81] And I will suggest that a

[81] I borrow these terms from Lomasky.

multivalent moral theory of rights could appeal to the component concepts that I identified in previous chapters. Eliminativism might pave the way for such a theory.

As I said before, this chapter has two goals: to show that a theory of rights is possible without the concept of the person and to show that it might be *preferable*. In §5.1 I will ask where rights come from and consider the possibility that their true source is the component concepts and not the concept of the person itself. If that is the case then eliminativism might not be fatal to rights talk. Then in §5.2 I will present the problem of exclusion and suggest that elimination of the concept of the person may be the best way to solve the problem. This would be one way of showing that eliminativism is not just possible but preferable. Finally, the remaining sections of the chapter will look at how these two goals manifest themselves in *specific* theories of rights, especially Kantian and neo-Kantian theories of rights that make heavy use of the concept of the person. If the theoretical analysis of the first section is correct, it should be possible to "rewrite" these theories without making reference to persons. If the theoretical analysis of the second section is correct, doing so should provide some specific advantages, namely a preferred solution to the problem of exclusion.

§5.1 WHERE DO RIGHTS COME FROM?

I promised to demonstrate in this section the possibility that a discourse on rights might be able to continue without the concept of the person. I have argued in previous chapters that the concept of the person is a cluster concept composed of several distinct components. Furthermore, I showed that the phenomena traditionally associated with the concept of the person (such as responsibility or self-concern) actually tracked the components. It was the components that were doing the argumentative work. Perhaps the same is true of rights. Perhaps the appeal to personhood in rights-talk masks a deeper connection to the *components* of the cluster concept.

Moral theories of rights that make explicit reference to the concept of the person are making an implicit reference to one or more *components* of the cluster concept. Theories that

attribute rights to persons are appealing—latently—to the component concepts that I have identified. Because it is the component that is doing the argumentative work—not the cluster concept—the moral theory can continue after elimination of the concept of the person. There are two kinds of argument that could demonstrate this. One would be an abstract argument that appeals to Parfit's Argument from Below. I will pursue that line immediately. But the other possibility is an argument by demonstration—i.e. showing that actual theories of rights that appeal to the concept of the person can be rewritten with the component concepts. I will delay this argument until the later sections of this chapter.

In previous chapters I identified the importance of Parfit's argument, which states that it is lower level facts which are morally significant. Parfit used this argument as part of his reductionist claim that the facts of personal identity could be reduced to facts about physical bodies and the mental and physical events surrounding them. I contrasted Parfit's Argument from Below with its opposite, Johnston's Argument from Above, which argues that the lower level facts are only important insofar as they constitute personal identity. I offered a defense of Parfit's Argument from Below and demonstrated several applications of this argument in the chapters on responsibility and self-concern. But now we see what might be the argument's most powerful application. It is the lower level facts of personhood which are morally significant; the fact that someone is a person is only morally significant because being a person involves certain lower level facts which are morally significant. Consequently, it might be said that persons have rights because of those lower level facts. If this were so, eliminating the concept of the person might seem less problematic than we had once thought. We can still make reference to those lower level facts. I shall demonstrate in later sections that many moral theories that attribute rights to persons are implicitly committed to the moral significance of those lower level facts.

Consider the following. Being a person involves many things: being a biological human being, exhibiting psychological continuity, and being a rational agent. All of these components are part of being a person. In the previous chapters my rationale for eliminating

the concept of the person was precisely that any account of personhood that tried to prioritize one component over the other would inevitably violate a major intuition about what it means to be a person. I also suggested that if you know the facts about biological human beings, psychological continuity, and rational agency, you know all of the facts that there are. Asking whether some being qualified for personhood was not a question about the being itself but a question about our use of the term 'person' and when it was appropriate to use it. Little or nothing hinged on solving the riddle of 'personhood' if you knew all of the underlying facts.

The same thing goes for rights. If rights are usually associated with persons, and facts about persons reduce to facts about biological human beings, psychological continuity, and rational agency, then it is these lower level facts which must track our attributions of rights. If knowing the facts about the components is all there is to know, then how could something's being a person, in and of itself, be the source for an attribution of moral worth? Asking, in addition to these facts, whether something qualifies for personhood is a question about our use of the term 'person' and not a question about the being. We already know what there is to know about him.

Let us consider the issue from the reverse. The Argument from Above claims that what matters is the higher level facts. Being a biological human being, with psychological continuity and rational agency, only matters when these facts constitute being a person. In the absence of being a person, these lower level facts are not significant. So even if some of the components are present, they are insignificant because in that case they do not constitute being a person. The significance migrates from the top down and stems from a being's status as a 'person.' The components do not matter in themselves, they only matter when they constitute being a person.

But it is unclear how the significance could migrate from the top down in any non-circular way. In order for this to be the case, personhood would have to be something other than just the presence of the components, otherwise it would have to be the case that the

significance came from the components; there wouldn't be any other place for it to come from. One way for the significance to come from above would be if personal identity was given by some deep further fact, as Parfit calls it. In this case, the presence of the components would not be significant in itself but only because it also signaled the presence of the deep further fact of identity. In cases when it does not signal that presence, the components lose all derivative importance. While this alternative is possible, it is rather implausible. We have already questioned the existence of a deep further fact of personal identity by looking at some Parfitian thought experiments. If the only way to support the argument from above is to ontologically commit oneself to a deep further fact, this is indeed a heavy price to pay. But perhaps there is another way to support the Argument from Above.

A second way for the significance to migrate from above would be if personhood is a conventional notion that is applied to a being within a social context by a given community. To be a 'person' is to be recognized as a 'person' by the rest of the community. In the absence of this recognition, the lower level facts mean nothing. The lower level facts only mean something when status is conferred upon them by the community by labeling a being a 'person'. In the absence of a larger moral community with a system of mutual recognition with other persons, the components themselves are meaningless. The components would mean nothing if an individual is stranded on a deserted island. In this case, the Argument from Above makes no reference to a deep further fact.

But this second possibility for the Argument from Above is a poor one. Even if this system of mutual recognition exists, we are still obliged to ask the question, what is this system of recognition based upon? The system of recognition and conferral of the moral status of personhood must be based upon something, i.e. there must be criteria by which the conferral is made. Which beings get the recognition and which beings do not? These decisions must be based on something deeper than the system of mutual recognition itself, otherwise they are *purely* conventional, in which case it is unclear how they can be anything other than arbitrary. They would not be morally justified. So while the Argument from

Above is possible, it seems rather implausible in the case of persons. Unless one is willing to accept a deep further fact, it is difficult to understand how the importance could come from above.

If we accept the Argument from Below we should recognize that rights stem not from the concept of the person itself but from the lower level facts, i.e. the components of the cluster. This recognition would transform our understanding of rights. At the very least, this will help provide a better process for deciding the controversial cases of rights in ethics and legal theory. Our method up to now of investigating rights is symptomatic of an inappropriate attachment to the Argument from Above. Consider how controversial or marginal cases of moral rights are adjudicated in ethical theory. The traditional avenue has been first to decide the metaphysical question of whether these individuals are "persons". (Usually some element which is typical of personhood is absent, thus making the issue metaphysically controversial). Once a verdict is reached on this question it supposedly entails various rights claims because an individual's rights flow from the moral significance of her status as a person. But the Argument from Below and eliminativism change this procedure for the better. Whether we decide to call someone a 'person' in a marginal case cannot be morally significant. What *is* morally significant is the lower level facts that constitute being a person. But if it is the lower level facts which are morally significant in marginal cases (not the higher level fact of whether the term 'person' is appropriate), then there is no reason to suggest that the same is not true in every case. It is the lower level facts that are morally significant when it comes to attributing rights. So if we know the lower level facts nothing hinges on whether we call someone a 'person'.

Later in this chapter I will show how the Argument from Below applies to *specific* moral theories of rights. Theories that attribute rights to persons are appealing—latently—to the component concepts that I have identified. I will consider Kant and the moral theorists in the "neo-Kantian" camp and then I will briefly consider some other theories of rights, including utilitarian and religious accounts. The working hypothesis will be that Kantian

moral theories appeal to an agent-centered definition of person and so will be unaffected by elimination of the concept of the person just so long as the concept of rational agency remains. And the non-Kantian theories can make use of the other components. In both cases they will not be fatally hurt by elimination of the concept of the person since the components are still at their disposal. But this demonstration must wait. I promised a theoretical explanation for how eliminating the concept of the person might make things better for a theory of rights by solving what I call the problem of exclusion.

§5.2 THE PROBLEM OF EXCLUSION

First, the problem: I want to lay out a general diagnosis of a problem facing many moral theories that ascribe rights to persons. Most of these theories take an essential attribute or attributes to be characteristic of persons. The moral theory then goes on to show that this attribute is an appropriate foundation for a moral right. The conclusion follows that persons have rights and sometimes, the account goes on to derive the *content* of these rights. This general argumentative structure can be found in many moral theories that attempt a derivation of rights.

The problem, of course, is that the structure of the argument requires denying rights to any being who does not carry the essential attribute or attributes that are argued to be morally significant. Consider the following example. Many moral theories, which I will outline below, ascribe rights to persons because of their capacity for rational agency. From this fact the content of a theory of rights can be derived. How this is to be done differs from theory to theory (as we shall see later in the chapter), but the point is that the derivation goes top-down from the concept of agency. As a consequence of this kind of theory, non-agents are excluded from the class of beings who are morally entitled to rights. Consider a patient in a hospital after a horrific car crash. He has suffered severe brain damage from the accident and, as a result, has no higher brain functions. He cannot speak or communicate in any way and shows no evidence of higher cognitive activity. Not only can he not perform actions

because of the state of his body, there is no evidence that he is performing mental actions in his head (math, geometry, deductive entailments, etc..) This state of affairs is confirmed by medical diagnosis which shows severe trauma to the upper regions of the patient's brain. That's why the patient is not capable of rational agency as it is usually defined by the moral theory that attributes moral rights.

But he is not brain dead. There is no damage to the lower portions of the patient's brain that control biological and animal functions. The patient's heart beats normally, the biological functions of his body continue under the direction of his lower brain, and most importantly, the patient can now breathe without help from a ventilator—so it is turned off. An apnea test was performed by a neurosurgeon. The surgeon disconnected the ventilator and watched to see if the patient breathed spontaneously within five seconds. He did, which is evidence of the "apnea response"—a function controlled by the lower brain. The brain detects a lack of oxygen in the blood flowing to the brain and in response stimulates the lungs to start breathing. Passing the apnea test means that the patient has *some* brain activity. He is not, in other words, brain dead.

