9-2015

Persons as Self-consciously Concerned Beings

Benjamin Abelson
Graduate Center, City University of New York

PERSONS AS SELF-CONSCIOUSLY CONCERNED BEINGS

by

BENJAMIN ABELSON

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2015

This manuscript has been read and accepted for the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Graham Priest

May 18th, 2015 _____

Chair of Examining Committee

John Greenwood

May 18th, 2015 _____

Executive Officer

John Greenwood

Linda Alcoff

Jesse Prinz

Marya Schechtman

Supervisory Committee

Abstract

PERSONS AS SELF-CONSCIOUSLY CONCERNED BEINGS

by

BENJAMIN ABELSON

Advisor: Professor John D. Greenwood

This dissertation is an analysis of the concept of a person. According to this analysis, persons are beings capable of being responsible for their actions, which requires possession of the capacities for self-consciousness, in the sense of critical awareness of one's first-order desires and beliefs and concern, meaning emotional investment in the satisfaction of one's desires and truth of one's beliefs. The persistence of a person over time requires uninterrupted maintenance of those capacities. This view is in conflict with the more popular account of persistence in terms of the continuity of distinctive psychological states. Furthermore, this account of personhood has the consequence that contrary to most alternative conceptions, the possession of rights to life and good treatment and the concern for others are neither necessary nor sufficient for being a person. In chapter one I explain and argue for my account of personhood in terms of self-consciousness and concern, illustrating that a being lacking either capacity would not be capable of responsible action and therefore would not be a person. In chapter two I argue for the claim that the persistence of a person requires only that those capacities are maintained uninterruptedly. Chapter three concerns the ontology of persons. There I argue for a Reductionist view of persons and defuse the objection that such a view necessarily slides into Eliminativism. In chapter four I draw a distinction between the concepts of 'person' and 'self,' arguing that the latter is not unique to persons and is best understood in neuro-cognitive terms. The fifth and final chapter deals with the implications of my account of personhood for ethics, as regards rights and concern for others.

Acknowledgements

This dissertation is dedicated to Raziel Abelson and Marie-Louise Friquegnon, who are my wonderful parents as well as masterful philosophers. I hope to have absorbed some small portion of the tremendous wisdom they have endeavored to impart to me throughout my life.

I am extremely grateful for my supervisor, John Greenwood, without whom I could not imagine having put this thing together. Nor can I imagine having a more patient, encouraging and challenging guide for the journey. I am also very grateful for the rest of my committee: Linda Alcoff, Graham Priest, Jesse Prinz, and Marya Schechtman for being with me every step of the process, providing me with invaluable feedback on various drafts of the project.

I am also greatly indebted to Christa Davis Acampora and David Rosenthal, the two philosophers who had the greatest influence on my thinking in my undergraduate and graduate studies, respectively.

I am doubly thankful for my amazing girlfriend Ana Talushllari, not only for her being unbelievably loving and supportive, but also for contributing to the project by sharing her rich knowledge of animal behavior and cognitive ethology, providing me with cutting edge research in those areas, and for lending her graphic design talents to the construction of Figure 1, which appears on page 211 in chapter four. And a very special thanks to our uncanny cats Bisou and Casanova, who were a constant source of inspiration as well as live-in case studies for how fascinating, intelligent, and worthy of our concern non-persons can be, and to Ana's parents Drita and Veri who were immensely accommodating and tolerant of my presence in their home while I wrote the vast majority of this work.

Special thanks go to my older siblings, Gabriel and Maris, who have been fantastic exemplars of how not only to be a good person, but a cool one as well. And also to my nephews and niece: Bodhi, Harry and Samantha, who are my favorite developing persons in the whole world.

Huge thanks to the following persons for reading and commenting on substantial portions of this work or otherwise influencing it enormously: Robert Black, Jonah Goldwater, Jessica Gordon-Roth, David Nagy, Nickolas Pappas, Arina Pismenny, and Richard Sorabji

Thanks to these persons, who impacted the project in various ways: Mark Alfano, Sophia Bishop, Charlene Blades, Frank Boardman, Brian Bollard, Richard Brown, Gregg Caruso, Ross Colebrook, Carl Craver, Bryan Danielson, Abraham Dickey III, James Dow, Cory Evans, Leonard Finkelman, William Fisk, Charles Goodman, Javier Gomez-Lavin, Richard Hanley, Katherine Hartling, Alan Hausman, Hyun Hochsmann, Eva Kittay, Thomas Kivatinos, Michael Levin, Eric Mandelbaum, Pete Mandik, Florence Nasar, Shaun Nichols, Jake Quilty-Dunn, Rick Repetti, John Richardson, Angelika Seidel, Elisabetta Sirgiovanni, Henry Shevlin, Leana Shugol, Sandeep Sreekumar, Thomas Teufel, Peter Unger, Denise Vigani, Alexis Vigo, and Thomas Whitney.

Table of Contents

Introduction

I. What is a Person?

The question “What is a person?” might ring oddly in ears unfamiliar with the philosophical debates surrounding it. Someone who poses such a query is likely to receive a reaction of puzzlement or even ridicule. Among the more helpful responses to it is the counter-question, “Isn’t a person just a human being?” This is helpful because the philosopher asks the former question only when he or she has in mind a distinction between the concept of a person and that of a human being. To illustrate this distinction, the philosopher is likely to mention fetuses or individuals in irreversible vegetative comas, as examples of humans that are non-persons on the one hand, and intelligent space aliens, artificial intelligences, or superevolved non-human animals, as examples of non-human persons, on the other. The first hand holds examples of genetically human creatures that don’t seem to meet the criteria for personhood and the second hand holds examples of non-humans who do seem to be persons. Now the original question can be recast in light of these examples: What do the human non-persons lack and the non-human persons have in common that is essential to being a person? This question might sound a bit less odd to the layperson, particularly if he or she has had any exposure to science fiction or the more publicly represented bioethical issues, such as abortion, euthanasia, and animal rights, or legal issues concerning the status of corporations. Often one speaks of the qualities of ‘humanity’ when one really has in mind features that could be possessed by something that is not genetically *homo sapiens*. The assignment of ‘human’ to the biological species concept and ‘person’ to the more abstract metaphysical notion is a matter of philosophical convention, but there

is a conceptual distinction to be drawn that is not a matter of convention but represents a genuine difference in meaning between biologically specific humans and trans-specific persons. Thereby, I follow the philosophical tradition in stipulating that in the context of this study, ‘human’ denotes belonging to the biological species homo-sapiens and ‘person’ denotes belonging to the trans-specific class.

For thinkers at the dawn of modern philosophy and science, particularly substance dualists such as Rene Descartes, Joseph Butler (1736), and Thomas Reid (1785), personhood was understood as depending on the possession of an immaterial, immortal soul. Persons were thought to be the unique possessors of minds, free will, and morality. According to Descartes, human beings are the only creatures to possess those characteristics (though he didn’t use the term ‘person’), by virtue of having a soul that is separate from their physical or mental components and properties, standing behind those components and properties, and evidenced by their unique capacity for language. He regarded all non-human creatures as *mere* mechanisms, without thought, will, or feeling. For him, only human beings oversee their own bodily mechanisms from the executive seat of the soul. The strict duality between soul-possessing persons and mechanical non-persons has since been rejected by most philosophers, and the notion of a separate soul with a causally undetermined will has been largely abandoned by philosophers with naturalistic inclinations. Even without such jettisoning, John Locke (1690), as a kind of proto-functionalist, saw that it didn’t matter what sort of stuff a being was composed of, but rather how that stuff was organized and what it could do. For a

non-physical substance to constitute a personal soul it would have to have the capacities constitutive of personhood. He says that

> *'Person* stands for… a thinking, intelligent being, that has reason and reflection, and can consider it self as it self, the same thinking thing in different times and places; which it does only by that consciousness, which is inseparable from thinking, and as it seems to me essential to it. (Locke 1690, II.XXVII.335.10)

So what matters is not that the soul is immaterial or separate, but that it possesses those person-constituting capacities, i.e. thinking/consciousness, intelligence, reason and reflection. If it is conceivable that matter could instantiate such capacities, then there is no reason to insist on the existence of a non-physical soul. Descartes may have objected that there are qualities of conscious experience that can *only* be instantiated in an immaterial substance, but he had no good way of explaining how those uniquely mental properties could have any relation to the purely mechanical processes that they appear to cause and be caused by.

If one recognizes the similarities in behavior and shared neurofunctional architecture between human beings and other creatures, then one should, as Locke seemed to, regard human persons as on a continuum with other animals rather than marking a radical ontological break. Moreover, because Locke understood persons in terms of their characteristic functions, he considered it possible that some other animal species, e.g. a super-intelligent parrot, could manifest those characteristics and therefore be rightly regarded as persons. In the Lockean spirit, I will here endeavor to draw a modest, but principled ontological distinction between persons and non-persons, though the line may not be in quite the place it has been often supposed to be. Some

creatures that are commonly regarded as *mere* animals, e.g. chimpanzees and dolphins[1] might turn out to be persons. Unfortunately, epistemic constraints may make it so far impossible to decide with certainty whether or not such creatures possess the capacities constitutive of personhood, but after deciding what those capacities are, there will be evidence to suggest that some beings have a higher likelihood than others of possessing them. Nevertheless, I will make no decisive claims about which beings, beyond the paradigmatic cases, are, and which are not persons. In some cases it is likely that there is insufficient evidence to judge either way with any authority. However, even if one cannot be certain which beings possess the capacities constitutive of personhood, one can still come to a clear understanding of what those capacities are, which, if any being realizes them, make that being a person.

To explain the concept of a person in terms of the possession of certain capacities that distinguish persons from other beings is to understand the concept as a metaphysical one, one that carves out different categories of being. However, one might also draw a distinction not only of nature but also of moral status, between persons as being the sorts of things which bear rights, responsibility, and moral awareness and non-persons which do not have that status. According to the moral concept of a person, only persons can be responsible in such a way that they can be reasonably praised and blamed for their actions. Such responsibility is often taken to be the ground for taking persons to be of special value, thereby having a unique claim to rights. For this reason,

[1] India has already declared that dolphins are legal, non-human, persons.

individuals who seek to establish protection for the rights of beings as diverse as dogs (Berns 2013), rivers (Messenger 2012), and human fetuses, have based their claims on the premise that these beings are persons. Finally, some writers have built a capacity for moral awareness and empathy into the concept of a person. Susan Wolfe (1987), for instance, has claimed that “sanity” is a necessary condition for responsibility, and therefore personhood, where sanity includes a recognition of the moral properties of situations.

More often than not, these two concepts of personhood, metaphysical and moral, are assumed to be bound up together. The difference in metaphysical attributes is supposed to account for the distinction in moral status. In this study (particularly, the concluding chapter) I will contest this assumption in its strongest forms by offering reasons for reconsidering the relation between the two concepts of personhood and arguing that while personhood is inextricably bound to notions of responsibility, much of the additional moral status commonly attributed and often thought unique to persons does not, in fact, follow from the metaphysical concept of a person.

Recently, some writers have worried that if the metaphysical concept of a person is ambiguous, then the moral conception has no clear foundation and the concept of a person should therefore be removed from the arena of ethical discourse. Gordijn (1999), for instance, thinks that since disagreement over criteria for personhood is intractable, its employment in bioethical debates over abortion and euthanasia is inevitably question-begging and therefore, irresponsible, as each disputant in the debates has his

or her own pet conditions of personhood tailored to supporting the ethical conclusions the disputant is arguing for. I agree with Gordijn that ‘person’ is not particularly useful for solving bioethical debates, but not for the reasons he supposes. I will argue that disagreement over the necessary and sufficient conditions of personhood is not as intractable as Gordijn thinks it is and that once we have sorted out those conditions, it will become clear that much of the moral status assumed for persons does not follow from the metaphysical concept, so that the bioethical disputes Gordijn is concerned with cannot be settled solely by appeal to personhood, but for different reasons than Gordijn provides.

The characteristics which have been assumed by various writers to be constitutive of persons do seem to form a rather heterogeneous and possibly inconsistent set. Gordijn offers the following list, which is representative if not exhaustive of that set and includes Locke’s aforementioned criteria:

1. The capacity to experience pleasure and/or pain;
2. The capacity to have desires;
3. The capacity to remember past events;
4. The capacity to have expectations with respect to future events;
5. An awareness of the passage of time;
6. The property of being a continuous, conscious self, or subject of mental states, construed in a minimal way, as nothing more than a construct of appropriately related mental states;
7. The property of being a continuous, conscious self, construed as pure ego, that is, as an entity that is distinct from the experiences and other mental states that it has;
8. The capacity for self-consciousness, that is to be aware of the fact that one is a continuing, conscious subject of mental states;
9. The property of having mental states that involve propositional attitudes such as beliefs and desires;
10. The capacity to have thought episodes, that is, states of consciousness

involving intentionality;
11. The capacity to reason;
12. The capacity to solve problems;
13. The property of being autonomous; that is of having the capacity to make decisions based upon an evaluation of relevant considerations;
14. The capacity to use language;
15. The ability to interact socially with others; (Gordijn 1999, 353)

However, I will attempt to show that most of the items on this list hang together in rather clear ways and the ones that don't fit are dismissible for good reasons. After offering my own account of necessary and sufficient conditions, I will look back at this list and attempt to demonstrate that the account is inclusive of some of those features listed above, and in the cases where it is not, argue for why the excluded conditions are rightfully excluded, for they are not necessary for being a person.

II. Objections to seeking criteria of personhood

The present study is an attempt to discern the necessary and sufficient conditions of personhood, in other words, the conditions any being must meet in order for us to accurately call it a person in a usefully consistent way. Before explicating my methodology for navigating the conceptual terrain, I must contend with various objections that have been raised to the very idea of trying to establish necessary and sufficient conditions of personhood. One sort of objection claims that 'person' is best thought of as an 'open-textured' concept, one that does not have exhaustive conditions of application. Choosing a particular closed set of conditions, the objection continues, could only be done arbitrarily and no matter where the line was drawn, it would unduly

restrict the potential future scientific employment of the term.

Waismann (1945) refers to the “open texture” of concepts as part of a general critique of verificationism as propounded by Mackinnon (Mackinnon 1945)[2]. While the idea of providing necessary and sufficient conditions for personhood is independent of Mackinnon’s verificationism, Waismann’s point nonetheless poses an important challenge to it. Waismann explains what he means by “open texture” in several different ways. I will consider only some of them, because it is not entirely clear that there is a single notion in play (‘open-texture’ itself may be open-textured) and only some senses of the term are relevant to the present discussion. To begin with, Waismann claims that “The failure of the phenomenalist to translate a material object statement into terms of sense-data is… due... to the ‘open texture’ of most of our empirical concepts.” (Waismann 1945, 121) He offers as an example the attempt to verify the statement “There is a cat next door” and wonders what would count as sufficient verification.

> I go over to the next room, open the door, look into it and actually see a cat. Is this enough to prove my statement? Or must I, in addition to it, touch the cat, pat him, and induce him to purr?... can I then be absolutely certain my statement was true?... What for instance should I say when that creature later on grew to a gigantic size? Or… it could be revived from death… Shall I in such a case say that a new species has come into being?... Have we rules ready for all imaginable possibilities? The fact that in many cases there is no such thing as a conclusive verification is connected with the fact that that most of our empirical concepts are not delineated in all possible directions. (Waismann 1945, 122)

So the fact that there is no exhaustive set of conditions that would allow us to decide with certainty whether or not a statement is true is supposed to be explained by

[2] Mackinnon’s paper was published together with the replies from Waismann and Kneale which are listed together under ‘M’ in the bibliography.

the fact that most of our concepts “are not delineated in all possible directions.” If one thinks that the meaning of a statement can be explicated entirely in terms of the conditions of its verification then it would make sense to hold the nature of the concepts themselves at fault for our lack of certainty in when to apply them. This can be seen as problematic even if we aren’t trying to translate our statements into talk of sense data. Any attempt at offering exhaustive necessary and sufficient conditions of a concept must take into account a perhaps inexhaustible number of potential counterexamples. If faced with rapidly growing or undead things that otherwise resembled cats we would have to make a choice between recognizing different sorts of cats or saying that these new things were not cats at all.