Under any theory of rights that appeals to rational agency for a foundation of moral rights, the patient does not qualify for rights. Consequently there are no constraints on our conduct towards him. Or rather, if there are constraints, they are not constraints stemming from a claim that *he* has against us. His exclusion from the moral community violates our intuitions that the patient is still a viable source of some moral rights and that some constraints, if not necessarily the same ones as before (when he was healthy), still apply to our conduct towards him. We can't simply do what we want with him. We can't torture and kill him. He is still a living human being, after all, even if he is now severely defective. But our moral theory claims that he is no longer the subject of right claims because he is no longer a rational agent.

I want now to broaden this particular situation into a general problem. We can do this by identifying the cause of the problem. A theory of rights which appeals to the moral

significance of rational agency will exclude from the moral community human beings who are no longer rational agents. The problem is the identification of a single essential element which is supposed to carry with it moral significance. Therefore *any* moral theory of rights which appeals to the moral significance of one essential element will naturally exclude from the moral community any human being who lacks this essential element. Doing so may very well violate our ethical intuitions that these beings ought to be objects of moral respect. I call this the Problem of Exclusion.

Lomasky locates the source of the problem with the distinction between univalent and multivalent moral theories. I have already described the former variety; it locates moral significance in one essential attribute of personhood. The latter variety allows for more than one source of moral significance and consequently more than one source of moral rights. Lomasky explains what I call the Problem of Exclusion this way:

> A univalent theory of basic rights centered around project pursuit will hold that all and only those beings who are project pursuers are rights holders. A multivalent theory need not. It can recognize that project pursuers possess rights *and* that other beings who do not pursue projects are also rights holders. A univalent theory is tidier, but the costs of simplicity are high: characteristically, the inability to accommodate a range of moral phenomena not captured by the favored criterion yet the suppression of which is strongly counterintuitive. The upshot may be a torturous effort to assimilate untoward cases under the one criterion recognized by the univalent theory through strategies that forfeit the desired univalence. For example, if the basis of rights is found in *personhood*, understood as the possession of some elevated level of mentality, it may be noticed that the account denies to very young children the status of persons. They then are not rights holders. The result is uncomfortable; it seems *prima facie* dubious that infants merit no further moral attention than need be paid to a piece of cheese or a puppy.[82]

Lomasky is suggesting that the price of simplicity is too high and that a multivalent theory of rights is preferable to the alternatives. What are the alternatives? They include biting the bullet, an appeal to concepts like potentiality or proportionality, or a disjunctive account of personhood.

The first option is biting the bullet. This means that the univalent theorist simply accepts the fact that the theory denies rights to a being who does not exhibit the relevant attribute. In accordance with this theoretical finding, the theorist can take this opportunity

---

[82] Lomasky, pp. 39-40.

to revise his moral intuitions. After all, moral intuitions are not infallible. The univalent theorist might argue that he is engaging in some form of reflective equilibrium. He starts with his moral intuitions as a guide for setting up a moral theory. Once the theory is created, though, it turns out that it accounts for most—but not all—of our moral intuitions. This suggests that we should revise our moral intuitions that the theory cannot explain. This is biting the bullet.

The second option is the most popular among univalent moral theorists. They appeal to concepts such as potentiality or proportionality in order to justify the ascription of rights to the excluded being—still on the basis of the essential characteristic. This time, though, the essential characteristic is combined with concepts such as potentiality or proportionality. So, consequently, a young infant who is not a rational agent is a *potential* agent and deserves respect on that ground alone. Also, a patient in a coma is a potential agent in the sense that he or she might recover. Consequently the patient retains the rights he once had and also has rights by virtue of the agent he will once again be in the future.

Also, one might ascribe certain *limited* rights to a being if they fall below the relevant standard but still exhibit some lower degree of the essential attribute. Consequently, the being should be accorded a lower level of rights *proportionate* to their level of the essential attribute. Consequently, children and other human beings receive rights in proportion to their rational agency. The older that they get and the greater the degree of rational agency exhibited, the greater are their rights until they reach the age of maturity and join the moral community with *all* of its rights and privileges. Until then, though, they have the rights in proportion to their status.

The third strategy to solve the problem of exclusion is to appeal to a disjunctive account of persons. With this avenue there is only one way to qualify for rights (being a person) but there are multiple ways to qualify for personhood. Although the theory is univalent on the level of the value theory it is multivalent at the metaphysical level. For example, the exclusion of the car crash victim could be dealt with by defining personhood as

rational agents *or* living human beings (biologically defined). Any rational agent would be a person whether or not they are biological human being (viz. a sophisticated exemplar from another biological species), and any human being would be a person whether or not they are a rational agent (viz. a newborn infant or a patient with no higher cognitive brain functions). Since the patient is a person under this definition, he qualifies for moral rights. The problem of exclusion would appear to be solved.

There are various costs associated with each of the three strategies that I have outlined above. Biting the bullet may be possible in some circumstances, but reflective equilibrium is a two-way street. Yes, it is true that we should consider revising our ethical intuitions if our theory shows them to be unwarranted. But when the revision to our ethical intuitions becomes more than minor, reflective equilibrium requires us to go back to the theoretical side of the equation and attempt revisions there. This is warranted in this case because the ethical intuitions are so strong. We are strong in our conviction that many individuals, even at the metaphysical margins, are proper objects of moral concern. We just aren't always sure why. That in itself is cause to attempt a revision in the theory.

While the appeal to concepts such as potentiality and proportionality is promising, it only applies to some of the cases. Potentiality only works in cases where the being in question will gain the essential characteristic in the future (e.g. children), and proportionality only works when the essential characteristic is already present in lesser degrees. But these two concepts will not cover all of the pressing cases that make up the problem of exclusion. For example, the car crash victim described earlier will never be a rational agent and does not possess a lower level of rational agency that would be relevant to the principle of proportionality. So the problem of exclusion still remains in that case.

The third proposed solution was the disjunctive account of persons. This strategy ends up conceding much ground to the eliminativist. Offering a *disjunctive* account of persons is tantamount to conceding that 'person' is a cluster concept. If the definition of personhood is genuinely disjunctive, i.e. if either an agent *or* a human being can qualify as a

person, then either an agent or a human being can qualify for rights. This would seem to solve the problem of exclusion because no one would be excluded from the moral community. But we should remember the Argument from Below from the previous section. It is the disjuncts themselves which must be morally significant, not the overall concept of which they are disjuncts. And if it is the disjuncts that are doing the argumentative work here, not the concept of the person, it is now unclear why we created the disjunctive definition in the first place. Why not simply do away with the concept of the person and stick with the disjuncts? As I hope to show in this chapter, the best solution to the problem of exclusion is precisely that—to recognize that the concept of the person is a cluster concept with distinct components, several of which may carry moral significance. This encourages us to eliminate our concept of the person and facilitate a switch to what Lomasky calls a multivalent theory of rights.

Where did the problem of exclusion come from? The problem says something about our theoretical need to offer a unified account of rights. Moral theorists, like all philosophers, value simplicity and elegance in their theoretical machinery. Simplicity is to be valued over complexity as an intellectual and aesthetic virtue. When it comes to rights, it is simpler to argue that they must all come from the same place. We must figure out what we are and why this endows us with rights. The assumption has always been that what we are is persons and that there is a particular element to personhood that is morally significant. But solving the problem of exclusion requires us to question that assumption. Perhaps there is more than one element to our existence which is morally significant. This may result in a moral theory that is less elegant. But the less elegant theory may turn out to be the right one.

§5.3 KANTIAN THEORIES OF RIGHTS

For the moment I want to divide up the landscape of moral theories into those that justify rights directly on some deontic ground and those that allow rights *only* by virtue of some consequentialist calculation. The former category is made up of Kant and the modern

theorists he inspired, including Rawls and other contractarians such as Scanlon and Gauthier. Another moral theorist who I include in the Kantian camp is Alan Gewirth, who derives rights from what he calls the generic features of action itself. All of these theorists are vaguely Kantian in the sense that their notion of rights is derived from the enlightenment concepts of human freedom, reason and rational action—either directly or indirectly.

In the second camp are utilitarians—particularly rule utilitarians—who reject natural rights but might accept a consequentialist foundation for a political or legal system that confers rights upon persons. (As for act utilitarians, although they usually don't express their moral theories in terms of rights, they do nonetheless have a notion of an object of moral concern that ought to be included in the class of beings whose utility should be maximized. This object is usually taken to be persons—although this is a matter of some controversy in moral theory, *viz.* Singer's version of utilitarianism.) We need to investigate whether this arena of discourse can continue without the concept of the person. A third miscellaneous camp includes religious doctrines that attribute rights to persons on some theological ground. These theories, while rare in contemporary moral philosophy, are nonetheless popular in the public debate of controversial subjects such as abortion and euthanasia. Also to be considered is the effect of eliminativism on the widening discourse of human rights. These theories will be considered in §5.5 and §5.6.

The striking thing about the first camp—Kant and those he inspired—is the degree to which the theories make explicit reference to agency for their ground. And in fact, it is the *rational* deliberation that encompasses purposive action that is the relevant aspect for Kant, Gewirth, Rawls, Scanlon, and Gauthier. We can therefore explore the degree to which these theorists are actually attributing rights to agents, not persons. If this is so, eliminativism poses little problem to the project of attributing rights, since agency is one of the component concepts that remain after elimination.

Kant sets the ground at the beginning when he claims that for a moral system to be universal and necessary it must be purged of all empirical considerations. Kant writes that

"the ground of obligation here must therefore be sought not in the nature of man nor in the circumstances of the world in which man is placed, but must be sought a priori solely in the concepts of pure reason..."[83] Already we see the connection between morality and reason—a connection which forms the basis for much of the history of moral philosophy. As it turns out, this connection between morality and reason, which begins with Kant (and which I will follow all the way to its natural conclusion in Gauthier's marriage of rational choice and social contract theories), is the story of rational agency.

Let us investigate this aspect of Kant more closely. Kant's moral prescription that "I should never act except in such a way that I can also will that my maxim should become a universal law"[84] is explicitly a prescription about *action*—and it is precisely one's ability to act in accordance with the categorical imperative that serves as the ground for one's rights claims. Chiefly, that claim is a right to respect which one's nature as a rational animal affords one. That right is also expressed in Kant's dictum that persons have a right to be treated as ends in themselves, as they would be in the Kingdom of Ends. He writes that "rational beings are called persons inasmuch as their nature already marks them out as ends in themselves, i.e. as something which is not to be used merely as a means and hence there is imposed thereby a limit on all arbitrary use of such beings, which are thus objects of respect." What exactly is that nature that marks persons as ends in themselves? Kant is quite clear about this. "The ground of such a principle is this: rational nature exists as an end in itself."[85]

So it is the capacity for reason, and furthermore the capacity to act freely and to legislate in accordance with the categorical imperative, that is the ground for a Kantian right. But what reason is there to attribute this capacity or faculty to the concept of the person? The concepts that Kant has traced here are precisely the ones I have associated with

---

[83] Immanuel Kant, *Grounding for the Metaphysics of Morals*, translated by James W. Ellington (Indianapolis: Hackett Publishing Co., 1993), p. 2.

[84] Kant, p. 14.

[85] Kant, p. 36.

agency—not personhood. Kant is emphasizing the *rational relations* between an agent's action and the universal legislation under which he acts. This is the hallmark of agency. Indeed, consider what Kant means by his ideal Kingdom of Ends. He says that "A rational being belongs to the Kingdom of Ends as a member when he legislates in it universal laws while also being himself subject to these laws." Consequently, "autonomy is the ground of the dignity of human nature and of every rational nature."[86]

Notice the language here. The Kantian right to respect is not limited to human nature but rather to *rational* nature—presumably wherever it can be found. It is evidence then that it is the capacity for *rational action*, subject to the demands of morality, that in turn make a being an appropriate object of moral respect. All of these rational relations are preserved in the account of agency that I have explored in previous chapters: the pooling of information, deliberation on the basis of that information, and formulation of a project, plan or response. It is precisely one's capacity for this kind of deliberation that makes one an object of a right-claim. It is by virtue of one's status as an agent that allows you to be in the Kingdom of Ends. There is nothing here essentially about Kant's concept of the person that is not antecedently provided by the concept of agency. That's because Kant is appealing to our faculty of the will:

> Our own will, insofar as it were to act only under the condition of its being able to legislate universal law by means of its maxims—this will, ideally possible for us, is the proper object of respect. And the dignity of humanity consists just in its capacity to legislate universal law, though with the condition of humanity's being at the same time itself subject to this very same legislation.[87]

Some might make the following objection: Although the rational relations that Kant describes as being the grounds for moral respect are the same as found in our concept of agency, Kant elsewhere demonstrates that the project of inquiry and of the rational relations that compose it requires a fundamental unity of consciousness, an ego which Kant describes as "formal" because it refers not to an ontologically defined ego but rather to a single

---

[86] Kant, p. 40-1.
[87] Kant, p. 44.

phenomenological point of view. The objector goes on to note that this fact relates agency to both psychological notions and physical animal nature—notions which can only be brought under the rubric of the concept of the person. Hence Kant's notion of rights—when explored at this deeper level—demonstrates its hostility to the eliminativist strategy.

But there's a ready response to this objection. The objection does not so much attack eliminativism as it does my contention—borrowed from Rovane—that agency is primarily a concept of rational relations that can be distinguished from both phenomenological unity and the physical body. Remember, we asserted that the rational relations implicit with agency can hold across phenomenologically unified beings and within a single biological organism. And Kant most certainly did not hold this view. It is *this* idea that the objection, following Kant, seeks to demolish.

Secondly, even if Kant is right, it is still possible for the two components to come apart. While Kant might have said that the demands of cognition of one being *require* mental unity, this doesn't mean that all physical beings must have cognition. *If* they have cognition, then it must be unified by the demands of reason, according to Kant. But if they don't have cognition (*viz.* a hospital patient with no higher cognitive functions), the conditional does not apply because the antecedent does not obtain.

There's also little in Kant's analysis of the unity of consciousness that requires the concept of the person. (Remember: the inspiration for eliminativism was the belief that no account of personal identity ought to prioritize our psychological intuitions over our biological intuitions, or vice versa.) The same ideas could very well be expressed with the component concepts: biological animals and their psychological continuants, and agents (which may or may not coincide with the former). More importantly, our concept of agency is perfectly suited to ground Kant's notion of the dignity of human nature. Both highlight rational nature leading to action and the capacity for rational thought and freedom, deliberation and legislation.

§5.4 NEO-KANTIAN THEORIES OF RIGHTS

In this section I will consider the contemporary moral theorists who draw their inspiration from Kant in one way or another, either by an emphasis on action alone or by an emphasis on reason and the social contract: Gewirth, Rawls, Scanlon and Gauthier. Although there are other neo-Kantian moral theorists, I take this group to be representative of the recent history of non-consequentialist moral theory. As we shall see, all of them attribute rights to persons on the basis of their status as agents. Consequently, all of these accounts will survive an elimination of the concept of the person just as long as the concept of the rational agent is still available. In this section I want to make two points: (1) their theoretical foundation for ascribing rights to persons is based exclusively on their status as rational agents and would thus be unperturbed by elimination of the cluster concept; and (2) elimination of the concept of the person provides a viable solution to the problem of exclusion for these neo-Kantian theories.

Gewirth argues that the very content of morality (and the establishment of certain rights) can be derived from the normative structure of action itself. All agents, insofar as they engage in free action, are implicitly committed to the evaluative and deontic judgments that are logically implicit in all of action. Accordingly a certain normative moral principle—indeed a *supreme* moral principle—logically follows from what Gewirth calls the generic features of action: voluntariness and purposiveness. In fact, any agent who rejects or denies this supreme moral principle does so on pain of logical contradiction. He cannot do so with rational consistency. The structure of Gewirth's argument can be put as follows: First, every agent makes an implicit claim about the goodness of his purpose and hence about the goodness of his freedom and the necessary means of action (including his well-being). Secondly, he is committed to a deontic claim that he has certain rights to freedom, well-being, etc.... Thirdly, since he must claim these rights for himself he must accept them for all agents.

The first step of the argument encompasses the rather Kantian notion that when an agent wills an end he also wills the means. For Gewirth, these are the generic goods of action. "If a rational agent is to claim any rights at all, could anything be a more urgent object of his claim than the necessary conditions of his engaging both in action in general and in successful action?"[88] The generic features of action yield what Gewirth refers to as generic rights: the right to freedom, i.e. non-coercion and non-interference from others, the right to well-being, i.e. goods necessary for "maintaining and obtaining" what one regards as goods. All other rights can be derived from these generic rights.

I trust that by now that I have succeeded in fulfilling the first modest goal for this section: showing that there is nothing in Gewirth's derivation of moral rights that requires the use of the concept of the person. It is now time for the second, less modest goal: demonstrating the *advantage* of eliminativism in solving Gewirth's exclusion problem. The problem of exclusion can be found in the case of human beings who do not qualify as rational agents as Gewirth defines them—and consequently do not qualify for rights. Although Gewirth offers proposals to deal with this objection, I think that eliminativism deals with them better.

Gewirth characterizes the generic rights of action as 'human' rights because they can be ascribed to all human beings because they are all "actual, prospective or potential agents." I previously looked at this conceptual strategy in §5.2 and suggested that it has shortcomings. In this case I think that Gewirth is wrong that *all* human beings are actual, prospective or potential agents. Some human beings may not be prospective or potential agents *at all*. This would be a problem because our intuitions suggest that these human beings are entitled to some moral rights even if they aren't agents. I think that an evaluation of Gewirth's argument will demonstrate that this is a problem and that his proposed solution is effectual in some cases but does not apply in others.

---

[88] Gewirth, p. 63.

Gewirth offers a specific argument to solve his exclusion problem. He calls it the principle of proportionality. Agents who exhibit a lower than normal degree of rational agency are entitled to lesser rights *proportional* to that degree. So children and young infants would receive fewer protections proportional to their rational nature. This accords nicely with our intuition that paternalism towards children is appropriate because they are incapable of identifying, expressing or protecting their own interests. This would seem to solve his exclusion problem.

One might think that Gewirth is then committed to the following corollary: agents who exhibit a higher than normal degree of rational agency are entitled to more rights proportional to that degree. This would be objectionable. But Gewirth claims that there is a relevant disanalogy between the original application of the proportionality principle and its alleged corollary. In other words, those with superior levels of intelligence and who engage in a more sophisticated kind of agency are not entitled to more significant rights. That's because there's an absolute standard with no degree as to who qualifies as a purposive agent. There are degrees in approaching that standard, according to Gewirth, but that's a different thing. His example is the voting age. Qualifying to vote is an absolute standard of being at least 18 years of age, although there are degrees of approaching that standard. The property of being at least 18 years of age admits of no degrees; the property of being close to 18 years of age *does* admit degrees because you can be closer or further away from the age of 18. Consequently, those who fall below the standard of rational agency may have degrees of not meeting the standard, but those who do meet the standard do so absolutely without any degrees. So superior agents don't get more rights than regular agents:

> The point is that once a person is an actual or prospective agent, he has the generic rights in full; but if he does not fully attain to the generic features and abilities of action, then he has the generic features and abilities of action, then has the generic rights in proportion to his degree of attainment of agency.[89]

And then later:

> Since for any prospective agent the point of having and respecting the generic rights derives

[89] Gewirth, p. 141.

> from their necessary contribution to action in pursuit of purpose fulfillment, persons who are not capable of such action have the right to freedom only to the extent to which their free conduct would not directly interfere with the conditions of their own freedom and well-being or with the generic rights of other persons.[90]

While Gewirth's account provides a plausible solution for the exclusion problem concerning children and other prospective agents, it does little to address those human beings who will never be agents in the future. I concede that Gewirth's answer works for agents with lower degrees of rational agency, say Alzheimer patients and children, but it does not resolve our ethical intuitions about living human beings who aren't agents now and won't be in the future. Gewirth's argument excludes the living human beings that I described in §5.2, e.g. patients who have no *higher* cognitive functions but who are not brain dead because they pass the apnea test (which is evidence of lower brain function). And denying all rights to these entities violates our ethical intuitions.

And so it is here where we can see the virtue of the eliminativist strategy. Having eliminated the concept of the person we no longer need to hang everything on whether something qualifies as a person or not. With a diversity of component concepts at our disposal we can evaluate which of the components is morally significant, in what way it is morally significant, and then attribute rights on that basis. For example, we might attribute some rights to agents—as Gewirth does—but attribute other rights to the biological concept of a human being. Still other rights might stem from the psychological capacity to feel pain—the utilitarian standard.