Waismann’s problem seems to be that our current way of defining cats in terms of characteristic physical features, behavior, dna, who they can mate with to produce fertile offspring, etc., don’t provide us with clear guidelines for how to deal with the weird possibilities. Any further specification of conditions that create a rule for deciding in some cases could continually be challenged by further cases, so that “we can never eliminate the possibility of some unforeseen factor emerging, we can never be quite sure that we have included in our definition everything that should be included, and thus the process of defining and refining an idea will go on without ever reaching a final stage. In other words, every definition stretches into an open horizon.” (Waismann 1945, 125) The problem in the case of persons is that the kind of science fictional possibilities that the concept must account for may be of endless variety and we might

not be able to anticipate some of them in advance. We could then never be sure if our definition of the concept of a person were inclusive of all possible persons to the exclusion of all possible non-persons.

However this conclusion need not be unwelcome, if one has a modest vision for the task of providing necessary and sufficient conditions. If, rather than attempting to decide all and only the cases in which a concept can be successfully applied, one is merely describing the way a concept is actually applied while perhaps also suggesting that its application could be refined for the purpose of greater clarity or precision, then such possibilities need not be taken as undermining that project. Rather, these sorts of cases make the project of definition interesting, in that one must attempt to make one's definitions responsive to as many odd cases as possible while finding principled reasons for drawing the lines where one does. In other words, the as yet only possible, but nevertheless imaginable cases challenge one to provide a definition inclusive of such cases, but that definition need not be *exhaustive* of all possible cases that may arise.

To illustrate Waismann's claim that many of our empirical concepts are open-textured, Sclafani (1967) points out that the necessity of positing imaginary numbers could not have been anticipated until the development of the quadratic equation. If our concept of 'number' had been too rigidly defined, then we could not have included these very helpful entities within it. Another example is the case of motion pictures. The rise of motion pictures as an art form could not have been anticipated in advance of the

invention of the right sort of camera. For that reason, one should think of the concept of art as open in texture, so as to allow new, as yet unimagined, forms to be included in it. So, if the concept of a person is similarly open in texture, one may object that establishing necessary and sufficient conditions may unduly rule out as yet unimagined variations on what we now consider to be persons. These two examples are instructive in different ways. In the case of number, Sclafani holds that there was a previously clearly defined conception of number that was modified in order to include the new entities, yielding a revised but still clearly defined concept.[3] This required a choice - extend the definition to include the new entities or refuse to include the new entities and say that there are two different concepts of number - the old outdated one, and the newly refined one. In such a case there are various theoretical virtues to be considered. For mathematicians the utility of considering imaginary numbers as numbers was worth paying the cost of conceptual mutilation.[4] Because of the possibility of such cases we must only think of our concepts as closed relative to what has so far been anticipated, but subject to revision should some as yet unanticipated possibility come to light. Once these possibilities are realized, the prior concepts may be modified to accommodate them. In other words, the concepts *change* to accommodate the new phenomena. So I

[3] It is likely, however, that this is not the actual truth about the history of imaginary numbers. It might be a more accurate story to tell about the introduction of *irrational* numbers following the discovery by Pythagoras of the incommensurability of the diagonal of a unit square with its sides, because it forced the Greeks to understand numbers in more theoretical and less concrete terms. (As explained by Priest 1998) Imaginary numbers, on the other hand, have a more complicated history of use, acceptance and explanation. Still, the example as stated, is helpful for illustrating the idea of concept revision.

[4] In some places Waismann's view seems to foreshadow the Quine/Duhem holism thesis, the latter of which I take to be perfectly compatible with my view of concepts.

am happy to concede that whatever necessary and sufficient conditions of personhood I am able to establish, they will only cover the as yet anticipated cases and that these conditions are revisable in light of any as yet unanticipated possibilities.

However, there need not always be conceptual revision in the light of new as yet unanticipated possibilities. In the case of art, it seems that the resources for deciding whether or not film should count as art were already provided by the pre-cinema concept, which already had to be sufficiently general to include painting, sculpture, music, photography, etc. The conditions of application of a concept may be general enough that novel instances fit happily under the old concept. My general response to the issue of open texture as a challenge to providing necessary and sufficient conditions of concepts is that we need not understand settling on a definition of a concept as closing the matter for evermore. New possibilities may arise, even new actualities that would force us to choose between extending, modifying, or even abandoning our concepts, but that doesn't mean the concepts need go undefined in the meantime.

One other way one might challenge the project of defining personhood in terms of necessary and sufficient conditions is by claiming that it is best treated as a 'family resemblance' concept which is often associated with a 'prototype theory' of concepts, in opposition to the 'definitional view'. Armstrong, Gleitman, and Gleitman (1999) characterize the definitional view as follows: "a smallish set of the simple properties are individually necessary and severally sufficient to pick out all and only, say, the birds, from everything else in the world. Membership in the class is categorical, for all who

partake of the right properties are in virtue of that equally birds; and all who do not, are not." (Armstrong, Gleitman, and Gleitman 1999, 227) The definitional view is the general strategy of the present study, with the caveat that I don't expect or require that the criterial properties of persons be 'simple' or even decomposable into 'simples,' but only that persons are definable in terms of necessary and jointly sufficient properties or capacities. Also, while my approach is committed to there being a clear line demarcating where an individual that meets it is fully a person, I admit there could be greater or lesser degrees of personhood approaching that limit. Still one might wonder if 'person' is really amenable to such a clear-cut definitional strategy. The proposed alternative is to treat it in terms of family resemblance, where there is no specific set of properties that are all necessary for inclusion under the concept, but where different instances of a concept have different combinations of some but not all of the properties generally associated with the concept. Beginning with Wittgenstein (1958), 'game' has been the typical example of a family resemblance concept. Baseball, solitaire, and ring-around-the-rosy are all games, yet there is no single feature or set of features they all share in common which distinguish them from all non-games, though all games do share some features in common with some other game(s). That there are examples of a concept, some of which are more and some less prototypical, is often taken to be evidence that the concept is best handled by the family resemblance model. For example, one might think that baseball is a more prototypical example of a game. The other games are such insofar as they resemble the prototype, though each does so in

respect of different subsets of features to different degrees.

Armstrong et al. cite some empirical data that prototype theorists take to be evidence for considering something a family resemblance concept. The data cited shows that for some kind-concepts such as ‘game,’ ‘fruit,’ and ‘bird’ subjects consistently rank different examples as better or worse instances of the kind in question. The theory suggests that the examples indicated to be the best are the prototypes, and the less typical examples are ranked in terms of their resemblance to the prototypes. However, it is clear that not all concepts that have prototypes are family resemblance concepts. For instance, ‘odd number’ has perfectly clear criteria of application, and yet, Armstrong et al point out, respondents display the same ranking behavior for those concepts as they do for the ones reasonably supposed to be cluster concepts. They write:

> Are there definitonal concepts? Of course. For example, consider the superordinate concept *odd number*. This seems to have a clear definition,a precise description; namely, *an integer not divisible by two without remainder.* No integer seems to sit on the fence, undecided as to whether it is quite even, or perhaps a bit odd. No odd number seems any odder than any other odd number. But if so, then the experimental paradigms that purport to show *bird* is prototypic [and therefore, not definitonal] in structure in virtue of the fact that responses to ‘ostrich’ and ‘robin’ are unequal should fail, on the same reasoning, to yield differential responses to ‘five’ and ‘seven’ as examples of *odd number*… [T]he facts are otherwise. For graded responses are achieved regardless of the structure of the concept…. [S]ubjects... judged *3* a better odd number than *501*. (Armstrong et al. 1999, 234-237)

That there are commonly recognized prototypes of a concept is often taken to be evidence that the concept cannot be defined in terms of necessary and sufficient conditions, and can only be analyzed into relations of family resemblance. However,

Armstrong et al.'s findings show that instances being ranked as prototypical or exemplary of a concept does not imply that the concept is best understood in terms of family resemblance, nor vice-versa - e.g. actual families may not have prototypical members. In summary, ranking behavior with regard to a concept is not evidence for it being a family resemblance one. So just because there are more or less prototypical persons does not entail that 'person' is a family resemblance concept.

Ranking behavior is not the only reason why someone might think a concept is a family resemblance one and not one with clear necessary and sufficient conditions of application. One might only point to places in which two different accepted usages conflict with one another, so that it is just not clear whether or not the concept applies. In such cases it seems to me that there are not one but two concepts with the same name, so one must decide which is the primary use of the term and find an alternative label for the concept which is abstracted from the secondary usage.

Timothy Chappell (2011) offers a different sort of objection to the idea of establishing criteria for personhood, claiming that our "normal decision procedure" when choosing whether or not to regard a being as a person is not to check off a list of criteria. For instance, we treat human children as persons long before they manifest most of the features usually thought constitutive of personhood. He argues that, therefore, one should be a "humanist" about persons, holding that being human is sufficient, though not necessary, for being a person.[5] Chappell allows that other

[5] Schechtman (2014) makes a similar claim.

species, space aliens, or “spooky refrigerators” could come to be regarded as persons. He even concedes that in those cases we may resort to our list of criteria, among other “factors and reactions, most of them defying explicit articulation” in order to decide on a particular inclusion, but that “it would be utterly misleading to generalize from thought experiments about these special and rare cases – almost all of which, to date, are imaginary – to alleged conclusions about the normal cases” (Chappell 2011, 19).

Chappell’s concern about establishing criteria for personhood is, similarly to Gordijn’s, a worry about the ethical implications of such an approach. ‘Person’ has commonly been supposed to denote all and only the members of the “primary moral constituency, (PMC): some class of creatures who exclusively and equally share in the highest level of moral rights and privileges” (Chappell 2011, 2). Therefore, Chappell infers, if one is a criterialist (which I take to be synonymous with definitionalist) about persons, and includes self-consciousness, rationality, intelligence, etc. to be criteria of personhood, then many genetically human beings, including the severely cognitively impaired, infant children, and even normal humans who happen to be asleep at the moment, do not count as persons, because they do not display any or all of those properties. He worries that such beings don’t count as persons they may be discriminated against and mistreated on the grounds that they are not part of the PMC. In chapter five I will more thoroughly address this worry, by arguing that one can deny that small children and also the severely cognitively disabled are full-fledged persons while still granting that as human beings, they are the bearers of all the rights afforded

to other human beings that they are capable of enjoying. The objection to seeking a definition of the metaphysical concept of a person which stems from a worry about the moral status of children or the cognitively impaired can be defused if it can be shown that persons are not the unique bearers of rights. But that argument aside, there is less reason to think that criteria of personhood would rule out as persons many of the human beings Chappell is concerned for.

For normal human beings who sometimes happen to be unconscious, as most are prone to be at regular intervals, they can easily be included in a definition of persons if it is stated it in terms of the *capacity* for exhibiting or instantiating those properties rather than the actual exhibiting and instantiating themselves. There is a further distinction to be drawn in the case of children, by speaking of their *potential* for such capacities. I will have more to say about capacities and potential in chapter one, when discussing my own view of the capacities necessary and sufficient for personhood.

Furthermore, while Chappell may well be right that we treat some beings as persons before they fully realize those properties characteristic of persons, we don't just do this for human children, but also household pets. People scold their dogs and cats as if they were responsible for their actions just as they do their own children, but that doesn't mean that those beings *are* persons. In the case of human children, we have good reason for treating them as persons, because there is strong evidence that the more you treat a child like a person, the more likely it is that it will become a full-fledged person - that (as demonstrated by the example of feral children) if a child grows up

without being treated like a person, the person capacities will be impaired.

Moreover, Chappell's argument generally conflates two different senses of "criteria." If one means by that term the properties people actually use for deciding whether or not someone is a person, then he is right to point out that most people in actual situations don't check off a list of criteria in making such decisions, but if by criteria we mean the features a being must have in order to *be* a person, independently of the epistemic access one has to that fact, the point about decision procedure is irrelevant. In other words Chappell is mixing up the conditions of empirically recognizing persons with the conditions for being a person. Even if the former are not available on a particular occasion, the latter may still exist and be explicable.

There are many other objections that I will have to grapple with in the course of this study. Here I have only listed those which are posed against, to borrow the title of Chappell's essay, "the very idea of criteria for personhood." I hope to have shown that none of them is fatal to the project, and that it would be highly advantageous to establish defined criteria for the application of the term 'person.' But to return to the central issue at hand, what does bear on the question of what a person is? How are we to decide which features are essential to being a person? The following section will be an explication of my method for this study, which can be summed up as empirically informed abstraction from primary use.

III. Discussion of method

Talk of necessary and sufficient conditions will likely evoke thoughts of the old style of conceptual analysis which was the hallmark of the ordinary language philosophers of the mid-20th century, such as Gilbert Ryle, J.L. Austin, and P.F. Strawson. Jose Luis Bermudez (2005) somewhat deridingly, characterizes conceptual analyses as

> purely a priori. They are neither justifiable nor answerable to any empirical facts that we might discover about the phenomena in question. They are obtained by reflecting on the various components of our conceptual scheme, by trying to identify relations of dependence between particular concepts and by constructing thought experiments that will test our intuitions and hence (so the theory goes) provide guidance as to how to understand particular concepts… constructing sets of necessary and sufficient conditions that would pick out all and only the situations in which we would intuitively say that [the concept in question applies].

The complaint against such a method of philosophizing is two-fold. First of all, by proceeding entirely "from the armchair" as it were, "purely *a priori*" analysis is supposed to be isolated from any new empirical information that might shed light on a subject of inquiry. However it is unclear that there is a definitive line between our pre-theoretical intuitions of what things are and what we discover about them empirically. Common sense is related dynamically to scientific discovery in such a way that we should regard our concepts as revisable in light of new developments. I am sympathetic to this line of objection and so wish for the account of the concept of a person that I will offer to be responsive to and continuous with contemporary scientific data. Our so-called "intuitive" responses to definitional questions are informed to varying degrees by our

understanding of the science of our time and cannot be thought to emerge out of a conceptual vacuum. However, once all the available relevant empirical data has been collected, there is still analytical work to be done.

Secondly, one might object that the attempt to define a concept by appealing to strange or purely hypothetical situations or thought experiments reaches far beyond the situations in which our concepts are actually employed, and is, therefore, ill-suited to the purpose of establishing necessary and sufficient conditions of application for the actual cases in which we would be warranted in applying a particular concept. As Bermudez puts it at greater length:

> our ordinary conceptual scheme developed to provide a framework for thinking about the types of objects and situations that we tend to encounter, and we can expect it to be silent on such questions as whether or not [for example] to attribute knowledge to someone who finds himself in a region that he knows to be full of fake barn facades made from papier-mâché and correctly identifies the object in front of him as a barn, even though he has not first checked to rule out the possibility that it might be a papier mache barn façade. (Bermudez 2005, 7)

Bermudez suggests we relax the constraints on what counts as a successful conceptual analysis, by restricting our analysis to how a concept is applied in the situations for which it has been designed, and ignoring the unlikely hypotheticals. To ask that our account of a concept have clear criteria of application in all conceivable situations, according to Bermudez, goes beyond the warrant of analysis of our everyday concepts, and is, rather, “a refinement or sharpening” of them.

However, it is not clear to me that we can draw any kind of principled distinction between cases for which our “conceptual scheme was developed” and those to which it

does not extend. Conditions of “ordinary use” are encountered woven together with imaginative exercises and scientific revelations in a complex tapestry that is without discrete boundaries between designs. Therefore, I depart from Bermudez by taking thought-experimental cases to be legitimate elements for theorizing. While I agree that we shouldn’t take our so-called intuitions about such cases to be an independent tribunal whose verdict would then be imposed on a concept in all cases, I do think we should understand part of the “integrative” task Bermudez speaks of to include making our concepts as rich as is tenable in the sense of being consistently applicable in the most conceivable situations.

On the other hand, I recognize along with Bermudez that there will be situations where we should curtail a concept so as not to risk contradiction in use. For example, while often in common usage and always for Locke, ‘self’ refers, reflexively, to ‘this person’, I will argue (in chapter four) that the term ‘self’ be given a distinct analysis from that of ‘person’ when doing metaphysics, because making such a distinction will be extremely useful without requiring so great a departure from how the two terms are commonly employed as to make them unrecognizable. ‘Person’ designates an objective being, one that can be an object of empirical investigation by others. ‘Self’ designates the internal, subjective representations that persons (but also animals and other organized beings) generate to monitor their own internal states which may more or less accurately represent those states. In persons, the ‘self’ usually includes one’s sense of social, moral, and narrative identity.

My general point is that there is need for both art and science in the analysis of concepts. Science will provide ever new threads that must be appropriated into our shared conceptual quilt, creating new folds in the fabric to be massaged or snipped by reflective analysis. The ideal should be level panels and clear, discreet seams, but only where the texture of the raw material allows.

Given the obstacles to be encountered when trying to weave and trim the conceptual fabric of personhood, or any other difficult concept, one must prioritize based on what one takes to be the most important uses to which the concept can be put. Hence, the method I will employ is abstraction from primary use, which, in the case of personhood, asks the question: which properties do persons possess such that the concept serves the purpose for which it is primarily utilized? There are legal purposes, metaphysical purposes, and moral purposes, with plenty of overlapping aspects between them. It has become rather common to speak of three distinct concepts of a person along those lines, though the overlapping makes drawing the lines difficult if not impossible. The moral notion is ideally supposed to ground how the legal one is defined and the metaphysical concept is often held to provide a foundation for the moral.

There is a long history of understanding the primary usefulness of 'person' to be for making judgments of responsibility. The concept of a person, in its modern form (i.e. not for defining angels, explaining the trinity, or demarking theatrical roles[6]), was given

[6] This latter usage is, arguably the actual historical origin of the term. According to Martin & Barresi (2006, 29): "The Greek word '*prosopon*' originally meant playing a role in a drama or in a religious ceremony. However, with the rise and democratization of the Greek city-states, the

its first thorough explication by John Locke in 1690. He understood it as a 'forensic' concept. By 'forensic' he meant that its purpose is to track which individuals are responsible for which actions.[7] We hold human beings, the paradigmatic persons, responsible in ways that we do not hold dogs, cats, bears, lizards, birds, or pigs, and for good reasons that will be explicated in what follows. Furthermore, if faced with an antagonistic member of an alien species, we would wonder after capturing it, whether it is the sort of thing that is responsible for its actions. Our answer to that inquiry would help us to decide the appropriate way to engage the alien. We can also speculate about whether an intelligent computer could be responsible for its actions. These considerations figure in most, if not all situations for which we wonder whether or not something is a person. They are considerations about what it would take for a being to be responsible for its actions, what capacities it must possess in order to act in a way such that it would be reasonable to hold it responsible. Therefore, it is reasonable to proceed on the supposition that tracking responsibility remains the best candidate for

---

word began to acquire a wholly secular meaning, which had to do with social and legal roles. Certain kinds of citizens were recognized as having rights and duties that distinguished them from others. In earlier Greek thought about people and society, the emphasis was on these roles. Only slight attention was given to the individuals who occupied the roles. People were regarded as little more than placeholders. However, when the Greek city-states declined, there followed a period of pessimism during which the traditional emphasis on harmonious relationships in the polis among essentially replaceable individuals waned. Cynics and Stoics, in particular, emphasized inner resources for adaptation to the general malaise. This gave rise to a new emphasis on individualism. The Latin term '*persona*,' from which the English term *person* derives, acquired its modern meaning from within the context of this latter development." As I see it, the "modern meaning", resulting from the new emphasis on individualism, consists in the term being primarily employed in forensic contexts.

[7] Jessica Gordon-Roth has suggested to me that Locke's understanding of person as a forensic concept has been overemphasized and that he had other uses for the term in mind, but it seems to me (and Gordon-Roth does not necessarily disagree) that the forensic usage is still, for him, the primary one.

the primary use of ‘person’.

Schechtman (2014) provides a good example of an alternative method in stark contrast to my method of abstraction from primary use. Rather than fishing out the one primary use of the term, Schechtman’s project is to develop a conception of personhood that weaves together all the various practical purposes for which it is commonly utilized. Drawing on the work of Lindemann (2001 and 2009) she seeks conditions of personhood that go “beyond the sophisticated forensic concerns that are usually the focus of practically oriented discussions of personal identity,” because

> treating someone as a person does not only involve treating her as a moral or rational agent, but includes the full range of everyday behaviors that make up the lives of human persons… taking for granted that persons wear clothes and are given names rather than numbers, or that they are referred to as ‘who’, rather than ‘what’. The social recognition that constitutes our identities… goes far beyond the acknowledgment of rights and responsibilities… To recognize someone as a person is not to make a particular kind of judgment about her, but rather to treat her in the myriad ways that this form of life entails, those that involve moral responsibility and autonomous agency and those that do not. (Schechtman 2014, 72)

These considerations lead Schechtman to her distinctive position on personhood and personal identity, which she calls the “Person-Life View” (PLV). According to PLV, to be a person is to live the characteristic life of a person, which requires the social recognition that confers personhood upon individuals, according them a place in “person-space” in addition to possession of the characteristic capacities of persons, i.e. self-consciousness, rationality, etc. which make them appropriate targets of forensic practices, because the recognition and possession of capacities are interdependent and conceptually inextricable -- the “capacities do not develop without the interactions and

activities that make up a human life" (Schechtman 2014, 116). Because Schechtman wants a conception of personhood that is sensitive to all the practical contexts and purposes in which it is actually employed, she is led to a view that is inclusive of all human beings, who from birth are recognized as persons, and which seems to exclude many (but not all) other sorts of being that might possess the capacity for responsible action but have a radically different form of embodiment. Such beings are excluded if they are unable to interact in ways characteristic of persons (ways that go beyond forensic practices), and hence cannot be accorded places in human person-space or else have their own analogous social relations among one another. Schechtman only allows for a non-human to be a person if one is capable of "living a person-life within the social infrastructure that defines such a life," in which case "we cannot but include her in person-space," though this does not require that such a person "engage in the canonical forensic interactions," as is the case with "humans of atypical developmental trajectories." (Schechtman 2014 132) However, I see things the opposite way, where the particular social infrastructure in which a being finds herself is irrelevant to her personhood, whereas whether or not she can engage in forensic interactions is essential. While I agree that as a matter of contingent fact, the capacities constituent of human personhood develop in the context of typical person-lives, such that outside of such a context human beings are unlikely to manifest those capacities, I don't take this to be a conceptual or necessary truth.

I take my own project to be more about understanding personhood as a

‘metaphysical’ concept (as opposed to epistemic or honorific) than Schechtman’s, (though she does not see our projects as differing in this way), in the sense that I constrain my analysis to the features that all persons intrinsically possess, as opposed to also taking as essential the ways that we treat them or our ways of socially engaging with them. Of course our ways of treating them and engaging with them often depend on the features they intrinsically possess, but, contrary to Schechtman’s view of the situation, not all such practices consistently track genuine features. I see personhood not as a status that must be conferred, but a set of a capacities that a being either has or doesn’t have regardless of whether or not it is recognized by anyone else. The practices associated with responsibility, I hope to show, do, at least ideally, track intrinsic features of persons that are essential to their personhood in a potentially consistent way, whereas the features Schechtman invokes seem to have more to do with arbitrary or culturally relative traditions. While she does take forensic capacities to be the “most salient and distinguishing characteristics of persons” she thinks they only have that importance because “they lead to and guide the development of the social and cultural infrastructure that characterizes person lives.” (Schechtman 2014, 131) However, it seems to me that if the same capacities in different environmental conditions would lead to very different sorts of lives, we would have no less reason to call them ‘person lives’. Such lives might be radically different from human person lives, but the whole point of the concept of a person is to delimit the possibilities of persons that are not human and may be very different from humans in all respects other than

their personhood. What matters is whether or not these possible lives include the forensic practices.

For Schechtman:

> Intelligent balls of light energy that could not take human form, do not need sustenance from the environment, and do not reproduce in anything like the way animals do would undoubtedly have a social organization so different from ours that it is exceedingly difficult to see how we could understand their form of life or engage with them in forensic interactions. Here we are no longer talking about nonhumans with the capacities of persons, but rather about beings with other kinds of capacities that may be equally or more sophisticated than our own. Such beings would not be persons according to PLV although they may be highly intelligent. Lest this seem chauvinistic, we should recall that our goal is to understand the nature of beings like us and to explicate the conditions of their individuation. At some point we are no longer talking about "beings like us" and that is all the denial of personhood amounts to in this context. (Schechtman 2014, 134)

I agree that there are conceivable beings as intelligent, or more intelligent, than us that would not be persons. I also agree that in some sense when analyzing the concept of a person the idea is to understand what 'beings like us' are. However, the 'like us' doesn't refer to just any similarities. Tool use puts us in a class with crows and chimpanzees, to the exclusion of cats and dogs, but there is no reason to think that's a class of persons. My task is to figure out which features we have that make other beings 'like us' in the respect that they are persons, and for that reason I must be selective in choosing which features are relevant. I hope to have given sufficient reasons for selecting the forensic capacities as the relevant ones. No doubt, there may be highly intelligent beings that differ from us so greatly that they aren't persons, but only insofar as those differences make them incapable of the kind of responsibility I take to be

essential to personhood. It may be extremely difficult to imagine how a ball of light energy could be put on trial. Perhaps we could never include it in our forensic practices. However, that doesn't rule out the possibility that there could be balls of light energy that do have capacity to take responsibility and hold each other responsible for some actions. In such a case, those beings would be persons. Without such capacities they would not be so.

However, according to Schechtman, we rightly consider all human beings to be persons, even if they are disabled in ways that make them incapable of responsibility, and that this conferral of personhood on those human beings and not on, e.g. household pets, is non-arbitrary, for it is grounded in our expectations about normal human beings vs. beings of other species. She writes:

> PLV sees humans with atypical development prognosis as persons for much the same reason that it sees human infants as persons -- because there is a default expectation that such infants will develop into beings with the full complement of forensic capacities; an expectation which is over-ridden in the atypical cases but does not disappear or cease to do work even when we know the expectation will not be met. (Schechtman 2014, 123)

Because atypical humans are atypical of a kind of being that normally engages in forensic activities, their inability to perform such activities must be excused or explained, whereas the inability of one's dog to engage in such practices need not. This, according to Schechtman, gives the atypical human a default position in person-space. However, I fail to see how that is not merely an artifact of our attitudes and particular cultural practices as opposed to a fact that grounds our attitudes. One can imagine (and unfortunately may not need to imagine long) a callous society in which cognitive

disabilities can be diagnosed before birth. If allowed to come to term, beings judged to be incapable of forensic interactions would be branded and then treated as pets, left unclothed, fed from a trough and given a litter box or taken for walks to relieve themselves. What reason, besides our feelings of outrage could we offer such a society for why they should not treat such individuals in that way? If we say, as Schechtman seems to suggest, that they are persons because we expect beings genetically like them to be persons, the reply would likely be that in these cases the expectations weren't met and why should unmet expectations determine the way things really are? Schechtman assumes that if non-humans were to develop the cultural infrastructure necessary for person-lives, then "individuals among them who fail to possess those capacities would nevertheless be persons within their own infrastructure," (Schechtman 2014, 135) but I see no reason why that need be the case. The attitudes toward atypical species-members and the role accorded them in society could be largely different (and has been) from that accorded in our own.

Schechtman ends her discussion of developmentally atypical humans by considering a possible world in which we have the same expectations toward dogs that we actually do toward cognitively disabled humans and human infants. She thinks that such a situation is "truly unimaginable" because "it is not evident what would be involved in having those expectations of ordinary dogs. Would the birth of every dog be viewed as a kind of surprise? Would there be foundations and research initiatives to determine what could be done to prevent dogs from being born with only the capacities typical of

dogs?" (Schechtman 2014, 137) I think I can imagine such a case. I can certainly imagine (and there may even be such people) who view dogs as abominations and would set up foundations to see that they are never born. If such measures were in place, then a dog birth would be a surprise. Moreover, imagine a situation in which most dogs grew to be super-intelligent, developing forensic capacities, and so were persons under both Schechtman's criteria and my own. Regular, non-responsible dogs might then be seen in very much the same way as developmentally atypical humans (even if they were still typical dogs). The smart dogs might even campaign for the rights of their genetic brothers and sisters - that they not be bought and sold and used as servants. But they might not. They might refuse to give the atypical dogs a place in person-space or even take them be have morally relevant interests, regarding them as mere pets or worse. Nothing about the concept of a person tells us what role in society developmentally atypical members of biological species whose normal members are paradigmatic persons.

We do often use the terms 'person' and more often 'personal' or 'personality' when talking about whom we clothe and name, but we do not do so in a consistent and principled way, so that such usage muddies the task of conceptual analysis. If one is, as Schechtman seems to be, trying to account for all the ways in which the term is actually employed under a single concept, then such muddying is unavoidable. However, if one is willing to be somewhat revisionary, holding that that the ambiguities of ordinary language can be remedied when doing metaphysics by regimenting what falls under

different terms in minimally mutilating ways, then I think choosing one primary usage, i.e. responsibility, that appeals to a clear set of intrinsic features, is the thing to be done. In general, I take metaphysics to be the task of delimiting everyday concepts in such a way that ambiguity is minimized and scientific prediction and explanation is optimized. Often this requires deciding, among competing uses of a term, the ones that it is most useful for and finding other terms to cover the rest. My task is not to figure out what we call persons in all everyday contexts but to distill from such contexts a single primary usage that could be applied in many different sorts (e.g. social, biological) of possible situations.

Chapter 1: Necessary and Sufficient Conditions for Personhood

I. Personhood and Responsibility

The method of abstraction from primary use locates where a concept is most useful and then abstracts from that use the conditions necessary and sufficient for the purpose served. For the reasons noted in the introduction I take the primary use of the concept of a person to be for marking the distinction between those beings that can be responsible for their actions from those which cannot. It follows then, that necessary and sufficient conditions of personhood are necessary and sufficient conditions for being an agent such that an agent is responsible for at least some of its actions. However, this method of abstraction should not be entirely insulated from empirical data and so in distinguishing between beings that can be responsible and those which cannot, it will not do to appeal to any scientifically dubious entities or powers, such as a self that is separate from any of one's physical or mental properties, or a mysterious free will that transcends physical laws of causation. To use a philosophical buzzword, the notion of responsibility, and therefore personhood, I seek to explicate, is a *naturalistic* one – one which must be potentially explicable in the terms of our best natural science. That's not to say that I know what is potentially explicable and what isn't, but the dubious metaphysical notions I have mentioned are ones that are partially defined as being outside of the jurisdiction of scientific investigation and therefore *a priori* objectionable to a naturalist. So, given that restriction, the only notion of responsibility available for my task is one that assumes the unreality of an unobservable

self, and is compatible with the possibility of complete physical determinism. That does not mean that the account of personhood I end up with will not be consistent with the possibility of libertarian free will or a transcendent self, but only that it will not require it. The account may be neutral on the question of the existence of such occult entities or powers.

One might object that compatibilistic responsibility isn't really responsibility, or at least not responsibility in a genuinely *moral* sense - that it does not account for whether or not someone truly deserves to be rewarded or punished for what she does, because it does not answer the question of whether or not it was really "up to her" in an ultimate, buck-stopping way. I agree that in that sense, the question of responsibility is left unanswered by the compatibilist. The kind of responsibility accounted for by any compatibilist analysis is not of the sort that would, for instance, get God off the hook for the evil in the world or justify purely retributive reward and punishment, but I maintain, along with the compatibilists, that there is another, more ordinary conception of responsibility which we employ more or less consistently in our daily lives, particularly in legal contexts, that allows us to make judgments about which cases someone "was in control of what she did and understood the consequences" and in which ones she was not and did not. This sense of responsibility lies between mere *causal* responsibility of the sort that even inanimate objects (e.g. a rock being responsible for the breaking of a window) are capable, on one end of the spectrum, and the buck-stopping ultimate responsibility on the other.