Gewirth has a response to this objection, but it is problematic. Gewirth has a formula for extending some rights not just to human beings who are no longer agents, but even non-human animals who were never agents. Here's how it goes: Gewirth contends that his attribution of generic rights on the basis of action accords with the utilitarian standard of feeling pain or pleasure, citing Bentham's question: Can they suffer? Gewirth contends that general utilitarian principles can be derived from the normative structure of action because

> Every rational agent, being also an animal, has various feelings, and these affect his pursuits

[90] Gewirth, p. 142.

> of his purposes and the physical background that makes those pursuits possible. The suffering of pain is a debilitating experience, and the PGC [principle of generic consistency] prohibits its wanton infliction on others as violating the rights both of freedom and of well-being.[91]

While this follows logically from his account so far, Gewirth goes on to suggest that because PGC prohibits the infliction of pain to persons it also prohibits the same towards animals, since animals are susceptible to pain to varying degrees. According to the Principle of Proportionality they should receive such rights proportional to their capacity to feel pain.

But this seems strange. The basis for the general utilitarian principle was not a utilitarian foundation but rather a deontic foundation derived from the generic features of rational and free agency. And animals do not meet Gewirth's standard for voluntary and purposeful agency. How can the utilitarian principles apply to a case when the underlying foundation is absent? If animals are not rational and free agents, they have no rights to freedom and well-being and the rights which derive from them. While Gewirth is correct to point out that all rational agents are also animals, it is certainly not the case that all human animals are agents; this is precisely the point of the exclusion problem. Gewirth goes on to qualify his position by stating that when the well-being of animals conflicts with those of humans, it is the rights of the latter that take priority because their rights are proportionally higher (according to the Principle of Proportionality). But it is unclear why animals should have any rights at all when the underlying facts which ground moral rights (rational agency) are absent. It seems as if Gewirth is trying to smuggle in some utilitarian considerations that would restrict our conduct towards biological entities and help solve the exclusion problem. That's fine, but he should not claim that these considerations are logically derivable from the generic features of rational action.

Gewirth's attempt to solve the exclusion problem appears to be based on a misguided application of the Argument from Above. What is important are the underlying facts and the facts in this case are that some animals do not meet the criteria for rational agency and therefore should not have generic rights derived from them according to Gewirth's own

[91] Gewirth, p. 144.

theory.[92] Consistency demands that we deny rights to human beings who are no longer agents (because they lack higher cognitive functions). And in the final analysis this exclusion should count against Gewirth's theory of rights. As I have argued, the cause of this predicament is the ascription of rights to persons, the identification of personhood with a single component of the cluster, thus entailing that beings without the essential component are not persons and don't have rights. Eliminativism avoids this unfortunate dilemma by blocking the first step of the argument: rights need not be ascribed to persons but might be ascribed to the underlying facts, the components. And different rights might stem from different components.

Rawls' notion of a moral person who is a participant in the original position is largely based on the notion of a rational agent. By putting all participants of the original position behind a veil of ignorance, Rawls ensures that the participants are ignorant of all contaminating factors which might allow them to tailor the social contract to their advantage. Not only is each participant fully rational, each participant is exactly the same: a rational agent *simpliciter*. Accordingly, the rational outcome of the bargain should be the same from each participant's point of view. One only need consider the negotiation from one agent's point of view because the points of view will turn out to be exactly similar. So there's a link—which will become more explicit in Gauthier—between Rawls' theory of justice with rational choice theory: "Understood in this way the question of justification is settled by working out a problem of deliberation: we have to ascertain which principles it would be rational to adopt given the contractual situation. This connects the theory of justice with the theory of rational choice."[93]

---

[92] I should be clear that I am not rejecting (though I am neither accepting) the derivation of rights from the generic features of action that Gewirth has offered. There is much to recommend this derivation. What I *am* claiming is that Gewirth's account, insofar as it appeals to an agent-centered account of personhood, does not cover the entire spectrum of beings that we consider to be valid sources of rights claims. I have diagnosed the source of this problem as a confusion about the concept of the person. Once that confusion is resolved, however, it may turn out that the rest of what Gewirth has to say about agents and their rights is largely correct.

[93] Rawls, *A Theory of Justice*, p. 17.

But what does Rawls mean when he talks about persons"? In *Justice as Fairness*, Rawls explicitly defines persons as having two moral powers: a capacity for a sense of justice and a capacity for a conception of the good, i.e. to *rationally* pursue a conception of the good in the form of projects or a life plan. Although Rawls goes to great lengths to distance this conception of personhood from any concept found in the philosophy of mind, metaphysics of psychology, he goes to fewer lengths to distance his notion of personhood from that of the rational agent found in rational choice theory, writing that

> This conception of the person is not to be mistaken for the conception of a human being (a member of the species Homo Sapiens) as the latter might be specified in biology or psychology without the use of normative concepts of various kinds, including, for example, the concepts of the moral powers and of the moral and political virtues. Moreover, to characterize the person, we must add to these concepts those used to formulate the powers of reason, inference and judgement. These are essential companion powers to the two moral powers and are required for their exercise and for the practice of the virtues.[94]

When Rawls here talks about adding the powers of reason, inference and judgment, he might as well be talking about rational agency instead of moral personhood. So it would seem as if little in Rawls' theory of justice (and consequently his derivation of basic rights agreed to by the parties of the original position) would be threatened by an elimination of the concept of the person. Since rational deliberation is the key concept in the original position, the concept of the agent will serve as an adequate replacement for the concept of the person.

Rawls' version of the social contract has come under scrutiny from a number of critics. Dworkin has argued, quite forcefully, that since hypothetical contracts have no binding force of their own, the role of the social contract in Rawls' account must be to bring to light additional reasons for the fairness of the outcome. In that sense, of course, the social contract becomes superfluous, according to Dworkin, because one might as well appeal directly to the additional reasons themselves—they, not the social contract, provide the justification for the fairness of the outcome. These additional reasons constitute a "deeper

---

[94] John Rawls, *Justice as Fairness: A Restatement*, edited by Erin Kelly (Cambridge: Harvard University Press, 2001), p. 24.

theory," deeper that is than the original position in the sense that the deeper theory is prior to—not the result of—the deliberation of the original position.[95]

Rawls has responded by clarifying that the original position is meant to *model* a deeper theory which might be collected under the term "justice as fairness." Specifically, he notes that it models what are "fair conditions under which the representatives of citizens, viewed solely as free and equal persons, are to agree to the fair terms of cooperation whereby the basic structure is to be regulated."[96] The position is the result of the deeper theory, not the other way around. Or more precisely, both are mutually enforcing through the process of reflective equilibrium.

Others influenced by Rawls have moved away from the theoretical structure of the original position and have offered alternate versions of the deeper theory. It might be instructive to see if these moral theories can survive elimination of the concept of the person and still provide a foundation for right claims. Scanlon's deeper theory centers around "reasonable rejectability." Although this notion is frequently used in connection with the concept of the person, we shall soon see that it is more specifically attached to the concept of rational agency. Scanlon argues that "morality applies to a being if the notion of justification to a being of that kind makes sense,"[97] a standard he cashes out in the following way: the being has a conception of the good and a point of view. In the absence of these conditions justification to a being would make little sense.

It is clear that both of these restrictions are agent centered. The first one was that *justification* to the being makes sense. This is another way of saying that we can provide reasons to a being which she cannot reasonably reject. And engaging in reason-giving practices is one of the hallmarks of rational agency. One offers justification to a deliberating rational agent who is capable of being *moved* by reasons. This is precisely the notion of

---

[95] Ronald Dworkin, "Justice and Rights" in *Taking Rights Seriously* (Cambridge: Harvard University Press, 1977).

[96] Rawls, *Justice as Fairness*, p. 17.

[97] T.M. Scanlon, "Contractualism and Utilitarianism" in *Ethical Theory*, edited by James Rachels (New York: Oxford University Press, 1998), p. 360.

agency that I identified earlier. The second constraint outlined by Scanlon—having a point of view—is also clearly agent-centered. It is agents who have a distinct point of view on the world from which they gather information, deliberate and reason. I have identified this, following Rovane, as the *rational* point of view.

Unfortunately, Scanlon is not working with a distinction between the rational and phenomenological points of view—a distinction which has been collapsed here. That's why he goes on to claim that

> contractualism can explain why the capacity to feel pain should have seemed to many to count in favor of moral status: a being which has this capacity seems also to satisfy the three conditions I have just mentioned as necessary for the idea of justification to it to make sense. If a being can feel pain, then it constitutes a centre of consciousness to which justification can be addressed. Feeling pain is a clear way in which the being can be worse off; having its pain alleviated a way in which it can be benefited; and these are forms of weal and woe which seem directly comparable to our own.[98]

But I think Scanlon is wrong here. It is wrong to say that a being who can feel pain constitutes a center of consciousness to which justification can be addressed, because only a confusion between the phenomenological and rational points of view would lead someone to say this. A being that can feel pain is identified by a unified phenomenology, but a being to whom justification can be addressed is a purely *rational* relation that need not—as I have argued earlier—coincide with a phenomenological center of consciousness.

This confusion between the phenomenological and rational points of view is one of the causes of the problem of exclusion. Remember, the whole point with the problem of exclusion is that not all beings who can feel pain are rational agents. A univalent theory tells us that only the latter have rights, but our intuitions tell us that the former should have rights too. Indeed, it is not even true that all rational agents are unified centers of consciousness. Group agents are the obvious counter-example here. (Though it is the case that all agents are endowed, at some level, with consciousness.) The two classes of beings do not directly overlap. Agents are the ones to whom justification is possible. But those who feel pain and who have a phenomenological center of consciousness are also entitled to moral

---

[98] Scanlon, p. 361.

concern. This becomes clear once you recognize that justification to an agent is a *rational* relation.

Eliminativism would help solve Scanlon's exclusion problem. We would be able to distinguish different sources of right claims. Under contractualism, a rational agent has rights because she is an appropriate subject of justification. For utilitarianism, the object of moral concern is a center of consciousness that can feel pain. Since these two things can diverge (either in thought experiments or marginal cases of everyday life), it is best to keep the sources of the rights-claims separate. That way we can correctly identify that some rights may stem from an agent's ability to be the subject of justification (such as political rights), while other rights might stem from a biological being's capacity to feel pain (i.e. the right not to be tortured).

We might be vulnerable to the following objection: every agent to whom we can attempt justification must be, at the very least, *composed* of phenomenological entities to whom pain can be attributed, so there is no relevant divergence between the two. All agents, even corporate agents, have the same raw materials—entities that can feel pain. But the objection misses the point. While it is true that all agents must be endowed, at some level, with consciousness and must therefore be subjects of painful or pleasurable experiences, the reverse is not necessarily true. Not all subjects of painful and pleasurable experience are agents to whom we can direct justification. Furthermore, the *identity* (in the sense of personal identity) of an agent may very well diverge from the identity of a biological entity with a unified consciousness and a phenomenological point of view. That's why the being that is capable of receiving justification (according to the theory) may not be the same being who is *owed* justification (according to our intuitions).

In Gauthier's version of contractarianism, the specter of rational agency is even stronger. The idea is that the social contract can be justified on purely rational grounds to each agent. In short, each agent has a reason to accept constraints on her behavior because doing so

leaves everyone better off. The model for this rational choice is, of course, the Prisoner's Dilemma, where the rational decision for each individual actor leads to a global results that is worse off for everybody. Cooperation would have improved every one's lot. Gauthier claims the same for the social contract:

> No one, of course, can have reason to accept any unilateral constraint on her maximizing behavior; each benefits from, and only from, the constraint accepted by her fellows. But if one benefits more from a constraint on others than one loses by being constrained oneself, one may have reason to accept a practice requiring everyone, including oneself, to exhibit such a constraint.[99]

Gauthier's account is an attempt to explicitly link social contract theory with rational choice theory—a process that began, latently, with Rawls. According to Gauthier, the constraints of morality are irrational if viewed from the point of view of a single rational agent. But, like the Prisoner's Dilemma, this leads to a result that is worse for all. When the constraints are looked at from the point of view of a collection of rational agents cooperating with each other, then the constraints are mutually beneficial and consequently rationally justified.

Gauthier's social contract theory is the natural conclusion of Rawls' programme of viewing the person as a rational, deliberating agent. In this case, the concept of the person is stripped down to its bare, deliberative essentials. And the rationality of the contracting parties can be modeled with the machinery provided by game theory and the Prisoner's Dilemma. Obviously, then, Gauthier's social contract theory is more wedded to the concept of the rational agent than it is with the concept of the person. More importantly, though, Gauthier's use of the concept of the rational agent—and its connection to rational choice theory—helps to bring into focus the fact that this entire line of moral theory from the beginning (back to Kant) has been a house built on the foundation of the concept of the agent, not persons.

One final point about the connection between the concept of agency and rights: In previous chapters I supported the claim that psychological continuity and phenomenological unity

---

[99] David Gauthier, "Why Contractarianism?" in *Ethical Theory*, edited by James Rachels (New York: Oxford University Press, 1998), p. 118.

were not necessary for rational agency. Consequently there could be agents composed of more than one human being and multiple agents within a single human being. If indeed it is the concept of the agent that is the foundation for moral rights, then a new problem has developed. Are group agents to be accorded the same rights as other agents? Should corporations have a right to life? Is dissolving a corporation a violation of its moral rights? Is it murder? This seems absurd and rightly so. This is the opposite of the exclusion problem. Instead of having a class of beings wrongly excluded from the moral community, this time we have a class of beings wrongly included in the moral community. Call this the problem of *inclusion.*

First of all, I have never claimed that corporations or multiple personalities are agents; this is an empirical claim. My only point, following Rovane, was that *if* these entities meet the criteria we established, then they should count as agents, even if they do not correspond one-to-one with biological human beings. So my claim is completely consistent with a state of affairs where no such entities exist and the problem of including them in the moral community does not arise in the first place. And if entities were to exist that met the criteria for group or multiple agency, this would be good evidence that a univalent, agent-centered moral theory is problematic since it fails to deliver answers that accord with our intuitions in marginal cases. In the case of group agents such as corporations it yields too many rights; in the case of human beings who aren't agents it yields too few rights. One solution to this problem would be to switch to a multivalent theory of moral rights and then we would not need to confer rights on the basis of a single aspect of personhood. If we do away with the concept of the person and replace it with its component concepts, we can attribute some rights to agents and some rights human beings. Or at the very least our conceptual machinery will allow us a sensible vocabulary with which this possibility might be explored.

Another solution to the problem of inclusion might be to take the alleged absurdity of giving rights to group and multiple agents as evidence that these entities are

metaphysically impossible. The fact that they would be entitled to moral rights is reason enough to deny their existence. While this would surely solve the problem of inclusion, there is much to recommend against it. First of all there are the myriad reasons brought forth in the previous chapters. Secondly, there is the fact that Gauthier's version of the social contract *supports* the interpretation of rational agency as built purely on rational relations, not phenomenological relations. The rationality of deliberation is everything. Phenomenology and psychological relations are nothing. Even in the context of moral theory, there is reason to support a definition of rational agency that supports at least the conceptual possibility of agent identity diverging from animal identity. So it would seem as if we might think twice before throwing out the idea of group and multiple agents.

I hasten to point out that there is another solution available. We could question the assumption that a moral theory of rights should be univalent and that the concept of agency is the single source for all rights. We could instead take this as evidence that a proper moral theory of rights should be multivalent. Some rights do indeed issue forth from the concept of agency but other rights issue forth from other components, viz. the biological human being and psychological continuity. At this point we have defended the claim advanced earlier in this chapter that we ought to recognize that there might be more than one conceptual source for moral rights. Now we have come to realize that different rights might come from different components. The content of moral rights might differ depending on the component concept in question. Different rights might come from different places.

§5.5 DISTRIBUTIVE JUSTICE WITHOUT PERSONS

Now to a different point about rights: The motivation for Rawls' social contract theory is at least partially dissatisfaction with utilitarianism. Rawls famously argued that utilitarianism ignores the *separateness* of persons. When maximizing either overall or average utility, utilitarians may distribute an unequal amount of benefits to some members while giving others an unequal share of burdens. This hardly seems fair to the least advantaged

because *they* cannot be compensated by benefits given to *others*. They are separate beings whose moral status makes demands on us. Hence the need for distributive justice. In this section I want to examine and respond to two objections that eliminating the concept of the person will cause irreparable harm to distributive justice. This will lead to a new objection that agency is the only relevant component for moral rights and that eliminativism is therefore unwarranted. I will respond to this objection by offering an example showing that some rights require us to separate out the components of the cluster. Lastly, I will show that eliminativism might help distributive justice even if some rights presuppose that the components stay linked together. In short, just so long as this does not hold true for all rights, the advantages of eliminativism will be vindicated.

This brings us to our first objection. Since distributive justice requires a just distribution of goods over a lifetime, how will we calculate this distribution without the concept of the person? A coherent concept of personal identity is required so that we can properly re-identify one person over time. Without such a theory it is impossible to determine an equitable distribution of goods. After all, the temporal boundaries of a single person—where one person ends and another begins—is crucial for distributive justice. If our theory of personal identity tells us when the life of a person ends and a new life begins, then the burdens of the former cannot be compensated by benefits to the latter. That's precisely why we need a theory of personal identity. And if you eliminate the concept of the *person*, you also eliminate *personal identity*.

But remember, we stated in chapter two that eliminating the concept of the person does not entail giving up a theory of identity over time. Rather, the identity criterion of personal identity is transformed into three corresponding identity criteria for each of the components that were pulled from the cluster. So in principle there should be no problem; just because we have eliminated *personal* identity does not mean that there are not corresponding accounts of identity over time for the remaining component concepts. We

can still identify particular agents over time, biological human beings over time and phenomenological continuants over time.

For example, the moral theorist working on distributive justice might be particularly interested in the component concept of the agent. This theorist might also be anxious that the concept of the agent is ill-suited for an account of identity over time, similar to personal identity, that would serve as an adequate basis for distributive justice. The theorist might conceive of an agent as, say, the initiator of a single action performed at a single moment in time. At first glance it might seem as if an account of agent identity over time is not forthcoming because actions are performed at discrete moments in time. Where is the temporal element that might yield an account of agent identity over time?

It is important to recall the account of agency presented in chapter three. I pointed out that the existence of an agent's rational life plan would serve quite well for an account of agent identity. And it is most important to realize that actions are not just discrete and temporally localized events. When we talk about agents performing actions at discrete points in time, it is important to remember that the individual actions stand in certain *rational* relations. One action, once completed, leads to a second action. Together they form the basis for a project or plan—extended in time—thus allowing the agent to prioritize opportunities in the face of an ultimate, future goal. Projects that are related together form the basis for an overall life plan. So in other words we are working with a very rich notion of a rational agent as someone who can pool information, deliberate and implement a project, plan or response. Together, these are the features that allow for the development of a life plan. This is fecund material for an account of agent identity over time which might provide a suitable backdrop for distributive justice.

A second objection might be formed now. Independent of the issue of identity over time, there might be something essential about distributing to *persons*. After all, Rawls complained that utilitarianism ignores the separateness of persons, not the separateness of other entities. After eliminativism we have no persons to distribute to and consequently no

persons with which to make this criticism about utilitarianism. But we should determine the degree to which the concept of the person is essential for Rawls' criticism of utilitarianism and his account of distributive justice. Although Rawls himself formulates his account in terms of persons, he also uses the concept of agency, as I explored in §5.4. But for Rawls, of course, the concept of the agent cannot be separated from the concept of the human being. That's exactly what a person is, according to Rawls. However, it seems to me that this commitment—that human beings and rational agents ought to be subsumed under the rubric of personhood—actually stands in some tension with Rawls' other commitments about the rationality of the parties involved in the original position. Let me explain.

Rawls freely points out that his account of the rationality of the parties in the original position is borrowed straight from rational choice theory, as I discussed in §5.4. That being the case, it is unclear why this notion of the rational agent cannot be separated from the concept of the biological human being, since the latter plays no role in the deliberative structure of rational choice theory. So although Rawls limits the original position to biological human beings, it is unclear if this assumption is absolutely essential to his formulation of the original position, since it is the structure of rational agency—independent of biology—that is relevant for deliberation in the original position.

One reason Rawls might feel compelled to limit the original position to biological human beings is that his "conception of the person itself is meant as both normative and political, not metaphysical or psychological."[100] This signals that Rawls' conception of personhood is not *just* influenced by the deliberative demands of the original position—it is also influenced by our moral intuitions that the parties to the social contract are also the *recipients* of its moral rights. Indeed, Rawls makes clear in *A Theory of Justice* the similarity between the veil of ignorance and Kant's categorical imperative. And the similarities are deep. Like in Kant, where those capable of legislating in accordance with the categorical imperative are also those eligible for the moral respect afforded by the imperative, so too in

[100] Rawls, *Justice as Fairness*, p. 19.