The concept of responsibility I have in mind is one that is practically applicable in the absence of a solution to what we might call, borrowing somewhat from Chalmers (1996), the “hard” free will problem. It consists in the capacity to take a critical stance towards one’s own reasons for action and therefore be related to the actions themselves in a distinctive way substantially different from what most intentional beings are capable of. This sort of responsibility does not require that one have free will in the sense that one is the “ultimate cause” of one’s actions. There is a tradition, beginning with Kant, which assumes that being such an ultimate cause is necessary for responsibility. As R. Abelson (1988) puts it:

> I take it as non-controversial that the concept of a person entails the ability to perform some actions for which one can be held responsible. I also take it that responsibility for an effect entails responsibility for its cause, and non-responsibility for the cause of an effect entails non-responsibility for the effect itself. More technically, responsibility is ancestrally transitive along natural causal chains and non-responsibility is hereditarily transitive along natural causal chains. Now it follows from the principle of ancestral transitivity of responsibility and hereditary transitivity of non-responsibility that a person cannot be held responsible for any link in a causal chain unless he or she initiated that chain. (R. Abelson 1988, 75)

I do not deny that there is a coherent sense of responsibility that is transitive in the ways above described and of which one might wonder whether it is possessed by anyone. One might reasonably argue that this sense of responsibility is the only genuinely *moral* responsibility, because it is the only sort that accounts for the justification of retributive reward and punishment. I agree wholeheartedly with the premise of such an argument but am less certain of the conclusion. This is because I think there is another sense of responsibility that is not ancestrally or hereditarily

transitive, and which may have some moral import, though it does not provide a basis for the justification of retributive reward and punishment. What I have in mind is a notion that allows us to distinguish between two types of action. Both types of action may be fully determined by previous causes for which the individual could not have been responsible in any sense. However, the difference between the two types of action is as follows: On the one hand, there are actions an individual performs which are caused by her beliefs and desires, but those beliefs and desires have not been subject to internal scrutiny by the individual, such that they might have been modified by the consideration of reasons for and against having them. They are, for that reason no different in responsibility than reflexes or even the movement of the rock smashing the window. On the other hand, there are actions that are the result of desires and beliefs that have been subjected to scrutiny in the way that the first sort are not, as the individual has considered whether or not those particular desires and beliefs are themselves desirable or true. Actions one is responsible for in my sense of the term are of the latter sort. They are ones for which we may ask an individual's reasons and expect that the being in question can reflect on those reasons, and therefore, only an individual with the capacity to so reflect is one that is fit for trial or interrogation. Furthermore, only such an individual is fit to give consent and make legal commitments. That one is so responsible for some actions does not by itself entail that one should be rewarded or punished for such actions, although it might be the case that some rewards and punishments are only effective on an individual with the capacity for responsibility.

Furthermore, that an individual is capable of responsible action, in my sense, implies that the individual is capable of *taking* responsibility, at least implicitly, for her actions. I don't mean that such an individual necessarily "owns up" to actions for which she is responsible, admitting culpability, promising to redeem herself, etc., but only that such an individual recognizes (at least to herself) that she is responsible for the action in question. Such implicit responsibility-taking may be morally relevant in the sense that one who recognizes her own responsibility has morally committed herself to taking responsibility in the more explicit sense of "owning up," whether or not she does actually make good on that commitment.

One important worry that needs to be addressed here is that there might be many cases in which an individual's self-scrutiny is partially or wholly determined by some alien power, for instance, post-hypnotic or subliminal suggestion or even direct neural manipulation by a mad scientist. In such a case it might be inappropriate to say that an individual's actions, influenced by such ministrations, are ones for which the individual is responsible. However, while I agree that one may not be responsible for such actions, this poses no real objection to my analysis of responsibility in terms of self-scrutiny. An individual capable of self-scrutiny is one with the capacity for responsible action, but such an individual might fail to be responsible for any particular action resulting from self-scrutiny if such scrutiny is influenced by an alien force. The capacity for self-scrutiny is what grounds the treatment of an individual as a responsible being and hence a person, but does not entail that every action performed by that

being, even if resulting from self-scrutiny, is an action that being is responsible for.

If after considering the above points, one still thinks it inappropriate to call the notion I have been describing “responsibility,” then a perhaps more palatable alternative would be ‘authorship’. There are many beings who can perform actions, but only persons can be genuine authors of their actions, by acting in a way that results at least in part from the consideration of reasons. I think ‘authorship’ basically gets the same point across, but prefer ‘responsibility’ because I think the analogy between performing reason-sensitive actions and authoring a work of literature is imperfect. On might author a rhapsodic poem that involved no reflection upon reasons for writing the lines that were produced. Furthermore, ‘responsibility’ fits more naturally with considerations of persistence. A person can be “responsible for something done in the past.” An analogous formulation of that phrase involving authorship would be awkward at best.

Harry Frankfurt (1971) articulates a compatibilist notion of responsibility along the same lines as the one just sketched, that he also takes to be distinctive of persons. According to Frankfurt, while there are relatively many kinds of beings that possess desires, only persons have, in addition to their first-order desires, second-order desires, the content of which are the first order desires. In other words, not only do persons *want-to x,* but they may or may not *want-to want-to x*. For instance, I may have a first-order desire to smoke, but because I value my health I may want that I did not have that desire. Frankfurt writes:

> It is my view that one essential difference between persons and other creatures is to be found in the structure of a person’s will. Human beings are not alone in

> having desires and motives, or in making choices. They share these things with the members of certain other species, some of whom even appear to engage in deliberation and to make decisions based on prior thought. It seems to be peculiarly characteristic of humans [assuming that no member of another species is in fact a person], however, that they are able to form what I shall call 'second-order desires' or 'desires of the second order.' Besides wanting and choosing and being moved *to do* this or that, men may also want to have (or not to have) certain desires and motives. They are capable of wanting to be different, in their preferences and purposes, from what they are. (Frankfurt 1971, 323)

A second-order volition is a second order desire that one or more of one's first order desires be effective in one's actions -- that they constitute one's 'will'. The capacity to form second-order volitions, for Frankfurt is sufficient for the capacity for responsible action, but for any individual action of a person to be one she is responsible for requires "harmony" between her second order volitions and the first order desires that are effective in that action. When you want to have the will that you have, you are responsible for what you will. For Frankfurt, the capacity for second order volitions (harmonious or not) marks the difference between persons and mere animals or 'wantons' who have no attitudes towards what they do or do not will.

While I take it to be correct in its general outlook, I depart from Frankfurt's view in a few ways. First of all, he denies that there is a morally interesting conception of free will that does require one to be the ultimate cause of one's actions, whereas I do think there is such a conception, I just don't think it corresponds to reality. I agree in general with compatibilism that there is a sense of responsibility that is compatible with determinism. However, I disagree that it is the only coherent or interesting sense of responsibility. The incompatibilist notion of free will is coherent and morally relevant,

though I think it is false concerning beings in the actual world.

Secondly, Frankfurt's account of human motivation is rather oversimplified. Most conflicts of desire and volition are more complicated than simply wanting or not wanting to have another desire. To begin with, one might both want and not want to want a desire to various degrees. For example, I may be pleased that my previous first-order desire to smoke has dissipated because I value my health, though a part of me wishes I still had the desire because I have some nostalgia for my youthful devil-may-care attitude and lifestyle. From day to day and even moment to moment I may vacillate in what I consider my "real" desires. Beliefs in the desirability of some desire or other play a role in whether or not one wants to have them and those beliefs may also be held with differing degrees of conviction. On days when I can go running without wheezing and gasping for breath, my pleasure in doing so may buttress my desire not to desire to smoke because it has provided new evidence for believing that the desire to not want to smoke is my true want. On other days when I've had a few drinks with my friends and they go out for a puff, that conviction may be weakened - though I no longer have the first-order desire to smoke, I may wish, for that moment, to still have that desire. One may also be more or less aware of one's motivations, and more or less self-deceptive. It is possible that a person believes that he or she wants to want something and yet not really want it "deep down" or "high up". For instance one might wonder whether the dictates of one's conscience, constituted by desires to want or not want something, are not merely conventional mores that one has internalized without wholeheartedly

endorsing them. Furthermore, one can be committed to an action in such a way that one does it despite retaining some pull of desire to refrain, as may be the case for an individual who must put to sleep a terminally ill and suffering animal.

The condition of harmony between levels is too strong a condition of responsibility, for one is almost never entirely volitionally harmonious. However, one need not provide a more complex necessary relation in its place, for we can say that to be capable of responsible action one must only be able to judge reflectively one's first order desires and act according to those second order judgments whether or not one is conflicted. I suggest that such reflective judgment requires two distinct capacities which may exist independently of one another, but which must be possessed in concert for an individual to be a person. I refer to them as 'self-consciousness' and 'concern': i.e. a. awareness of the motives behind one's actions and b. emotional investment in the satisfaction of one's desires and truth of one's beliefs, at least insofar as they contribute to the satisfaction of one's desires.

Again, I do not purport to establish conditions that would make an individual responsible for any particular action, but rather the conditions a being must meet in order to be capable of performing some actions for which he or she is responsible. Schechtman (2014) seems to have this difference in mind when she distinguishes between a person as a 'forensic unit' and as a 'moral self.' The former term refers to the sort of being that is an appropriate object for inquiry about responsibility, or as Schechtman puts it, "a suitable target about which particular forensic questions can be

raised and judgments made" (Schechtman 2014, 14). The latter notion construes the person as only the performer of the actions he or she is actually responsible for, his or her limits "set by the very actions and experiences for which she is in fact held rightly accountable." (Schechtman 2014, 15) My conception of persons is as forensic units, beings who are sometimes responsible for their actions, not as 'moral selves,' because I take persons to exist and be persons even when performing those actions for which they are not responsible. For example if a person enters a temporary state of fugue in which he is no longer able to reflect on his reasons for action, the person does not cease to exist, but is merely temporarily unable to exercise his person-constituting capacities.

The general picture of persons as responsible beings is also not far from Locke's early characterization. He understood 'person' as primarily useful in tracking responsibility. Furthermore, while Locke explicitly defined persons as conscious, rational beings, his discussion of the relation between persons and responsibility shows that the kind of consciousness he was talking about is closer to what is now often referred to as 'self-consciousness' and that some type of 'concern' is also essential to the 'forensick' conception of personhood.

> *Person*… is the name for this *self*. Where-ever a Man finds, what he calls *himself*, there I think another may say is the same *Person*. It is a Forensick Term appropriating actions and their Merit; and so belongs only to intelligent Agents capable of a Law, and Happiness and Misery. This personality extends it *self* beyond present Existence to what is past, only by consciousness, whereby it becomes concerned and accountable, owns and imputes to it *self* past Actions just upon the same ground, and for the same reason, that it does the present. (Locke 1690 II. XXVII. 17)

My view of personhood, therefore, is essentially an updated and refined version of Locke's conception in light of subsequent developments in philosophy as well as psychology and other empirical areas of inquiry. I argue that the capacities for self-consciousness and concern are necessary and jointly sufficient for personhood. In unpacking what I mean by those terms and how they relate to one another I will depart significantly from Locke and my understanding of/suggestion for the relation between the concepts of 'person' and 'self' is a rather significant departure from Locke given the special meaning I assign to the latter term (in chapter four). Nevertheless, Locke's view provides a helpful outline for a model of personhood, though the details must be filled in rather differently than he had done. One great advantage of this view, which Locke failed to capitalize on is that it allows for a unified treatment of personhood and personal persistence over time, which will be explained in chapter two.

II. Self-Consciousness

Self-consciousness is the capacity for critical awareness of the reasons that lead to one's actions and recognition of those reasons as one's own. This capacity has long been thought central to the notion of personhood, and rightly so. Self-conscious beings can project their desires and beliefs into the past and future and so can be aware of the potential and actual consequences of their own actions relevant to those intentions. Responsible beings can be reasonably praised or blamed for at least some of their actions. One can only reasonably praise or blame, rather than merely scold or stroke,

someone who has the capacity to be aware of the desires and beliefs that lead to her actions. Only then can she be responsive to praise and blame, and either endeavor to modify her actions, attempt to restrain herself from acting in accordance with them, or conceal them and the actions performed on the basis of them from others. For that reason, self-consciousness is necessary for responsibility, and therefore personhood.

One focused analysis of the capacity for self-consciousness and its relation to personhood is contained in Daniel Dennett's "Conditions of Personhood." (1978) Dennett, following Frankfurt's intuitions about what distinguishes persons from non-persons, explicates a notion of self-consciousness that I will argue is necessary, though not by itself sufficient (for reasons other than Dennett's own) for being a person. It is helpful to look closely at Dennett's account because it explores the relations between self-consciousness and various other conditions that are often thought to be criteria of personhood, though I disagree with him on a few specific points that together form a significant difference in general outlook. Dennett introduces six 'themes' that he thinks are generally considered necessary conditions of being a person. He then goes on to explain the relations of dependence between them and to explore the question of why they are indeed necessary and whether or not they are together sufficient for personhood. Dennett understands 'person' to be both a metaphysical and moral concept and wonders whether the two concepts *coincide*, eventually deciding that while they are not really wholly distinct concepts, metaphysical personhood is necessary but not quite sufficient for moral personhood, so that they are "two different and unstable

resting points on a continuum." (Dennett 1978, 284) I will argue in chapter five that metaphysical personhood is not only insufficient for being a person in the full moral sense, but that it is not even necessary for some aspects of moral personhood. For now, I will put the morality to one side and focus on the metaphysics, specifically the explication of Dennett's themes and how they are related.

Dennett's six themes are: rationality, intentionality, being treated as something with intentions, reciprocity of that attitude toward others, verbal communication, and consciousness, or more specifically 'self-consciousness' of the kind that is the focus of this section. According to Dennett, the first three themes come together as one package, as possession of each implies possession of the others. Here Dennett appeals to the main line of argument from his earlier article "Intentional Systems" from the same volume (further developed in his later book *The Intentional Stance* (1987)), asserting that whether or not a being has intentionality, i.e. whether or not its behavior results from having beliefs and desires, is entirely a question of whether or not it is useful to explain its behavior by using intentional language. In other words, there is no fact of the matter about whether or not something is intentional beyond its being usefully treated as such. And its being usefully treated as such, is for Dennett, a matter of whether its behavior appears to be rationally directed toward an end. For instance, in the case of a chess computer:

> By assuming the computer has certain beliefs (or information) and desires (or preference functions) dealing with the chess game in progress, I can calculate - under auspicious circumstances - the computer's most likely next move, provided I assume the computer deals rationally with these beliefs and desires. The

> computer is an Intentional system in these instances not because it has any particular intrinsic features, and not because it really and truly has beliefs and desires (whatever that would be), but just because it succumbs to a certain stance adopted toward it, namely the Intentional stance, the stance that proceeds by ascribing Intentional predicates under the usual constraints to the computer, the stance that proceeds by considering the computer as a rational practical reasoner. (Dennett 1978, 271)

Dennett tells a plausible story about what sort of behavior should lead one to decide whether or not a system is intentional, but his extreme suggestion, that there is no fact of the matter about whether or not a being *really* has the beliefs and desires that it seems to, is less convincing. To begin with, one might object that there is a phenomenal experience associated with having intentions that a goal-directed being may or may not really have. However, to show that this is a real problem the objector must justify the premise that there actually is phenomenal feel that comes with having intentions. The objector is more likely to succeed with desires than with beliefs, but in any case, when one is invoking phenomenal experience, one is talking about *conscious* beliefs and desires, which is another matter from just ascribing intentions alone.

Another more serious doubt comes from the fact that beings can be deceptive, so one might doubt whether a being has the *particular* desires and beliefs that it seems to. However, to be deceptive one must have *some* intentions, so the possibility of deception raises no real problem for taking goal directed behavior to be evidence of intentionality in general. Still, if one can be perfectly deceptive, such that one's behavior, despite one not having a desire or belief, is impossible to distinguish from that of someone that does have it, then observable behavior cannot be all there is to say about

particular intentional states.[8] Furthermore, besides deceptive cases there seem to be other differences in intentional states that are behaviorally indistinguishable. For instance two people may pursue the same goal by identical means but due to entirely different motives. Failure to take these sorts of cases into account makes Dennett's extreme suggestion implausible in the same way as it does classical behaviorism.

Nevertheless, despite the dubiousness of the extreme suggestion, Dennett is mostly right about the *evidence* available to us for ascribing intentionality. For him, any kind of goal directed behavior is good enough reason for ascribing first-order, non-conscious intentional states to a being and I agree on this point as far as beliefs go, though not in the case of desires, because, as I will argue in the next section, their ascription additionally requires that the being in question demonstrate *concern*. Regardless, on this view, all kinds of non-persons, from computers to dogs and cats, can reasonably be ascribed intentional states.

Meanwhile, Dennett may be right that one cannot reasonably attribute intentions to something without assuming that the being in question is rational in a minimal sense of the term. However, it could be the case that an individual has irrational intentions, but that no observation of its behavior (other than spoken avowals) would be evidence of them, so that one's answer to the question of the possibility of irrational intentions depends on whether or not one accepts Dennett's view about the vacuity of the doubt

[8] To be sure, Dennett holds, plausibly, that genuine deception requires second-order intentions, as will be explained below, so the objection from deception may only require that there be a fact of the matter about whether or not someone has second-order intentions. Still, it would be strange if the question of the reality of intentional states only came in at the higher order.

about the ontological status of intentional states. So while I take both intentionality and minimal rationality to be necessary features of persons, I remain agnostic about whether there is any strong dependence between them. Either way, being a person requires being an intentional system, though not all intentional systems are persons.