Rawls the conception of the person is doing double duty: both as the negotiators of the social contract and its beneficiaries. That is why Rawls speaks of citizens as being free and equal persons; it is obvious that, for Rawls, the concepts of citizen and person are closely linked since both are moral and political.

But what happens if we step back from the Kantian assumption, look at the deliberative structure of the original position and stop asking our concept of the person to fulfill this double duty? When this is done, it seems at least plausible to argue that all agents, even if they don't correspond one-to-one to biological human beings, should be part of the original position. The justification for this would be that we are no longer considering whether they are appropriate objects of moral respect (that is to presume the *result* of the original position, after all), just whether they are capable of the deliberation required for the original position. Perhaps we could codify this sentiment by changing the parameters of the original position such that the biological boundaries of agents negotiating in the original position should remain behind the veil of ignorance. In other words, agents involved in social contract negotiations would not know if they correspond to a single biological human being, are composed of several biological human beings (viz. a corporation), or share a biological human being with other agents (viz. multiple personalities). All of this adds up to the following possibility: we might detach our biological notion of the person from our concept of the rational agent and let the bare notion of the rational agent form the basis for the deliberations of the original position. So we can conclude that only the notion of the deliberating agent is essential for the original position; the notion of biological human beings is inessential.

Now a different objection emerges. One might argue that since the concept of the rational agent is doing all of the work in the original position, this is the only component of the cluster relevant for rights. If this were the case, we might have good reason to keep the concept of the person at the center of rights-based discourse and adopt a more agent-centered definition of persons. If this were the case, the concept of the person might not be a

cluster concept after all; rather, persons might be agents first and foremost. Although this view seems possible, there are several factors counting against it. Consider a distribution of benefits to a community of agents. Although the distribution to various agents is equitable and fair, what if it turned out that a disproportionate number of painful experiences were distributed to some phenomenological subjects in exchange for a radically lower number of painful experiences to a second group of phenomenological subjects? Although the distribution of benefits to agents would be fair, this scheme seems radically unjust to us. And that is because when it comes to distributive justice, agency is not the only relevant component of the cluster.

So much for my claim that we can separate out the biological component in Rawls' account of distributive justice. Now I want to suggest a few reasons why this might be better than simply sticking with a unified notion of personhood where the components always go together. There might be certain advantages in pursuing distributive justice with the components instead of with the concept of the person. This might allow us to appropriately distribute different kinds of rights. In general, Rawls' two principles of justice are the equality of basic rights and liberties and the distribution of socio-economic benefits (in accordance with the difference principle). It may very well be the case that the liberties track the concept of the agent (because it is agents that can make use of liberties) while economic distributions may track the concept of biological human beings. The details must be worked out explicitly and it may turn out to be a more complicated story. But in general, here is how the breakdown might work: rights to act in a certain way should be attributed to agents who can exercise them, including group and multiple agents, while rights to certain kinds of treatment should be attributed to biological entities such as human beings (i.e. the *recipients* of action).

An example will help illustrate the alleged benefits of pursuing distributive justice with the components of the cluster instead of with the concept of the person directly. Take future directives. Consider a case where an agent has strong reasons for a particular attitude

and wants us to respect his wishes on the matter. Not only does he want us to respect his wishes now, he also wants us to respect his present wishes in the future. The wish represents a commitment that stands at the very core of his personality. So he formulates a directive stating that if he appears to change his mind in the future, he is to be considered insane. He wants to be committed to a mental institution or medicated. Now suppose that through some severe mental or physical process—say amnesia—the rational commitments of this agent change. In minor cases one might think that the agent has simply changed his mind, but in severe cases like this our theory of agent identity suggests that a new agent has emerged. The amnesia is so severe that there is no longer a commitment to a unified rational point of view that encompasses the pre- and post-amnesia agents. So according to our theory of agent identity, we have a *different* agent now, even though he belongs to the same biological human life as his predecessor. Now here is the interesting consequence. If we are to follow the desires of the new agent we might break our promise to the original agent—who might still have a valid moral right. According to distributive justice, we have to be careful about balancing the burdens of one with the benefits of another.

Here is a case where the components of the cluster concept come apart in a morally relevant way. We have two agents, each with moral rights, who belong to the life of the same biological human being. If we imagine the first agent, it seems clear that he has at least *some* moral rights. And any conceptual landscape that prevents us from even *examining* those rights would be a real detriment for the moral theorist. A proper analysis of the tricky moral issues involved is best achieved by keeping the components separate. Working with a unified notion of personhood where the components stay together, such as a naturalist conception of persons as human beings, would only obscure the moral issues here: the rights of the original agent who formulated the directive.

One final objection must be considered now. Some moral theorists might suggest that the concept of the person is essential for rights-based discourse, not because persons are parties to the original position, but because the rights themselves presuppose that the

components of the cluster are linked. Consider, for example, the notion of a parental right. Not only are we talking about the right of an agent to act in a certain way—making child-care decisions for their offspring—but the right is predicated on a biological relationship between a parent and their offspring. In this case, the right only makes sense if the components are linked together. If the two come apart—the agent does not stand in the proper biological relation—then the right itself begins to dissolve.

But is this an argument for retaining the concept of the person for our rights-based discourse? It is not obviously so, because eliminating the concept of the person in favor of its components is not a denial of the deep link between the components. No one would deny this link. It is simply a recognition that they *sometimes* come apart. In those cases the capacity to separate out the components is essential for a proper analysis of the moral consequences. The objection only works if all rights are like parental rights and require that the components stick together. But as shown with the previous example of future directives, this is clearly not the case. Presumably there might be more rights that follow the same structure and require the separation of the component concepts. It is also important to remember, though, that in cases where the components stay together—as in the case of parental rights—there is nothing in the eliminativist position that prevents us from charting the axiological consequences of the components staying together. The virtue of the eliminativist position is that it gives the moral theorist the advantage in cases where the components come apart, yet does not hamper the moral theorist when the components stay together.

In this section I have attempted to resolve an anxiety about eliminating the concept of the person. In addition to its role as a central concept in rights-based discourse generally, the concept of the person also holds a central place in distributive justice specifically. It would very much count against the eliminativist proposal if it were shown to prevent—or even hamper—distributive justice. But a consideration of the original position shows that what is inessential, i.e. the biological and phenomenological components, might be

separated out from what is essential, i.e. the agency component; there is nothing inherently person-dependent about bargaining from the original position. More importantly, though, the content of at least some rights indicates that their analysis might even require the separation of the components. Although Rawls wrote that utilitarianism ignores the separateness of persons, he might as well have said that it also ignores the separateness of agents, the separateness of biological human beings and the separateness of phenomenological continuants. And distributive justice might still rectify that shortcoming without the concept of the person.

§5.6 UTILITARIAN AND RELIGIOUS THEORIES OF RIGHTS

So far our examination of the effects of eliminativism on moral theory has concentrated exclusively on deontic theories of rights. It would be wrong to ignore consequentialist moral theories, however, since they represent a huge chunk of the theoretical spectrum. If eliminativism were to prove incompatible with utilitarianism, this would certainly be relevant in evaluating the eliminativist strategy. The reverse is also true, however. If eliminativism were to offer some conceptual clarity to utilitarian moral theories, this too would be relevant.

We should divide up the utilitarian camp into direct and indirect utilitarians. Indirect utilitarians (such as rule utilitarians) might use a utilitarian calculation as a foundation for a system of moral rights. The idea here would be that it would increase overall utility to recognize a system of moral rights. Consequently, it might seem as if the eliminativist strategy would pose a problem for this kind of utilitarianism. To whom would we attribute these rights if we can't attribute them to persons? But the answers provided in the previous sections apply just as much here. The component concepts previously identified—biological human beings, their psychological continuants and rational agents—might serve as adequate sources for rights claims. The only difference would be that the

ascriptions would be second-order and would be justified on the ground level by utilitarian considerations.

As for direct utilitarians, they have no notion of moral rights. Persons don't have any *stable rights* at all because direct utilitarians demand *every* action maximize utility; this demand for near-constant calculation is incompatible with the stability and fixity presupposed by a moral right. Direct utilitarians demand that rights be violated in every instance where doing so maximizes utility; this prescription is tantamount to denying the existence of moral rights in the first place. However, utilitarianism must appeal to some object whose feelings should be part of the utility calculation. Here is one avenue were eliminativism may improve the situation. For some utilitarians, the appropriate object of moral concern is persons—however that is defined—and other beings who can feel pain, such as animals, are not included. For other utilitarians, the appropriate object of moral concern is *any* being capable of feeling pain. That includes animals, according to theorists such as Peter Singer, and excludes human beings such as newborn infants who are severely disabled. The rationale behind this standard is Bentham's question: "Can they feel pain?" These are some of the most contentious debates in utilitarian theory today. Perhaps eliminating the concept of the person and dealing instead of the components of the cluster will provide a better structure through which this debate might continue.

Let us go back to the roots of utilitarianism for a minute. Utility was originally cashed out to mean happiness—and happiness is an explicitly psychological attribute. So even after the elimination of the concept of the person, direct utilitarianism would not be impoverished. We can still maximize the happiness of biological human beings and their psychological continuants. If the correct moral theory is to maximize the utility—cashed out in terms of the psychology of happiness—of human beings, elimination of the concept of the person poses a problem for it. If utility is cashed out differently, say in terms of *interests*, then the component concept being utilized is that of the agent. We can maximize the utility

of agents by giving them the necessary goods to fulfill their projects and plans. This would be a consequentialist version of Gewirth's moral theory.

What is the advantage of eliminativism to utilitarianism? Precisely that we can separate out the different concepts so that it is clear whose interests we are maximizing. It makes a big difference whether we should maximize the utility of every agent or whether we should maximize the happiness of biological animals. The two do not coincide. There are different ways to investigate this ethical controversy. The wrong way is to determine who should count as persons and then allow those entities to figure in the utility calculation. Remember the Argument from Below: it is the lower level facts which matter. Instead of trying to decide who should count as a moral person, we should look at the components and determine which of them are morally significant to the utilitarian and why. Then on that basis beings can be included or excluded from the utilitarian calculus.

A third camp—neither deontic nor consequentialist—includes religious doctrines that attribute rights to persons on some theological grounds. These theories, while rare in contemporary moral philosophy, are popular in the public debate of controversial subjects such as abortion and euthanasia. These theorists often want to attribute rights to persons—such as the right to life—including fetuses or hospital patients with severe brain damage. These theorists often find themselves in conflict with utilitarians such as Peter Singer. At the heart of their arguments is that these persons have a *right* to life because they are God's children and only He has the power to decide when they should die. As such, society has a *religious* duty to protect human life whenever possible because all persons have a basic right to life and dignity. Anything less is an affront to God.

These religious theories of rights can do just fine without the concept of the person. The locus of moral concern here is a suitably-endowed biological entity created by God. By suitably endowed I do *not* mean the psychological and cognitive capacities, but rather the *soul*. Biological animals with souls (i.e. humans) have a right to life because their souls are divine and will live with God for eternity. By virtue of this fact they deserve moral respect

and they have moral rights. Although I am not committed to the existence of the soul or the dualism this theory presupposes, it is important to point out that nothing in the eliminativist strategy would derail a theological theory of rights. After eliminating the concept of the person we can still talk about biological human beings and their psychological continuants—in this case, according to religion, their *souls*.

§5.7 HUMAN RIGHTS WITHOUT PERSONS

There is a final concern that I feel compelled to address. In various quarters—academic, political and international—there has been a growing appeal to the concept of human rights. There is a growing body of human rights scholarship and a growing collective of international organizations that use the nomenclature of human rights to press their moral claims. It would provoke severe anxiety among some if eliminativism were to prove fatal to these doctrines of human rights. They might shy away from the metaphysical prescription if eliminating the concept of the person would deprive them of the necessary tools to rally for universal rights; the pill would be too bitter to swallow. Our moral and political rights can only be secured by gathering under a banner upon which the dignity of *persons* is inscribed. Few will respond to political and moral rhetoric about "biological human beings, psychological continuants and rational agents." It's just too messy.

The first thing we must do is get clear about the different ways the word 'human' can be used. Under one usage, the term 'human' is a near synonym for person and is used interchangeably. When this happens, 'human' is just as much a cluster concept as personhood and the eliminativist strategy applies to it just as much. That's because the concept attempts to house, under one conceptual umbrella, diverse notions such as the biological concept *Homo sapiens*, psychological attributes, and considerations of rational agency. If the term 'human' is used in a context of providing justification for human rights, often one of these components is singled out as the ground for the moral claim. Under a second usage, the term 'human' is not meant to refer to a cluster concept such as personhood

but is meant to refer only to the specifically biological notion of the human being, i.e. *Homo sapiens*. In this case 'human' is *not* a synonym for 'person' at all and so it would seem as if eliminativism is no longer relevant. But while eliminativism does not directly apply in this case, the spirit of the strategy is still applicable. Even when the term 'human' is restricted to its biological interpretation, ethicists and human rights theorists must still determine who will count as a human being. So with either sense of the word 'human' the eliminativist strategy has an application to human rights theory. That's why I feel compelled to address the concern that human rights theory would be harmed by eliminativism.

The first response to this concern is that it is not really the burden of my project to demonstrate that political rhetoric can continue *easily* without the concept of personhood. I only need to demonstrate that it *can* continue *simpliciter*. But even if it were my burden, the burden could be met: we still have at our disposal the components of the cluster including the biological concept of human beings. The biological human being may be entitled to certain dignities and basic rights that could be called "human rights."

Of course, we would still need to define what a human being is in order for the concept of human rights to have an application. I shall not attempt to do this here. However, I think this is yet another situation where the virtues of eliminativism are clear. We can still talk about biological human beings, their psychological continuants and rational agents, and eliminativism allows us to separate these concepts. This will go a long way to offering a more nuanced understanding of the scope of human rights. Indeed, part of human rights scholarship is devoted to studying the *scope* of human rights, i.e. who should be included in the class of human beings. These right claims become difficult to adjudicate when considering biological human beings who lack certain attributes. (Do fetuses and the mentally disabled deserve human rights?) Consider our hospital patient described earlier in the chapter. Although he has no higher cognitive functions, he still has enough lower brain function to keep him breathing and to pass the apnea test. Although there is no

psychological continuity here, a functioning human being remains with lungs that breathe and a heart that pumps.

As was pointed out in the beginning of this chapter, the avenue used in human rights scholarship to adjudicate these human rights claims has often been to see if these individuals qualify as persons by an investigation into metaphysics or if they qualify as human beings by an investigation in the philosophy of biology. This then supposedly entails various claims about rights because rights flow from personhood and human-beinghood. The rights game is all or nothing. But eliminativism has the power to change this procedure for the better; our analysis has demonstrated the importance of being faithful to the Argument from Below. It is the underlying facts of the case which are significant. We should look at those facts and determine—directly—which moral consequences, in the form of rights, emanate from them. We can confer rights based on which components are present. Instead of engaging in metaphysics first and value theory second, we can take the facts of the matter already provided and go straight to moral theory. Remember, the Argument from Below suggested that it is the underlying facts which are significant, not how we describe those facts. Once we know the underlying facts of the matter, it is insignificant to ask if these facts are best described with the word 'person'. It would be better to evaluate those underlying facts and ask ourselves what morality demands of us when faced with those facts. An individual may be entitled to some human rights because he is a functioning human being and entitled to other rights because he is a rational agent. This accords well with our ethical intuitions that the complete denial of rights to human beings who are no longer agents is ethically problematic.

The end result is not just a better method of investigation for human rights scholarship, it might also help them deal with the problem of exclusion we well. In previous sections I suggested that eliminating the concept of the person would help facilitate a move away from univalent moral theories and towards a multivalent moral theory. Eliminating the concept of the person encourages this process by removing the cluster concept in favor of

components which may be distinct sources for moral claims. Finally, recognizing different sources for moral claims might lead to the realization that the rights game need not be all or nothing. Different rights might come from different places. This realization has the power to solve the problem of exclusion by attributing some rights—but not others—to marginal entities who have some of the components but not others. This would have the effect of reversing the method of investigation in human rights scholarship. Instead of trying to decide what counts as a human being or a person in order to see if they are legitimate bearers of moral rights (when in fact nothing hinges on this point), we could instead focus our attention on investigating which underlying facts are morally significant and why. We can then decide which rights should be matched up with those facts.

Bibliography

ADKINS, A.W.H. *Merit and Responsibility: A Study in Greek Values.* New York: Clarendon Press, 1960.

AYER, A.J. "Free-Will and Rationality." In *Philosophical Subjects: Essays Presented to P.F. Strawson*, edited by Zak Van Straaten. Oxford: Clarendon Press, 1980.

AYER, A.J. "The Concept of a Person." In *The Concept of a Person and Other Essays.* New York: St. Martin's Press, 1963.

BENNETT, JONATHAN. "Accountability." In *Philosophical Subjects: Essays Presented to P.F. Strawson*, edited by Zak Van Straaten. Oxford: Clarendon Press, 1980.

BLACKBURN, SIMON. "Has Kant Refuted Parfit?" In *Reading Parfit*, edited by Jonathan Dancy. Oxford: Blackwell Publishers, 1997.

BRADLEY, F.H. "The Vulgar Notion of Responsibility in Connexion with the Theories of Free-Will and Necessity." In *Ethical Studies*, pp. 1-57.

BROWN, JAMES ROBERT. "Thought Experimetns since the Scientific Revolution." In *International Studies in the Philosophy of Science* 1 (1986): 1-15.

BUTLER, JOSEPH. "Of Personal Identity." In *Personal Identity*, edited by John Perry. Berkeley: University of California Press, 1975.

CHISHOLM, RODERICK M. "Identity Through Time" and "Reply to Strawson's Comments." In *Language, Belief and Metaphysics*, edited by Howard Kiefer and Milton Munitz. Albany: State University of New York Press, 1970.

DANCY, JONATHAN, ED. *Reading Parfit.* Oxford: Blackwell Publishers, 1997.

DAVIDSON, DONALD. "Agency." In *Essays on Actions & Events.* New York: Oxford University Press, 1980.

DAVIDSON, DONALD. "Deception and Division." In *The Multiple Self*, edited by Jon Elster. New York: Cambridge University Press, 1986.

DAVIDSON, DONALD. *Essays on Actions & Events.* New York: Oxford University Press, 1980.

DAVIDSON, DONALD. "Freedom to Act." In *Essays on Actions & Events.* New York: Oxford University Press, 1980.

DAVIDSON, DONALD. "How is Weakness of the Will Possible?" In *Essays on Actions & Events.* New York: Oxford University Press, 1980.

DENNET, DANIEL. "Mechanism and Responsibility." In *Free Will*, edited by Gary Watson. New York: Oxford University Press, 1982.

DENNETT, DANIEL C. *Brainstorms: Philosophical Essays on Mind and Psychology.* Cambridge: MIT Press, 1981.

DWORKIN, RONALD. *Taking Rights Seriously.* Cambridge: Harvard University Press, 1977.

ELSTER, JON. *The Multiple Self.* New York: Cambridge University Press, 1986.

ELSTER, JON. *Ulysses and the Sirens: Studies in Rationality and Irrationality.* New York: Cambridge University Press, 1979.

EVANS, GARETH. *The Varieties of Reference.* New York: Oxford University Press, 1982.

FEINBERG, JOEL. *Doing & Deserving: Essays in the Theory of Responsibility.* Princeton: Prineton University Press, 1970.

FRANKFURT, HARRY G. *The Importance of What We Care About: Philosophical Essays.* New York: Cambridge University Press, 1988.

FRANKFURT, HARRY G. "Freedom of the Will and the Concept of a Person." In *The Journal of Philosophy* 68 (1971): 5-20.

GAIUS. *The Institutes.* Translated by W.M. Gordon and O.F. Robinson. Ithaca: Cornell University Press, 1988.

GAUTHIER, DAVID. *Morals by Agreement.* New York: Oxford University Press, 1986.

GAUTHIER, DAVID. "Why Contractarianism?" In *Ethical Theory,* edited by James Rachels. New York: Oxford University Press, 1998.

GEWIRTH, ALAN. *Reason and Morality.* Chicago: The University of Chicago Press, 1978.

GLOVER, JONATHAN. *Responsibility.* New York: Humanities Press, 1970.

HART, H.L.A. "The Ascription of Responsibility and Rights." In *Proceedings of the Aristotelian Society* (1948-9).

HART, H.L.A. and TONY HONORÉ. *Causation in the Law.* New York: Oxford University Press, 1959.

HART, H.L.A. *Punishment and Responsibility.* New York: Oxford University Press, 1968.