The fourth theme in Dennett's account is 'reciprocity.' Here he suggests that it is not enough for a system to be a person that it be intentional, but it must also be capable of recognizing other systems as intentional. Dennett defines a 'second-order intentional system' as "one to which we ascribe not only simple beliefs, desires and other Intentions, but beliefs, desires, and other Intentions about beliefs, desires, and other Intentions," (Dennett 1978, 273) However, I prefer to reserve the appellation 'second-order intention' for beings that are aware of their own first-order intentions, so as to distinguish *mindreaders* - those able to reciprocate the intentional stance by attributing beliefs and desires to others, from what I take to be a different sort of second-order intentional systems - those capable of self-reflection or metacognition, having second-order desires and beliefs that target their own first-order desires and beliefs. The ability to attribute intentions to others is an ability distinct from attributing intentions to oneself. Tests (e.g. as suggested by Lurz 2011a&b) that purport to demonstrate whether or not chimpanzees attribute desires and beliefs to others, do not by themselves provide conclusive evidence for whether or not the chimps are self-reflective as well, nor does evidence of self-reflection entail evidence of attributing mentality to others. Further, I don't see why a being must be able to attribute intentions to others in order to be a

person. It does seem to be a contingent fact about human beings that we are unable to develop the self-reflective states without growing up in a community of persons and treating other members of that community as persons. And attributing intentionality to others may have been evolutionarily prior to self-attribution. But neither point makes it inconceivable that there could be a self-reflective being that does not, or cannot recognize intentions in others. In fact, some autistic individuals are claimed to be that way. (Baron-Cohen 1997) Their being so does not imply that they are not persons.

The distinction between the two kinds of second order intentionality becomes crucial as Dennett goes on to claim that "genuine self-consciousness" (the sixth theme,) is not just a matter of self-reflective second-order intentionality, but also requires the capacity for verbal communication (the fifth). Here he invokes the Gricean account of verbal communication, which has as a necessary condition not just second-order intentions, but third order intentions, not of the self-reflective sort, but of the mindreading variety. According to Dennett, for Grice, this is because to successfully communicate one must intend that another person understand one's intentions.[9]

Being a person, i.e. a responsible being, requires self-consciousness, for Dennett, because for a being to be responsible for its actions, it must engage in the kind of "reflective self-evaluation" described by Frankfurt (Dennett 1978, 284). But to do this one's intentions cannot be merely implicit, rather one must consciously entertain and

[9] However, it is not clear that Grice thinks of such intentions as higher-order and not just first order and self-referential.

decide upon one's reasons for action. Unlike attributions of intentionality of the first order, it is not enough that one's actions show "an order which is there." (Dennett 1978, 284) And the only way of demonstrating that one is capable of such conscious reflection is via verbal communication. Dennett writes:

> *If I am to be held responsible for an action* (a bit of behavior of mine under a particular description), I must have been *aware* of that action under that description. Why? Because only if I was aware of the action can I say what I was about, and participate from a privileged position in the question-and-answer game of giving reasons for my actions. (If I am not in a privileged position to answer questions about the reasons for my actions, there is no special reason to ask *me*.) And what is so important about being able to participate in this game is that only those capable of participating in reason-giving can be argued into, or argued out of, courses of action or attitudes, and if one is incapable of "listening to reason" in some matter, one cannot be held responsible for it. The capacities for verbal communication and for awareness of one's actions are thus essential in one who is going to be amenable to argument or persuasion, and such persuasion, such reciprocal adjustment of interests achieved by mutual exploitation of rationality, is a feature of the optimal mode of personal interaction. (Dennett 1978, 282)

However, even if there is no way someone could *demonstrate* his capacity for self-consciousness without some verbal ability (by which I, and I suppose Dennett as well, mean to include writing and sign language), that does not mean someone can't *be* self-conscious - that is, aware of his first-order intentions, without being capable of communicating those intentions to someone else. The relationship between thought and language is a contentious issue in the philosophy of psychology. Jerry Fodor's claims about the innateness of concepts, which are tangential to his "Language of Thought" hypothesis, are extremely contentious, but it is a viable theory about some of the ways thoughts are structured and related to one another and how at least purely deductive

reasoning works. (Schneider 2009) It is plausible that if not all mentality, then at least consciousness with respect to the intentional content of mental states require some kind of language, (as Rosenthal 2005 suggests) and when one gets to evaluative self-consciousness of the sort I take to be necessary for personhood, the argument for requiring language is even more compelling. One reason for this is that conscious introspection (i.e. silent communication with oneself) seems to involve 'listening' to one's 'inner speech' and brain imaging studies show activity in the regions of the brain responsible for verbal communication during such activities (Prinz 2012, 159), so that some form of linguistic representation does seem to be necessary for self-consciousness.

While I am in agreement with Dennett that sensitivity to reasons and awareness of one's reasons for action are necessary for responsibility, and therefore personhood, I do not think it inconceivable that those reasons could be confined to the mind of an individual agent. Being conscious of one's intentions likely requires something like language, or a language-like scheme of representation, but not necessarily the ability to communicate those representations to others. Again, it is likely the case that no human being could develop self-consciousness without interacting with other individuals, specifically persons, and indeed Nietzsche may have been right in attributing the original source of human self-consciousness to societally imposed internalization of value judgments or to the need to communicate one's thoughts for deceptive or cooperative purposes (Nietzsche 1887/1967), but that does not rule out the conceptual

possibility of a being that is an island of self-consciousness unto itself. Despite the social influence on the formation of one's character and the external origin of one's values, one can be self-conscious without there being anyone else around with whom to communicate. For this reason, it may not be necessary that one have the capacity for verbal communication in order to be self-conscious. However, it is plausible that such communication is necessary to provide *evidence* of self-consciousness.

At this point I need to explore the relationship between self-consciousness and consciousness-simpliciter, or if there is no such notion, then one or two other senses in which philosophers use the term 'consciousness'. There are quite a few theories of consciousness on the market these days, and this is not the place to go over all the various arguments for or against each, though I must say something about how well some of the views mesh with how I understand self-consciousness. It may seem as if one genus of views about consciousness, the higher order theories, the most prominent of which is David Rosenthal's higher order thought (HOT) theory, fits most neatly with what I have said about self-consciousness. The HOT theory claims that for a mental state to be conscious one must have a second-order mental state that takes that first-order state as its content, thereby making the first order state conscious. However, some writers have objected that while this might account for a certain kind of consciousness, perhaps "access consciousness" (Block 2002) or "transitive consciousness," (Mandik 2013) it doesn't explain basic phenomenal consciousness, or

'what it's like' to be something with consciousness (Nagel 1974).[10] This objection has particular force if one wishes to claim, as I do, that animals such as dogs, cats and elephants aren't persons by virtue of the fact that they lack self-consciousness (as I will argue later on, the other capacities constitutive of persons are ones that are shared to some degree with animals) because it now seems wrong to deny, as Descartes did, that those animals lack any kind of conscious awareness whatsoever. In 2012, a group of scientists signed a declaration to the effect that a wide range of animals have consciousness. The declaration states that birds, for example,

> appear to offer, in their behavior, neurophysiology, and neuroanatomy a striking case of parallel evolution of consciousness. Evidence of near human-like levels of consciousness has been most dramatically observed in African grey parrots. Mammalian and avian emotional networks and cognitive microcircuitries appear to be far more homologous than previously thought. Moreover, certain species of birds have been found to exhibit neural sleep patterns similar to those of mammals, including REM sleep and, as was demonstrated in zebra finches, neurophysiological patterns, previously thought to require a mammalian neocortex. Magpies in particular have been shown to exhibit striking similarities to humans, great apes, dolphins, and elephants in studies of mirror self-recognition. (Low et al. 2012)

And it concludes:

> We declare the following: "The absence of a neocortex does not appear to preclude an organism from experiencing affective states. Convergent evidence indicates that non-human animals have the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states along with the capacity to exhibit intentional behaviors. Consequently, the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Nonhuman animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates."

[10] Prinz holds that any kind of consciousness that does not involve what-it's-like-ness, is not worthy of the term 'consciousness.' Prinz 2012 (35)

In denying that dogs and cats are self-conscious I do not mean that they are incapable of affective states, nor that they lack consciousness of some sort. One way to accommodate the claim that *mere animals* are not persons primarily in virtue of the fact that they lack self-consciousness with their possessing basic phenomenal consciousness, is to understand the distinction between the two kinds or levels of consciousness in terms of the presence or absence of higher order thought: the basic phenomenal kind that animals possess and that could be accounted for with a theory that restricts itself to first-order mental states versus the self-consciousness distinctive of persons that requires higher-order thought. However, if one endorses the HOT theorist's view that all consciousness requires higher-order thought then one must demand more than just higher-order thought for self-consciousness. Rosenthal himself does not deny consciousness in "non-linguistic" animals. He thinks they may have HOTs, however crude, but "the HOTs of such beings would not result in their being conscious of their mental states in the rich way we're typically conscious of ours… Moreover, the HOTs of creatures without language might never make them conscious of their mental states in respect of the intentional content of those states." (Rosenthal 2005, 6) So if one can distinguish between rich and impoverished varieties of higher-order thought, then one can maintain that higher order thought is necessary for consciousness and that animals are conscious, but that persons are distinctive in having the capacity for a richer kind of higher-order thought, one that makes one conscious of one's first order states in respect of their intentional content. As alluded to

parenthetically above, Rosenthal otherwise suggests that the difference might be that only persons have third-order thoughts. He writes:

> We are the only creatures we know of that we regard as persons, but we can easily imagine discovering others that we would classify with ourselves in that way. It is not, of course, that only persons have conscious mental states; many non-human animals presumably do, as well. There is no reason to deny to animals without language the capacity to have suitable higher-order thoughts. The relevant higher order thoughts do not require much richness of conceptual resources or syntactic structure. But we have no reason to suppose that animals other than persons are aware of whatever higher-order thoughts they may have. And if none of an animal's higher-order thoughts are conscious, it will lack the particular kind of reflective consciousness that involves some measure of rational connectedness in the way it is aware of [the awareness of][11]its mental states. (Rosenthal 2005, 146-7)

I don't see any *a priori* reason for ruling out non-linguistic higher-order thoughts of a crude sort, particularly if a clear distinction can be drawn between "conscious self-referential thinking" and "sensory metacognition." The latter, according to Lau and Rosenthal (2011) might better reflect higher order representations in animals that possess a pre-frontal and parietal cortex, "for instance the ability to rate confidence appropriately to distinguish between correct and incorrect perception". One researcher, J. David Smith has run numerous studies testing various animals for metacognition, as he defines it, aka the capacity for monitoring or regulating their own cognitive states. One type of study tested whether or not animals would reject a task that was presented to them, knowing that the task was difficult or impossible for them.

[11] Given, as will be explained below, that Rosenthal doesn't think consciousness plays a very important role in behavior or in the relations between first order mental states, I found this last part of the passage odd. In personal communication he agreed that it was a bit misleading as stated and suggested that the text in brackets be added to make it a clearer expression of his view.

> First, animals were given difficult perceptual discriminations: the difficulty potentially created uncertainty in their minds. Second, animals were given an additional response—beyond the discrimination responses—with which they could decline to complete any trials of their choosing. This response—sometimes called the Uncertainty Response (UR)—allows animals to report on, or cope with, the difficulty. If animals monitor cognition accurately, they should prospectively recognize difficult trials as error-risking and decline those trials selectively. (Smith 2009, 389)

A study of this kind done with macaque monkeys indicated that the monkeys declined difficult tasks with a regularity strongly isomorphic to the results of human trials. In the human cases, participants reported that their UR reflected their own conscious uncertainty, suggesting that the monkeys too were aware of their own risk of error. Some of the tasks tested the monkey's memory for previously presented stimuli so that the monkeys' UR response suggested that they "judged the robustness of internal memory representations." (Smith 2009, 391) This aspect of the data is particularly important, because judgments of present stimuli are much easier to explain without attributing metacognition then are judgments of absent stimuli (though there are paradigmatic cases of metacognition in humans involving judgments of present stimuli. e.g. wavering over multiple choice options in an exam.) According to Smith, the studies show that the "minimum cognitive sophistication" that can be attributed to animals who demonstrate UR responses is "a controlled decision, on the threshold of perception or memory… involving controlled cognitive processing… at difficult decisional choice-points." (Smith, 394) Pigeons and capuchin monkeys, on the other hand, showed limited evidence of URs, despite neither species being "behaviorally or associatively challenged." The metacognition that macaques, and also dolphins (Smith et al., 1995)

are seemingly capable of then seems to be a benchmark of cognitive sophistication. In the absence of linguistic competence these kinds of tests, combined with homologies in neural architecture may be the best evidence available for the presence of metacognitive capacities in animals.

So if macaques are capable of this kind of metacognition and have phenomenal consciousness, does it mean that they are also self-conscious? This question brings us back to Dennett's claim that the kind of reflective evaluation necessary for responsibility, in addition to higher order intentions, must involve genuine consciousness of one's intentions and reasons for acting, not just "an order that is there." For him, nothing short of first-person verbal avowals could count as evidence of such consciousness. Some chimpanzees who have been trained in sign language have behaved in ways that could be interpreted as just such evidence.[12] But on the HOT theory all there is to a state's being conscious is that there is a higher-order thought about it. Still, self-consciousness, or, for Rosenthal, "self-referential consciousness," could additionally require that the higher order thought itself be conscious by virtue of being the content of a third-order thought.

Now, if it turns out that some non-human animals are self-conscious, that doesn't annihilate the distinction between persons and mere animals. It could turn out that there are just more persons than we originally thought. For instance, the more we progress in teaching chimpanzees sign language, the more astute they become in communicating

[12] Peterson (2011) describes a case of a chimp lying about having defecated on the floor, implicating others, before finally giving up and admitting that it was her own doing.

their motives, the harder it will be to justify considering them non-persons. Again I find it unproblematic to view personhood as a matter of degree -- it seems that it must be so if we are to assume that self-consciousness and concern are traits that have evolved, since evolutionary changes are incremental. While I am agnostic about the necessity of third-order thought for self-consciousness, I am sympathetic to Rosenthal's caution in declaring a being completely HOTful or HOTless, and his openness to multiple ways of drawing the lines. As he puts it "Being a person, on this account, may be a matter of degree, but that is as it should be. Our distant ancestors doubtless had the distinctive characteristics of people to some degree, though not as fully as we do, and the same may be true of other creatures elsewhere." (Rosenthal 2005, 147) Rosenthal also uses third-order thoughts to account for the experience of the unity of consciousness (2003). I am sympathetic to HOT theory as an account of consciousness due to its elegance and explanatory power, and furthermore I find the account of unity it offers compelling as well as useful component in the conception of the self that I offer in chapter four. However, I take my conception of self-consciousness to be compatible with other higher order views of consciousness as well as first-order accounts. The kind of self-consciousness necessary for personhood requires consciousness and higher-order representation. On higher order views, consciousness and higher order representation are conceptually inseparable -- they are one and not two capacities, while on a first-order view of consciousness, such as Michael Tye's PANIC (Poised Abstract Non-conceptual Intentional Content) view (presented in Mandik 2014), where what makes

the difference between conscious and non-conscious thoughts is not the presence or absence of higher-order thoughts, but rather differences in the content of the thoughts themselves, consciousness and higher-order thought are distinct, separable (at least conceptually) capacities, though both are required for self-consciousness.

I have so far argued that self-consciousness, of the sort that involves consciously self-reflective second order desires and beliefs, is necessary for personhood. Next I will engage the question of whether or not it is also sufficient. Dennett provides his own reasons for thinking that it is not, which are connected to the question of the relation between metaphysical and moral personhood, and which I will discuss in chapter five. However, I have my own objection to the claim that self-consciousness is by itself sufficient for personhood, which is that to be genuinely reflectively self-evaluative in the way required for responsibility and, therefore, personhood, one must not only be consciously aware of one's desires and beliefs, but must also be concerned with whether or not one's desires are fulfilled or one's beliefs (at least those relevant to the satisfaction of one's desires) are true. A self-conscious, but wholly unconcerned individual would not be a person. I the next section I will explain exactly what I mean by 'concern', why it is conceptually distinct from self-consciousness and why it is necessary for personhood.

III. Concern

Imagine a self-conscious humanoid robot. It is aware of its first order intentions: its desires (in the minimal sense of goals to be attained) and its beliefs about the world relevant to satisfying those desires. However, the robot under consideration doesn't genuinely care whether or not those goals are fulfilled. It is not emotionally *invested* in them. Failure does not frustrate the robot, nor does success gratify it. It may even be able to change its goals given repeated failures that cause damage to it, or change its beliefs about how those goals are to be attained (as connectionist neural networks seem, in a crude way, to be able to do), but the damage incurred from the failures does not cause any feeling of sadness, anger or shame. And successes, even ones that greatly benefit the robot (such as saving its spare battery from being exploded by a bomb[13]), do not give the robot a sense of joy, accomplishment or relief. What such a robot is lacking, is what I mean by 'concern,' i.e. a range of affective investment in the attaining of one's goals and the truth of one's beliefs insofar as they are relevant to one's goals. By virtue of this lack of concern such a robot cannot be responsible for its actions and therefore, is not a person.

Speaking more strictly than I have until now, having a genuine desire for something, as opposed to merely being directed toward a goal, requires being *concerned* that some goal be attained. A self-conscious chess computer would be capable of representing its goals to itself so as to take a higher-order attitude toward

[13] As with the elusive R2D2, prospected in Dennett (1984)

them and will that they be enacted, believing that it has a desire to win, regardless of whether or not it is concerned that it win. But in order for it to really have such a desire it must have some emotional investment in winning or losing. The benefits of winning or costs of losing must be emotionally significant to the being itself. They must be tied to its emotional responses, such that it can become frustrated when it doesn't get its way, surprised when things turn out to be different than they had seemed, or satisfied and reassured if things go well and/or as predicted. Having a desire is not just being directed toward a goal, but requires being genuinely *motivated* toward attaining a goal. A torpedo may modify its behavior in reaction to feedback about its success or failure in attaining a goal, but it gives no indication of appreciating the circumstance as satisfying or frustrating for itself. The torpedo does not fear failure, nor does it hope for success. It cannot feel happy or sad for what it does or what it is. In that sense, it lacks the capacity for concern.

Martin Heidegger makes much of the German terms '*Sorge*' in his philosophy, which is often translated as 'care,' and '*besorgen'* which is closer to 'being concerned with' and *Fürsorge* which is specifically care or concern for others. For Heidegger, *Sorge*, *besorgen*, and *Fürsorge* describe one way of "being-in-the-world," part of the general condition of human beings, or "Dasein." Inwood's *A Heidegger Dictionary* (1999) explains *Sorge* and *besorgen* as follows:

> *Sorge*, 'care', is 'properly the anxiety, worry arising out of apprehensions concerning the future and refers as much to the external cause as the inner state' (DGS, 56)... *besorgen* has three main senses: (a) 'to get, acquire, provide' something for oneself or someone else; (b) 'to attend to, see to, take care of'

> something; (c) especially with the perfect participle, *besorgt*, 'to be concerned, troubled, worried' about something. The nominalized infinitive is *das Besorgen* , 'concern' in the sense of 'concerning oneself with or about' something. 3. *Fürsorge*, 'solicitude', is 'actively caring for someone who needs help'. These three concepts enable Heidegger to distinguish his own view from the view that our attitude towards the world is primarily cognitive and theoretical. Descartes's and Husserl's 'concern for known knowledge' (*Sorge um erkannte Erkenntnis*) is only one type of concern, and not the primary, or a self-evidently appealing, type (XVII, 62; LXIII, 106).

The third sense of *besorgen*, is closest to 'concern' in the way I have defined it. The explanation provided of the role the term plays in Heidegger's philosophy, distinguishing his view about the way we are engaged in the world, from the purely "cognitive and theoretical" conception of earlier philosophers, fits well with my contention that persons are essentially emotional in addition to being self-conscious. However, the distinction for me is not in terms of the kind of attitude we take toward the world, or *what* we're concerned with when we inquire into it, but rather that we are concerned at all about the satisfaction of our desires or truth of our beliefs and not just receivers and manipulators of information.

In popular discussions and depictions of artificial intelligence (as found in countless science fiction stories) automata are described as "doing only what they were programmed to do," "not possessing a will *of their own*," etc. What seems to be expressed in these sorts of phrases is that part of what a *mere robot* is lacking, which would be necessary for it to be a person, is something like free will. Now, one might be tempted to explain what the will-less robot is lacking in cognitive terms that appeal to something like self-consciousness. However, this approach does not match the

depiction of automata in many of these informal treatments. In science fiction, computers and robots can be depicted as hyper intelligent, hyper-rational and completely self-conscious, but beings for whom praise and blame would be meaningless due to their lack of emotional investment. The ship's computer in *Star Trek: The Next Generation* is one example. It is presented as entirely aware of its own beliefs (more so perhaps than an ordinary human, because nothing is repressed or rationalized away). It has goals in the form of operations it is commanded to perform and ones it performs based on its own judgment of a situation, and can report on those goals. However, nothing that it does or that happens to it has any emotional impact on it. Holding it accountable, praising or blaming it, rewarding or punishing it, would be pointless, because it is not concerned about the satisfaction of its desires or truth of its beliefs. Without such concern, it cannot be a person.[14] In contrast, there is another character from *Star Trek: The Next Generation*, the android Data, who most would agree is a person, though that very question is debated, in a fairly philosophically sophisticated manner for television, in the episode "The Measure of a Man." There, it is Data's enjoyment in possessing his medals, his friendship with Commander Ryker, his possessiveness of the book Captain Picard gave him and of his memento of love interest Tasha Yar, and his distress that he could lose all the memories he cherishes if

[14] Marvin the Depressed Robot from Douglas Adams' *The Hitchhiker's Guide to the Galaxy* seems robotic despite his depression, because he has no *range* of emotions. He can't not be depressed. Nothing can affect his emotional condition, so he may not be responsive to praise and blame in the way required for responsibility. R2D2 and C3P0 from *Star Wars* seem to be self-conscious and concerned. It is only their mechanical appearance that marks them as robots.

he is taken apart, that convinces us he is a person.[15] That is because having a "will of one's own," in the sense relevant to the folk distinction between persons and mere automata, means, in part, having not just 'preference functions', but genuine desires, which require affective engagement. (The other part being, as Frankfurt first pointed out, and was elaborated on in the previous section, what distinguishes mere animal will from the will of persons: self-consciousness.) At one point in the episode, the character Bruce Maddox who wishes to take Data apart in order to study him, but is faced with resistance from Data and his comrades, inquires, for comparison, whether they would allow the ship's computer to refuse a refit. However, such a question is moot, because the ship could and would never make such a refusal, because it has no concern for its own well-being. It might caution the crew that doing so would be dangerous for their own welfare, but it would not care if they decided not to take its advice.

If one is unimpressed by these science fictional examples of the self-conscious but unconcerned, there are some actual cases to consider. Some individuals have experienced complete apathy or "akinetic mutism" which is associated with damage to or sectioning of the cingulate cortex. (Prinz 2012, 42 and 67) Such individuals are "mute, inactive, and utterly lacking in motivation but nevertheless perceive the world

[15] Though as Anderson (2000) notes, the official solution to the legal question of Data's personhood, as argued in the episode by Captain Picard, and similar to Putnam's (1964) assessment of the issue, is a kind of Pascal's Wager, namely that the consequences should we grant something personhood and be wrong are much less awful than the potential for evil wrought from wrongly denying a being and, worse, class of beings personhood. However, this solution wrongly assumes that to deny a class of beings personhood is to deny them the rights of life and fair treatment. On the other hand, concern on its own may be a ground for assigning those rights so that even if it was decided that Data was not a person due to not possessing self-consciousness one could still claim on his behalf a right to not be dismantled.

around them." (Prinz 2012, 41) After recovering from her condition, a patient of Antonio Damasio reported that "she didn't speak because she felt as if she 'really had nothing to say.'" (Prinz 2012, 42) Prinz further reports that these individuals "seem comatose -- they do not respond verbally or behaviorally to the world around them -- but they are actually fully cognizant of the world around them." (Prinz 2012, 67) And again, after recovery, they report having been "emotionally dead. They do not respond to the world around them because everything leaves them disinterested (sic)." Now it's clear from these descriptions that such akinetic individuals are fully aware of themselves and what subjective states they retain[16], given that after recovery they can report on what they experienced when mute. And they report being devoid of emotional engagement, and thus are devoid of concern. Another patient of Damasio's, named Elliot, who had a tumour near his frontal lobe removed, became completely emotionless and also incapable of proper decision-making. Damasio reports:

> He was always controlled… always describing scenes as a dispassionate, uninvolved spectator. Nowhere is there a sense of his own suffering, even though he was the protagonist… He was not inhibiting the expression of internal emotional resonance or hushing inner turmoil. He simply did not have any turmoil to hush. This was not a culturally acquired stiff upper lip. In some curious, unwittingly protective way, he was not pained by his tragedy. I found myself suffering more when listening to Elliot's stories than Elliot himself seemed to be suffering… I never saw a tinge of emotion in my many hours of conversation with him: no sadness, no impatience, no frustration. (Damasio 2005, 43)

These individuals all seem from their descriptions to be self-conscious but unconcerned. However, one thing I left out of Prinz's description is that he consistently

[16] This is consistent with Prinz's conjecture that their lack of emotion may be because "conscious experiences of emotionally significant bodily changes are lost." (Prinz 2012, 67)

refers to these individuals as “people,” but calling them people would seem to contradict my claim that concern is necessary for personhood. There are two points to make in response to this apparent conflict. First of all, much depends on how terminal the akinetic or otherwise emotionally-impaired condition is. The cases described by Damasio and Prinz are mostly ones where the patients recovered (otherwise we wouldn’t have the data of the patients’ reports.) In the introduction I addressed Chappell’s concerns about criterialism, specifically his worry that if we define persons in terms of high-level cognitive properties, then we must accept some absurd implications. For example, anyone who is in a coma, or even merely asleep would not count as a person because he or she would not be self-conscious. My response to this complaint was that if we define persons not in terms of their occurrent properties, but their capacities or dispositions for having the relevant properties then we can account for individuals remaining persons even when they are temporarily failing to instantiate them. We can further distinguish cases where a capacity is temporarily disabled, from ones where it is just not currently in use, like being asleep. In the latter case, we can say that a capacity is currently ‘intact’ and ‘poised,’ while in the former it is ‘intact’ but temporarily ‘impaired’. If we define persons in terms of intact capacities that could be in either condition, Chappell’s worry can be defused. So if an individual is only temporarily clinically apathetic, we can say he or she has an intact, though impaired, capacity for concern. In cases where there is no hope of the individual ever being concerned again, then individual is no longer, strictly speaking, a person. What the lack of concern would

imply, is that such an individual cannot be responsible for his or her actions, though it is likely that he or she would perform few actions anyway. Such a being might not even be totally mute or inert and therefore still, automatically, go through the routines of his life, but would have no emotional investment in the continuance of such habits, and would therefore be like the self-conscious chess computer, i.e. not a person. Pierre, the eponymous character of Sendak's (1962) children's book always says "I don't care," even when he is about to be eaten by a lion. He does not seem to be a person until he is removed from the lion's belly, and grateful at surviving the experience exclaims: "Indeed, I do care!"

Persons can be responsible because they are able to engage in actions that follow from their consciously considered beliefs and desires, but also because they can preemptively evaluate both the actions and the desires that yield them and decide whether or not to perform them depending on their judgment of future consequences. This requires both self-consciousness and concern. It requires concern because concern is what gives rational evaluation its motivating force. Consequences can only be weighed and considered in a way that is meaningful for an individual if that individual is concerned about them. They cannot be genuine reasons for action without such emotional significance. The self-conscious but unconcerned computer may calculate the best route towards a goal, but that is not the same as considering reasons and then acting upon the one that is most strongly motivating. Locke also thought concern was necessary for personhood, but he did not offer any explicit definition of the capacity, and

seemed to understand it rather differently from how I have described it, at least in how it relates to self-consciousness. For Locke, concern for happiness is necessarily present when a being is self-conscious. It is:

> the unavoidable concomitant of consciousness; that which is conscious of pleasure and pain, desiring that that self that is conscious should be happy. And therefore whatever past actions it cannot reconcile or appropriate to that present self by consciousness [i.e. self-conscious], it can be no more concerned in than if they had never been done: and to receive pleasure or pain, i.e. reward or punishment, on the account of any such action, is all one has to be made happy or miserable in its first being, without any demerit at all. (Locke, 1690 II.XXVII.346.26)

Whiting (2002) similarly seems to understand concern (though she also does not offer a precise definition of the term) as necessarily bound up with self-consciousness. She writes:

> What we have here is a *holistic package* whose components are functionally related to one another: consciousness in a normally embodied creature is (among other things) consciousness of pleasure and pain, the very essence of which engage their subject's *concern* in ways that lead the subject to *act* so as to increase the pleasures and diminish the pains of which it is *conscious*, and *consciousness* of such action and its basis in the subject's *concern* leads the subject to *impute such action to itself* in a way that renders intelligible the forensic practices of holding oneself and other subjects *responsible* for their actions. (Whiting 2002, 207)

So for Locke and Whiting, self-conscious, though unconcerned, creatures are impossible. Any creature aware of itself would necessarily be concerned for itself. But as we have seen, Locke and Whiting are wrong about this - the two capacities are conceptually distinguishable. A self-conscious system need not have emotional investment, though, as I have argued above, having genuine desires does require emotional investment. So a self-conscious, unconcerned being would have conscious

beliefs and goals, but no genuine desires.

A self-consciously concerned being is a being with desires that can be consciously scrutinized when deliberating over actions. On the other hand, we can imagine highly intelligent, fully self-conscious beings that are unconcerned. In addition to the apathetics described above, it may be the case that some ascetics wish, through practice, to become entirely unconcerned. They may see concern as a kind of weakness to be transcended or wish to reach a state wherein, as Flanagan (2009) describes *nirvana* (though I don't think most Buddhists would agree that this is what they are aiming at), "one ceases to exist as a desirer and the flame that one was is extinguished forever." On the current proposal that would mean that these individuals wish to transcend personhood. The *transhumanist* movement thinks that we will appropriate technology in a way that will make us no longer human. However, if one were modified in such a way that one were no longer self-conscious or concerned, it would be more apt to say that such modifications would make one no longer a person, and therefore would be *transpersonal*. This is not to say that such a being would be *less* than a person. There is no reason to think that persons occupy the highest rung on the ontological ladder. There might be beings that transcend personhood, that are in some way more than persons. This might be something like what Nietzsche (1886) has in mind when he envisions the Übermensch, who has overcome the human-all-too-human (read: personal-all-too-personal) and attained a higher level of being, one without desires and therefore incapable of responsible action; or to paraphrase Sartre (1956),

no longer condemned to responsibility.

Just as not all self-conscious beings need be concerned, not all concerned beings must be self-conscious. Many animals evidence a capacity for happiness and sadness, frustration and satisfaction in the absence of self-consciousness. Appropriating a couple of old terms, we can say they are *sentient*, but not *sapient*, where ‘sapience’ is meant to cover all the capacities that together are sufficient for personhood, including both concern and self-consciousness. These animals’ behavior demonstrates concern for their own well-being without offering evidence of a capacity for fully self-conscious awareness of their desires and beliefs. If you pet my cat Bisou when she doesn’t want you to, she will let you know of her displeasure by whining. If you persist she will seem to become increasingly agitated, whining louder and longer, even growling gutturally until finally lashing out with a scratch, bite and/or hiss. It is difficult not to infer from this behavior that the cat becomes increasingly emotionally charged as her desires continue to be frustrated. Likewise a dog that has waited all day for its owner to return home displays a joy and excitement when the beloved human finally arrives which has erupted out of a desire that has increased in its intensity over time. Behavior associated with fear is even more clearly exhibited by a large variety of species, not restricted to mammals. And for this reason most studies of animal emotion focus on it. In the presence of a threat to her well-being, a dog for example, Bisou stops in her tracks before quickly diving for shelter. Her hair stands up on end and she alternately hisses and growls. This escalating fear behavior shows that the cat is not

just shocked by the presence of a threat but emotionally aggravated by the thwarting of its desire that the threat go away. According to Makowska and Weary, experimental inquiry has provided evidence of amusement in rats. “Rats will seek out hands that have tickled them much more than hands that have petted them an equal amount of time… and will learn to press a lever for a tickling reward. When being tickled and during social play, rats emit 50kHz calls that may be indicative of positive affect.” Makowska and Weary 2013, 8) Panksepp and Burgdorf (1999) and Panksepp (2007) argue that these responses are analogous to human infants’ laughter, and since they only occur in environments where the rats are comfortable (i.e. ones where they have not been subjected to stressful stimuli or punishment) their response is not merely stimuli-dependent, but dependent on affective state as well.