HEMPEL, CARL G. *Aspects of Scientific Explanation: And Other Essays in the Philosophy of Science.* New York: The Free Press, 1965.

HURLEY, SUSAN L. *Natural Reasons: Personality and Polity.* New York: Oxford University Press, 1989.

JAMES, WILLIAM. *The Pricniples of Psychology.* New York: Dover Publications, 1950.

JOHNSTON, MARK. "Human Concerns without Superlative Selves." In *Reading Parfit,* edited by Jonathan Dancy. Oxford: Blackwell Publishers, 1997.

KANT, IMMANUEL. *Fundamental Principles of the Metaphysic of Morals* in *Great Books of the Western World.* Chicago: Encyclopaedia Britannica, 1952.

KANT, IMMANUEL. *Grounding for the Metaphysics of Morals,* translated by James W. Ellington. Indianapolis: Hackett Pubishing Co., 1993. First published in 1785.

KITCHER, PATRICIA. *Kant's Transcendental Psychology.* New York: Oxford University Press, 1990.

KITCHER, PATRICIA. "Natural Kinds and Unnatural Persons." In *Philosophy* 54 (1979): 541-47.

KORSGAARD, CHRISTINE M. "Personal Identity and the Unity of Agency: A Kantian Response to Parfit." In *Philosophy and Public Affairs* 18 (1989): 101-132.

KUHN, T.S. "A Function for Thought Experiments." In *The Essential Tension.* Chicago: University of Chicago Press, 1977.

LEWIS, DAVID K. "Survival and Identity." In *The Identities of Persons,* edited by Amelie Rorty. Berkeley: University of California Press, 1975.

LEWIS, DAVID K. *On the Plurality of Worlds.* New York: Basil Blackwell, 1986.

LOCKE, JOHN. *An Essay Conerning Human Understanding.* Edited by Alexander Campbell Fraser. New York: Dover Publications, 1959.

LOMASKY, LOREN E. *Persons, Rights, and the Moral Community.* New York: Oxford University Press, 1987.

LUCAS, J.R. *Responsibility.* New York: Oxford University Press, 1993.

MADELL, GEOFFREY. *The Identity of the Self.* Edinburgh: The University Press, 1981.

MARTIN, RAYMOND. *Self-Concern: An Experiential Approach to What Matters in Survival.* New York: Cambridge University Press, 1998.

MCDOWELL, JOHN. "Reductionism and the First Person." In *Reading Parfit,* edited by Jonathan Dancy. Oxford: Blackwell Publishers, 1997.

MCDOWELL, JOHN. *Mind and World.* Cambridge: Harvard University Press, 1994.

NAGEL, THOMAS. "Moral Luck" in *Free Will,* edited by Gary Watson. New York: Oxford University Press, 1982.

NAGEL, THOMAS. "Brain Bisection and the Unity of Consciousness." In *Mortal Questions*: New York: Cambridge University Press, 1979.

NAGEL, THOMAS. *The View from Nowhere*. New York: Oxford University Press, 1986.

NOONAN, HAROLD. "The Closest Continuer Theory of Identity." In *Inquiry* 28 (1985): 195-229.

NOONAN, HAROLD. *Personal Identity*. Brookfield, Vermont: Dartmouth Publishing Co., 1993.

NOONAN, HAROLD. "Wiggins' Second Thoughts on Identity." In *Philosophical Quarterly* 31 (1981): 260-8.

NOZICK, ROBERT. *Philosophical Explanations*. Cambridge: Harvard University Press, 1981.

OLSON, ERIK T. *The Human Animal: Personal Identity Without Psychology*. New York: Oxford University Press, 1997.

PARFIT, DEREK. "Comments." In *Ethics* 96 (1986): 832-872.

PARFIT, DEREK. "Later Selves and Moral Principles." In *Philosophy and Personal Relations*, edited by Alan Montefiore. London: Routledge and Kegan Paul, 1973.

PARFIT, DEREK. "On 'The Importance of Self-Identity.' " In *The Journal of Philosophy* 68 (1971): 683-90.

PARFIT, DEREK. *Reasons and Persons*. New York: Oxford University Press, 1984.

PENELHUM, TERENCE. "The Importance of Self-Identity." In *The Journal of Philosophy* 68 (1971): 667-78.

PERRY, JOHN, ED. *Personal Identity*. Berkeley: University of California Press, 1975.

PETTIT, PHILIP AND JOHN BRAITHWAITE. *Not Just Deserts: A Republican Theory of Criminal Justice*. New York: Oxford University Press, 1990.

PETTIT, PHILIP and SMITH, MICHAEL. "Freedom in Belief and Desire." In *The Journal of Philosophy* 93 (1996): 429-49.

PETTIT, PHILIP. *The Common Mind: An Essay on Psychology, Society and Politics*. New York: Oxford University Press, 1993.

QUINE, W.V. Review of *Identity and Individuation*, edited by Milton K. Munitz. In *The Journal of Philosophy* 69 (1972): 488-97.

RACHELS, JAMES. *Ethical Theory*. New York: Oxford University Press, 1998.

RAWLS, JOHN. *A Theory of Justice*. Cambridge: Harvard University Press, 1971.

RAWLS, JOHN. *Justice as Fairness: A Restatement*, edited by Erin Kelly. Cambridge: Harvard University Press, 2001.

RORTY, AMÉLIE OKSENBERG, ED. *The Identities of Persons*. Berkeley: University of California Press, 1976.

RORTY, AMÉLIE OKSENBERG. "A Literary Postscript: Characters, Persons, Selves, Individuals." In *The Identities of Persons*. Berkeley: University of California Press, 1976.

ROVANE, CAROL. "Branching Self-Consciousness." In *The Philosophical Review* 99 (1990): 355-95.

ROVANE, CAROL. "Rationality and Identity." In *The Philosophy of Donald Davidson*, edited by Lewis Edwin Hahn. Chicago: Open Court, 1999.

ROVANE, CAROL. "Self-Reference: The Radicalization of Locke." In *The Journal of Philosophy* 90 (1993): 73-97.

ROVANE, CAROL. *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton: Princeton University Press, 1998.

SCANLON, T.M. "Contractualism and Utilitarianism." In *Ethical Theory*, edited by James Rachels. New York: Oxford University Press, 1998.

SCHEFFLER, SAMUEL. "Ethics, Personal Identity, and Ideals of the Person." In *Canadian Journal of Philosophy* 12 (1982): 229-46.

SCHLOSSBERGER, EUGENE. *Moral Responsibility and Persons.* Philadelphia: Temple University Press, 1992.

SCHULTZ, BART. "Persons, Selves, and Utilitarianism." In *Ethics* 96 (1986): 721-45.

SCRUTON, ROGER. "Corporate Persons." In *Proceedings of the Aristotlean Society* 63 (1989): 239-66.

SHOEMAKER, SYDNEY, AND RICHARD SWINBURNE, EDS. *Personal Identity.* Oxford: Basil Blackwell, 1984.

SHOEMAKER, SYDNEY. "Parfit on Identity." In *Reading Parfit,* edited by Jonathan Dancy. Oxford: Blackwell Publishers, 1997.

SHOEMAKER, SYDNEY. "Wiggins on Identity." In *Identity and Individuation,* edited by Milton K. Munitz. New York: New York University Press, 1971.

SHOEMAKER, SYDNEY. "Persons and Their Pasts." In *American Philosophical Quarterly* 7 (1970): 269-85. Also reprinted in *Personal Identity.*

SHOEMAKER, SYDNEY. *Self-Knowledge and Self-Identity.* Ithaca: Cornell University Press, 1963.

SORENSEN, ROY A. *Thought Experiments.* New York: Oxford University Press, 1992.

STRAWSON, GALEN. *Freedom and Belief.* New York: Oxford University Press, 1986.

STRAWSON, P.F. "Chisholm on Identity Through Time (A Response)." In *Language, Belief and Metaphysics,* edited by Howard Kiefer and Milton Munitz. Albany: State University of New York Press, 1970.

STRAWSON, P.F. "Freedom and Resentment." In *Free Will,* edited by Gary Watson. New York: Oxford University Press, 1982.

STRAWSON, P.F. "P.F. Strawson Replies." In *Philosophical Subjects: Essays Presented to P.F. Strawson,* edited by Zak Van Straaten. Oxford: Clarendon Press, 1980.

STRAWSON, P.F. *Individuals: An Essay in Descriptive Metaphysics.* New York: Routledge, 1959.

STRAWSON, P.F. *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason.* London: Methuen, 1966.

SWINBURNE, RICHARD. "Personal Identity: the Dualist Theory." In *Personal Identity,* edited by Shoemaker and Swinburne. Oxford: Basil Blackwell, 1984.

TAYLOR, CHARLES. "Responsibility for Self." In *Free Will,* edited by Gary Watson. New York: Oxford University Press, 1982.

TAYLOR, CHARLES. *Sources of the Self: The Making of Modern Identity.* Cambridge: Harvard University Press, 1989.

THOMSON, JUDITH JARVIS. "Persons and their Bodies." In *Reading Parfit,* edited by Jonathan Dancy. Oxford: Blackwell Publishers, 1997.

THOMSON, JUDITH JARVIS. "Ruminations on an Account of Personal Identity." In *On Being and Saying: Essays for Richard Cartwright.*

UNGER, PETER. *Identity, Consciousness and Value.* New York: Oxford University Press, 1990.

WATSON, GARY. "Free Agency." In *Free Will.* New York: Oxford University Press, 1982.

WHITE, STEPHEN. *The Unity of the Self.* Cambridge: MIT Press, 1991.

WIGGINS, DAVID. *Sameness and Substance.* Cambridge: Harvard University Press, 1980.

WILKES, KATHLEEN. *Real People: Personal Identity without Thought Experiments.* New York: Oxford University Press, 1988.

WILLIAMS, BERNARD. "Bodily continuity and personal identity." In *Problems of the Self.* New York: Cambridge University Press, 1973.

WILLIAMS, BERNARD. "Are Persons Bodies?" In *Problems of the Self.* New York: Cambridge University Press, 1973.

WILLIAMS, BERNARD. "The Self and the Future." In *Problems of the Self.* New York: Cambridge University Press, 1973.

WILLIAMS, BERNARD. *Problems of the Self.* New York: Cambridge University Press, 1973.

WITTGENSTEIN, LUDWIG. *Philosophical Investigations,* translated by G.E.M. Anscombe. New York: Macmillan Publishing Co., 1958.

WOLF, SUSAN. "The Importance of Free Will," *Mind* 90 (1981): 386-405.