I use the term ‘concern’ instead of just saying that persons and animals have the capacity for *emotion* because ‘concern’ is not just emotion, but emotion directed toward the satisfaction of one’s desires or truth of one’s beliefs. One could imagine an individual who displays random emotional behavior that is not caused by anything in particular that the individual has experienced. One minute he or she is angry, the next minute ebullient whether or not anything good or bad is happening to him or her. This kind of condition, unfortunately, is all too common to varying degrees in people who suffer from bipolar disorder and related pathologies. It is only in the most extreme kind of case, however, where an individual’s emotions are never tied to their desires or beliefs, that we have reason to deny that it is a person we are dealing with. In this way,

concern, like intentionality in the first place, implies a certain kind of rationality. For emotional states to be states of concern, they must be responsive to the perceived relation between desired or believed states of affairs and the way the world actually is, though the perception of how things are could itself be mistaken.

Belief that animals have emotions is reinforced by neuro-physiological similarities with human beings. For instance, all mammal, and some bird and reptile brains have amygdalae (Ledoux 2002, 218), which have been shown to play a crucial role in human emotions, particularly fear, and are activated in situations where both animals and humans display seemingly emotive behavior. In creatures with damaged amygdalae fear response and conditioning is inhibited. However, despite the behavioral and neurological evidence, some writers call into question the confidence with which we can attribute emotions to animals. The main strand of these doubts is sometimes called the "credibility problem." The problem is that if emotions are understood as subjective feelings, they cannot be assessed in non-linguistic creatures who cannot report on them. This is because while animals may display behavior that looks similar to behavior associated with certain feelings in humans, this similarity may be misleading as an indication of the internal states of the animals. Ledoux's (2002) suggestion to avoid this problem is to define emotions not in terms of subjective feelings, but instead in terms of processes that contribute to particular kinds of behavior, given particular stimuli, along the lines of the information processing model in cognitive science. "Since emotions as processes can be studied in animals and humans alike, and since… emotional

processing underlies both emotional behavior and emotional feelings, a processing approach is a way out of the credibility problem." (Ledoux 2002, 205) So if one is skeptical about attributing feelings to animals, one can read "capacity for concern" as a bit of information processing, though I am fairly confident (as is Ledoux), that many do have such subjective states. While one may not be able to rule out skeptical doubts on that score, the situation is not much different in the case of our knowledge of the minds of linguistic creatures as well. And though we might not have any way to tell with absolute certainty whether or not an animal has the feelings inferred from its behavior, that doesn't mean there is no fact of the matter of whether or not it has them.

So by requiring both self-consciousness and concern, we embed persons (at least the human ones, but perhaps also some other sorts) in the animal continuum but at the same time mark them as distinctive. The shared capacity for concern explains our sense of camaraderie with animals and gives us reason to consider them as deserving of moral consideration. A New York Times Sunday Review opinion piece from October 5th, 2013, by Gregory Berns, claims that "Dogs are People, Too" on the basis of their possessing a brain area "caudate nucleus" in common with human beings. The author claims that "the caudate plays a key role in the anticipation of things we enjoy, like food, love and money," having found (by scanning the dogs in an MRI machine without sedating or restraining them, so that "if the dogs didn't want to be in the scanner they could leave") that "activity in the caudate increased in response to hand signals indicating food. The caudate also activated to the smells of familiar humans. And in

preliminary tests, it activated to the return of an owner who had momentarily stepped out of view." The author takes these findings to indicate that dogs experience emotion and argues for their positive treatment on those grounds, claiming "that dogs have a level of sentience comparable to that of a human child. And this ability suggests a rethinking of how we treat dogs… And this means we must reconsider their treatment as property." Now, as Alva Noë rightly points out in his reply to that article ("If You Have to Ask You'll Never Know" *NPR Blogs: Cosmos and Culture* October 11th 2013), we don't need to scan dogs' brains to be able to tell if they have emotions. It should be evident from their behavior alone. I agree that neurology doesn't tell us much about emotions that we shouldn't already have gathered from behavior. I grant that dogs do have emotions, concern, sentience, and thus are deserving of our consideration, perhaps even having a claim to rights. However, that doesn't mean we have to stretch the concept of a person to include them in our moral community or to recognize their status as ends in themselves and not property.

This is one way in which the moral connotations of the concept of a person need to be rethought. If persons are the sole bearers of rights then it will always seem absurd to grant dogs, cats and elephants rights, because they clearly are not persons. They cannot claim those rights for themselves - persons must claim them on their behalf, because persons are able to think critically about their own interests and those of others. Therefore, we should say that though dogs are not persons they deserve respect for their interests and have a claim to the recognition of their rights. The author

of the article suggests perhaps granting dogs a “limited personhood” and if one prefers to think of it that way then one may. Otherwise we can treat them as persons even if they are not strictly so, but, in any case, it is a mistake to think of dogs as full-fledged persons. There is a crucial difference between their capacity to monitor and control their desires and our own as regards responsibility. Noë, however, disagrees on this point. He thinks animals are responsible, citing animal trainer Vicky Hearne as arguing persuasively that “you can’t work with dogs unless you can take them seriously as, well, responsible agents. A-search-and-rescue, for example, or a seeing-eye dog, is a collaborator, not a tool.” (Noë 2013) However, while I agree that such creatures are not just tools, and may be considered collaborators, tool and person are not exhaustive categories and collaboration does not entail responsibility.

As I will argue in chapter five, my insistence on including concern as a necessary condition of personhood does not imply that concern for others is necessary, nor does it imply that persons are essentially moral, i.e. concerned with the experiences of others in the same way that they are concerned with their own. Persons are the sort of things that can reasonably be held responsible, but for that they need not be of the sort that do what is right, or even recognize the difference between right and wrong from a genuinely ethical perspective (i.e. not just following explicit rules and practical reason.) The capacity for the feeling of empathy, which is arguably an essential part of the biological foundation of morality, is present in many kinds of life forms, particularly mammalian, but is always unequally distributed among the members of any individual

species, including human beings, our paradigmatic persons. This could also be the case with any of the imagined non-human persons of science-fictional thought experiments. Persons are necessarily self-concerned, yet may entirely lack empathy. Requiring empathy would make the concept of a person far less useful for the classificatory purpose for which it is otherwise well suited.

IV. Rationality and other proposed conditions

Both the capacities of self-consciousness and concern bear complex relations to the concept of rationality, which is one of the most difficult folk psychological notions to pin down with any precision. Bermudez (2005) is right to use it as the prime example of a theory-cluster concept. As I said when discussing Dennett above, we ascribe most intentional states on the basis of goal directed behavior and the utility of such judgments requires that at least it is usually the case that such behavior is rational, i.e. that the individual's beliefs about how to accomplish his or her goals are largely justified, otherwise we wouldn't be able to use intentional terms to consistently describe what an individual is attempting to do. However, it also seems to be the case that one can have irrational beliefs, ones that depend on consistently unreliable sources of justification or that are fallaciously inferred from justified premises. If there were an individual for whom all of his or her beliefs were thus irrational, yet those irrational beliefs were fully conscious and the individual in question were also concerned in the sense explicated above, we would hesitate to call that individual a person.

To be a person, an individual's perceptions must give her mostly veridical information about the world. Even if someone lost all of her capacities for visceral sensation, if she were still able to think rationally, she would realize the reality of being shut off from the world. If she were then unknowingly hooked up to a computer simulation of reality she would still be able to reason based on what she experienced in the simulation. Therefore, a person must draw valid inferences from whatever sensory capacities she has. A minimal reasoning capacity would therefore require the ability to perform deductive and inductive inferences. However, people differ in the degree to which they possess those abilities which is why I hesitate to make a kind of rationality, more robust than that already implied by self-consciousness and concern, necessary for personhood. We just have to settle for an admittedly vague principle of "seeing things for how they are," though I will argue in chapter five that seeing things for how they are does not require seeing them in ethical terms.

Having explicated the account of personhood in terms of self-consciousness and concern with the minimal rationality implied by those capacities, I am now in a position to view it set against the sundry items on Gordijn's list of potential criteria. Gordijn's claim was that the disagreement over which of the items on the list should count as necessary and sufficient for being a person is intractable. I will attempt to show that the conditions on the list are either implied by the ones I have already considered or are not necessary for personhood, and in some cases internally incoherent to begin with. It is unlikely that this will actually settle the debates for the parties involved, but I hope to

demonstrate that it should.

1. The capacity to experience pleasure and pain.

It is tempting to say that the capacity for pleasure and pain is a necessary condition of concern and therefore already covered by the inclusion of the latter capacity. However, it is conceivable (and perhaps even actual) that there be individuals whose nervous systems are defective in such a way that they cannot ever feel physical pleasure or pain. Such individuals may still be concerned for their health and well-being, capable of happiness or sadness, though they are devoid of pleasure and pain sensations. Those individuals are clearly no different from ones who do feel those sensations in any way relevant to responsibility and therefore are persons. So the capacity for physical pleasure and pain is not necessary for personhood, though the capacity for positive or negative emotional states is.

2. The capacity to have desires, is clearly entailed by my account.
3. The capacity to remember past events.

This is a more difficult challenge. Sophisticated memory does seem to set persons apart and for many, including Locke, is the mechanism that confers responsibility over time. However, memory has turned out to be a heterogeneous group of functions that may be differently related to the capacities constituent of persons. Working memory seems integral to consciousness, and is characteristic not just of persons, but many animals as well. Episodic memory is the sort that has been traditionally thought to account for the identity relation among person-stages, but there are reasons to doubt the adequacy of memory-based accounts of personal identity that

will be explored in chapter two. It is enough to mention here that there have been empirical studies of dissociative amnesic individuals, whose ability to form episodic memories is impaired, that seem to demonstrate that episodic memory is unnecessary for much of what we consider integral to the responsibility associated with personhood. (Craver 2012) Craver employs a method of testing individuals with cognitive deficits against conceptions of personhood to see if they live up to them and finds that the ones without the capacity for episodic memory still "are able to track and respect the reward and punishment structures of their world well enough to guide adaptive choices." (Craver 2012, 468) Still, these individuals do possess declarative memory or propositional memory, just not the memory of experiences. Declarative memory is likely necessary for rationality and self-consciousness, so that one's beliefs can be based on knowledge of the past and general truths and one can hold one's desires and beliefs in thought long enough to scrutinize them.

4. The capacity to have expectations with respect to future events.

In his account of individuals with episodic memory deficits, Craver notes that the same neurofunctional processes associated with episodic memory play a role in one's ability to imagine future situations one might experience. However, while "episodic future thought can be used to modulate decision-making about hypothetical future rewards" an individual who is unable to form such thoughts is still able to hypothetically value future events and is "willing to exchange an immediate reward for a larger reward at a later time" but "cannot imagine how he would spend the money. After repeated

questioning he says he would put it in the bank." (Craver 2012, 468) So if 4. is taken in that robust sense, then it seems not to be necessary. In a more rudimentary sense, however, in order to be concerned for one's future experiences relative to one's desires one must be capable of some kind of expectation or anticipation of future circumstances, though not necessarily episodically.

5. An awareness of the passage of time.

It's hard to say exactly what such awareness consists in. If it means understanding of the difference between the past and the future, or A-series time, to use Mactaggert's (1908) designation, such understanding might be necessary for personhood because in order to be responsible for one's actions one must be able to distinguish an action from its causes and consequences in such a way that one can see that the consequences follow from the action, which is or is not justified depending in part on the events that led up to it. As Craver puts it: "It seems requisite for us to be an agent in any full sense that we recognize that possible futures lie ahead of us, that our pasts are irrevocable, and that the choices we make now will have consequences in the future" (Craver 2012, 464). It is likely also necessary for basic intentionality in the first place. Many beliefs are time-indexed, they are about how things are, were, or will be at various times. The same goes for desires. There are immediate wants, but also longings for how things may turn out eventually, and sadly, for there is nothing to be done about it, desires for what was but is no longer or even what might have been. So understanding of time, in that sense, is necessary for personhood but is already implied

by the notions of self-consciousness and concern.

6. The property of being a continuous, conscious self, or subject of mental states, construed in a minimal way, as nothing more than a construct of appropriately related mental states.

7. The property of being a continuous, conscious self, construed as pure ego, that is, as an
entity that is distinct from the experiences and other mental states that it has;

I'm taking these two together because they both concern the concept of a self, albeit in extremely different ways. Whether having such continuity or such a self as conceived of in (6) is necessary for personhood is a question I will explore in depth in chapter four, though it does seem that there has to be something to the 'self' in self-consciousness. One way to understand that term is purely reflexively, where 'self' refers to 'this person'. In that case the problem is just one of explicating the reference of 'this person,' at different times, which is the persistence problem which I will discuss in chapter two. Another kind of thing 'self' could mean is an internal cognitive representation of a person - something experienced, though not necessarily substantial. The question of whether a being needs to be capable of forming such a representation and what the nature of that representation is (i.e. narrative, imagistic, relatively permanent or fleeting, single or multiple, etc.) will be the subject of chapter four.

However, for reasons that were explained in the introduction, I see the idea of a self in 7. that is "distinct from the experiences and mental states that one has" as both too scientifically and philosophically untenable to consider it necessary for personhood. There is no evidence that such a thing exists. If it did, there is no coherent way of explaining how it could causally relate to the rest of the person and it is unnecessary to

explain any observable or phenomenal facts about persons.

8. The capacity for self-consciousness, that is to be aware of the fact that one is a continuing, conscious subject of mental states;
9. The property of having mental states that involve propositional attitudes such as beliefs and desires;
10. The capacity to have thought episodes, that is, states of consciousness involving intentionality;
11. The capacity to reason;
12. The capacity to solve problems;

The above capacities, 8-12 have been covered already either under the notion of self-consciousness, or in the case of 11-12, rationality.

13. The property of being autonomous; that is of having the capacity to make decisions based upon an evaluation of relevant considerations;

13., so long as it doesn't imply indeterministic free will, can be taken to be more or less equivalent to the sort of responsibility that I take self-consciousness and concern to be necessary and sufficient for.

14. The capacity to use language;
15. The ability to interact socially with others;

The above two capacities were both discussed in the section on self-consciousness, where I argued that 14. may be necessary for self-consciousness, though that does not imply the necessity of the ability to use language to communicate with others. By that same token, while it may a contingent fact about human persons that we could never develop sapience in isolation I don't see any reason to rule out *a priori* that some being could.

Having considered what I take to be a representative list of the capacities sometimes considered necessary for personhood and explained how they relate to self-

consciousness and concern (though some of the details of that explanation had to be postponed until later chapters), I hope to have shown that (aside from those who believe in separate egos) one who endorses any of the capacities on Gordijn's list should be satisfied that my account of personhood is inclusive of them while providing a both more succinct and complete picture. In the remaining chapters I will pick up the threads left dangling concerning persistence over time (chapter two), reductionism and Eliminativism (chapter three), the self (chapter four) and the moral implications of personhood (chapter 5).

Chapter 2: The Persistence of Persons

I. The problem of personal persistence and responsibility for past actions

So far I have claimed that persons are beings that are capable of being responsible for their actions. For that distinction to have any significance in practice, persons must sometimes be responsible for actions performed in the past. Otherwise, no one could ever be held responsible for any action, because responsibility for any action would be expunged immediately after the action was performed. For anyone to be responsible for any action performed by a person at an earlier time there must be a suitable relation between the person at the later time and the person at the earlier time, who performed the action, such that responsibility is maintained from the earlier person to the later one. Usually the relation in question is assumed to be that of *identity*, though some writers, such as Parfit (1984) have proposed slightly different candidates, such as *survival* or *persistence*.

One reason for picking an alternative relation to identity in the above general formulation is to allow for the sort fission and duplication cases (such as when a brain is split and each hemisphere is placed in a different body, or a person somehow splits like an amoeba, leaving two exactly similar persons who each have the same claim to being the original), in response to which Parfit (1984) morphs the question of "personal identity" into a question about "personal survival." If the relation holds between a person Y at a time t_2 and a person X at an earlier time t_1 that holds in an ordinary case of survival at the same time and to the same degree that that relation holds between a

second person Z at t_2 and the person X at t_1, then X would survive as Y and Z despite the fact that X is identical to neither Y nor Z. That there is no transtemporal *identity* in such a case should not, according to Parfit, concern us, because identity is just a special case of survival, and that latter relation is all that really matters. Identity is merely the one-to-one case of survival. I will grant this point to Parfit at this stage, assuming that a person one survives as but is not identical to could be responsible for one's present actions. Parfit's views about what grounds continued responsibility are complex and not entirely clear. I will deal with them in depth later on. However, I will assume at this stage that responsibility goes along with survival, so that whatever grounds responsibility in the special case of identity could also hold between a person at an earlier time and any person at a later time he or she survives as in branching cases, allowing for the possibility that two contemporaneous persons could both be responsible for something an earlier person did so long as that earlier person *persists* or *survives* as those later two.

A second reason for that last disjunction, is that there is a difference between *strict* identity and "identity in the loose and popular sense" borrowing a phrase from Butler (1736) and, later, Chisholm (1976).[17] In particular, if one is committed to a mereological or qualitative essentialist version of psychophysical reductionism, according to which a being cannot survive any change in its parts or properties, then one might contend that no being can be *strictly* identical from one moment to the next

[17] Though Butler and Chisholm themselves insist that persons are the only things that *must* be identical over time in the strict sense.

due to continuous mereological and/or qualitative change, though some relation holds between numerically distinct momentary person-stages that unites them into what can be loosely, or conventionally (invoking the Buddhist idiom) regarded as the same person over time. As with the point made in the previous paragraph, at this stage I will not rule out the possibility that such a loose or conventional identity is sufficient for maintaining responsibility. For brevity's sake, I will mostly speak of 'persistence', though I will use the phrase 'the same person over time' where that phrase should be understood to be similarly non-committal, and I will speak of 'identity' when discussing a particular writer who uses that term.[18]

The question I address in this chapter is: Given the conception of personhood developed in chapter 1, according to which a person is a being with the capacity for self-consciousness and concern, what conditions must be met for a person at one time, t_1, to persist as a person at a later time t_2, given that persons change mereologically and qualitatively over time? In answer, I motivate and defend an account of the persistence of persons over time in terms of the continued existence of a dynamic, organized being that instantiates and uninterruptedly maintains the capacities for self-consciousness and concern.

One proposed criterion of personal persistence that has been very popular among philosophers over at least the past forty years, is the "Psychological Criterion" defended by Parfit (1971/1984) and Lewis (1983), among others, which essentially holds that

[18] Chapter 3 will deal more thoroughly with the mereology of persons.

continuity of distinctive psychological features is necessary and, given an appropriate causal link between states, sufficient, for personal persistence over time. In contrast, I endorse a view I call the Core Psychological Criterion, according to which personal persistence requires the continuous realization of core psychological features of persons over time, but not continuity of distinctive features. Following Unger (1990), core psychology refers to the features possessed in common by all psychological beings, whereas one's distinctive psychology consists of the psychological features that an individual either possesses uniquely or else possesses in common with some but not all other psychological individuals. The core psychological features are general capacities, not specific psychological states. This is important to both Unger's view and my own, because capacities are the sorts of things that are maintained even when they are not in use and even when they have been temporarily disabled, so that a person would persist while asleep, while in a reversible coma, and even if cryogenically frozen. As I will stress later on, this is an advantage of Unger's view over the traditional Psychological Criterion, because it is difficult to see how one's distinctive psychological features, even the dispositional ones, would be continuous through such conditions.

My position is similar to the "Physical Criterion" defended by Unger (1990)[19], but differs from Unger's view in a few ways: First of all, Unger restricts the class of beings

[19] Mcmahan's 'Embodied Mind' (2002) account is also similar, but differs from mine in more or less the same ways as Unger's as well as in the additional respects that 1. the possible realizers of minds seem, in Mcmahan's account to be limited to organic brains and 2. Mcmahan thinks that the functional and organizational continuity instantiating a mind could persist through teletransportation whereas I think that such an event would necessarily interrupt such continuity. Furthermore, I don't take egoistic concern to be central to what matters in persistence the way Mcmahan does.

that could realize the core psychological capacities of persons to physical beings and takes maintenance of such capacities to require physical continuity (not without good empirical reasons), but I state my criteria of personal persistence in more metaphysically neutral terms, in the sense that I leave open the possibility that there could be non-physical persons that persist via non-physical causal continuity (though such continuity would have to be equivalent to physical continuity in crucial respects.) I am, along with Unger, committed to naturalism in the sense that I do not require for personhood or persistence the existence of any entity or feature that is in conflict with established scientific fact or is undiscovered so far by scientific method. In that class I include an immaterial substance, a transcendent or immutable soul, and non-deterministic free will. However, I leave open the possibility that persons may be composed of or possess such things.

Secondly, the core psychological features that Unger takes to be relevant to persistence are those which are minimally required to be a psychological being, “my capacity for conscious experience, my capacity to reason at least in a rudimentary way, and my capacity to form simple intentions.” (Unger 1990, 193) Conversely, my account requires of a persisting person the more demanding set of core psychological capacities of *persons*, i.e. self-consciousness and concern. Insisting on capacities beyond those shared by all psychological beings importantly distinguishes my view from Unger’s, because it follows from my view that an individual organism could continue to live as a psychological being despite failing to maintain the capacities necessary for personhood

and would then cease to be a person.

Another difference between Unger's account and mine is terminological, for he calls his account the "Physical Criterion", which is misleading given his focus on psychological capacities, despite his insistence that they be physically realized. This is important, because Unger's view is different from most 'physical', 'somatic', or 'bodily' views endorsed in the personal identity/persistence literature by writers such as Van Inwagen (1995), Olson (1997), and Thomson (2008), which take persons to be co-terminous with the lives of individual human organisms. Unger's view, on the other hand, allows for the possibility that one might transform into something that can no longer be described as a human body, without thereby ceasing to persist as a person. The "physical" epithet is even more misleading in describing my own account given my relative metaphysical agnosticism (as regards the possibility of non-physical beings), as well as my contention that one might cease to be a person while remaining a human organism if one ceased to maintain the capacities necessary for personhood. Instead, I suggest the view be called the Core Psychological Criterion.

The appeal of the Psychological Criterion seems to rest on the intuition that our distinctive psychological characteristics are essential to distinguishing one person from another, so that it is "what's on the inside" that counts. In other words, our physical appearance and attributes are not what we consider essential to ourselves as distinct individuals, but rather it is our psychological qualities - our beliefs, desires, values and preferences, that matter most to our sense of who we are. This intuition lies behind such

turns of phrase as “she is not the same person she used to be.” Such a phrase is usually uttered when someone has radically altered her beliefs, motivations, values or behavioral dispositions, which are generally understood to be psychological attributes. We can understand the question about the necessity of psychological continuity in terms of whether or not we mean it *literally* when we say that someone is “no longer the same person” due to that person’s change in psychology. Most writers’ way of deciding this question is to consider various thought experiments, as well as some actual cases, where someone has changed psychologically and to anticipate what one’s intuitions about identity are when presented with those cases. Through such an inquiry, one attempts to establish not only whether or not psychological continuity is necessary, but also which psychological features are essential and to what degree they must be continuous for a person to persist over time.

According to Parfit (1984) when we say of someone that he or she is ‘no longer the same person’, this may be a claim about both qualitative and numerical identity. “Indeed, on one view, certain kinds of qualitative change destroy numerical identity. If certain things happen to me, the truth might not be that I become a very different person. The truth might be that I cease to exist -- that the resulting person is someone else.” (Parfit 1984, 202) Conversely, according to the Core Psychological Criterion, no qualitative change to a person (short of making him or her a non-person) can make that person a numerically different person. I might cease to exist if changed into something that is not a person, but so long as the properties constitutive of a being person are

maintained, I remain numerically the same person.

My argument for endorsing the Core Psychological Criterion and rejecting the traditional Psychological Criterion is essentially that the Core Psychological Criterion, given an account of personhood in terms of psychological capacities, allows for a unified theory of personhood and personal identity, which is also largely continuous with accounts of the persistence of other organized beings, and embraces both the suppositions that the unique psychological features of persons underlies our persistence as persons while allowing that a person's psychological features may be highly variable over the course of her life. The traditional Psychological Criterion, on the other hand, cannot support the last premise without ad hoc adjustments, and more importantly, unjustifiably makes the persistence of persons radically unlike the persistence of any other kind of organized being. Furthermore, the Core Psychological Criterion accords with a conception of responsibility over time that reflects actual legal practices, the abandonment of which would be untenable.

After providing a historical overview of the development of the Psychological Criterion from Locke to Parfit and Lewis, I present Unger's arguments for his position against Parfit's expanded version of the Psychological Criterion as well as offering my own additional arguments. In some places these arguments follow the traditional method of approaching this topic by relying on intuitions about thought experiments. Some doubt has been cast on the legitimacy of such arguments (e.g. Wilkes 1988 and following her, Schechtman 2014) and they are surely less powerful than one would like

them to be, but there aren't many better sources of justification around when debating this issue. I have, however, appealed to theoretical considerations in addition to intuitions as much as possible. Furthermore, I have engaged with some of the data gathered from experimental philosophy's initial forays into this topic and am open to other empirical evidence that may be relevant, e.g. from neuroscience and physics. My primary method, however, remains the prevailing way of approaching this topic, which is that of conceptual analysis. However, the method is not just analytical, but in part revisionary, because it involves tightening up and clarifying the concept of a person and of personal persistence, which are vague and internally contradictory in ordinary usage.

II. Continuity of consciousness and the memory criterion

Locke is the first modern Western philosopher to develop a conception of personal identity in terms of psychological attributes, rather than mental substance. As discussed in chapter one, Locke takes the concept of a 'person' to be a forensic one, the purpose of which is to track responsibility over time. For him, such continued responsibility requires that a later individual have 'the same consciousness' as the individual who performed an action in the past. Both Locke's predecessor, Descartes, as well as his early critics, Butler and Reid, also take continuity of consciousness to be essential, but Locke understands such continuity in terms of the continuity of psychological attributes, specifically memories, rather than of the continued existence of a mental substance, although, as will become evident, he means something peculiar by

‘continuity of psychological attributes.’

For Locke, the persistence of persons cannot be accounted for in the same way that the persistence of other kinds of complex beings, such as biological organisms and man-made mechanical objects, can. The identity of compound, but unorganized bodies, or 'Masses of Matter', consists in the identity of their parts, so that a change in parts implies the destruction of one object and creation of a new one. (Locke 1690, II.XXVII.330.15) The identity over time of organisms and other organized bodies, which persist despite changes in their parts, consists in the continuity of their life, understood in terms of the maintenance of their organizational structure such that it enables them to maintain the functions associated with the sorts of things that they are. For living organisms these functions are generally limited to continued life, nutrition, generation and regeneration of cells. For instance, an oak remains an oak and therefore the same oak so long as it maintains an organization of parts

> as is fit to receive, and distribute nourishment, so as to continue, and frame the Wood, Bark, and Leaves, etc. of an Oak, in which consists the vegetable Life. That being then one Plant, which has such an Organization of Parts in one coherent Body, partaking of one Common Life, it continues to be the same Plant, as long as it partakes of the same Life, though that Life be communicated to new Particles of Matter, vitally united to the living plant, in a like continued Organization, conformable to that sort of Plant. (Locke 1690, II.XXVII.331.15)

Man-made objects, while they have no ‘Life,’ have more specific functions to maintain, e.g. a watch tells time and a table holds other objects aloft. This account applies to human organisms as much as it does to any others, whether oak, frog, cat or watch, however, Locke insists that the same human organism or “same Man” (or even

some other sort of organism, such as a super-intelligent parrot) could be at one time a different *person* than at another and therefore not responsible for things the Man did as the former person. Locke thinks that this sort of possibility is a consequence of his definition of a person as "a thinking intelligent being, that has reason and reflection, and can consider itself the same thinking thing, in different times and places; which it does only by that consciousness which is inseparable from thinking..." (Locke 1690, II.XXVII.335.10). Furthermore, Locke thinks that if "that consciousness" which allows one to consider oneself "the same thinking thing" at different times and places could be transferred from one organism or substance[20] to another, the person would go with it. "For should the Soul of a Prince, carrying with it the consciousness of the Prince's past life, enter and inform the Body of a Cobler as soon as deserted by his own soul, everyone sees he would be the same Person with the Prince." (Locke 1690, II.XXVII.340.10)

Locke's psychological account of persistence is purely first-personal. Consciousness is what allows one to distinguish oneself from everyone and everything else. Therefore, what makes me the same person as some person in the past is that I maintain the same consciousness. In other words, my consciousness extends backward to the thoughts and actions of that past person. What that amounts to, for Locke, is my remembering having been conscious of that person's thoughts and actions in the past:

[20] Gordon-Roth (2015) argues that persons must themselves be understood by Locke to be substances and not modes of other substances as other commentators have suggested. That debate is beyond the scope of this study.

> For, since consciousness always accompanies thinking, and it is that which makes every one to be what he calls self, and thereby distinguishes himself from all other thinking things: in this alone consists personal identity, i.e. sameness of a rational being; and as far as this consciousness can be extended backwards to any past action or thought, so far reaches the identity of that person; it is the same self now it was then; and it is by the same self with this present one that now reflects on it, that that action was done. (Locke 1690, II.XXVII.335.20)

An alternative account of continuity of consciousness, which Locke rejects, would consist in the continued existence of particular *states* of consciousness, so that I would be the same person as a person in the past if I continue to possess some of that previous person's conscious states. Locke rejects this proposal because he understands all conscious states, indeed all thoughts in general, to be of only instantaneous duration, such that none persist from one moment to the next. As we shall see later on, Locke's *atomistic* conception of thought has been one of the obstacles to establishing of satisfactory account of the persistence of persons in terms of psychological continuity:

> Because each perishing the moment it begins, they cannot exist in different times, or in different places, as permanent Beings can at different times exist in different places; and therefore no motion or thought considered as at different times can be the same, each part thereof having a different beginning of Existence. (Locke, 1690, II.XXVII.329.30)

Thus for Locke, the same consciousness over time can only be understood in terms of memory, specifically the sort which contemporary psychologists call 'episodic memory' or 'experience-memory.' More precisely, sameness of consciousness over time depends on a memory (a current state of consciousness) having the same content as the original conscious state; i.e. of the event remembered. If I can remember performing any action, which was consciously performed by some person in the past,

then I am the person who performed that action. If I remember giving the Gettysburg address, for example, then I am Abraham Lincoln.

Locke's view can be stated as follows:

A person Y at t_2 is identical to person X at t_1 iff

1. X consciously performed some action at t_1 and

2. Y at t_2 consciously remembers performing that action at t_1.

If Locke is right about this, then it follows that one and the same man, meaning human being or human organism, could be two different persons at different times and vice-versa - one and the same person could be embodied in different human organisms or men at different times. So long as memory is preserved, survival and/or transmigration of persons after death, reincarnation, and body switching are all possible.

However, Locke's insistence on experience-memory as necessary and sufficient for continuity of consciousness and therefore persistence of persons makes his view vulnerable to several objections. The first group of objections targets the claim that memory is sufficient for personal persistence. The most widely discussed of these objections comes from the epistemic character of the memory criterion. For Locke, we are the persons in the past whom we are conscious of having been. Butler (1736) objects to this account, because he takes remembering an experience to imply that one knows that one was the person who experienced the event remembered. He says that memory or "consciousness of what is past" is what "does… ascertain our personal identity to ourselves". For example, "by reflecting upon that which is myself now, and

that which was myself twenty years ago, I discern that they are not two, but one and the same self." (In Perry 2008, 100) Since memory is one's means of confirming whether or not one is identical to a person in the past, remembering having done something presupposes and therefore cannot constitute the fact that one is identical to the person who did it. If remembering that I was some person in the past means that I know myself to be that person, there must be some fact that makes me that person independently of my remembering it. To say I remember being a person in the past already presupposes that I am that person, and therefore cannot be what makes me that person, because in general, knowing that something is the case cannot be what makes it that it is so. "Consciousness of personal identity presupposes, and therefore cannot constitute, personal identity, any more than knowledge, in any other case, can constitute truth, which it presupposes." (In Perry 2008, 100)

Shoemaker, and later, Parfit, attempt to counter this objection by introducing the notion of *quasi-memory (q-memory)*, a faculty like memory but which does not presuppose the identity of the q-rememberer with the person who originally experienced the event q-remembered. Shoemaker defines q-memory as knowledge of

> past events such that someone's having this sort of knowledge of an event does involve there being a correspondence between his present cognitive state and a past cognitive sensory state that was of the event, but such that this correspondence, although otherwise just like that which exists in memory, does not necessarily involve that past state's having been a state of the very same person who subsequently has the knowledge. (In Perry 2008, 253)

Parfit refines this definition of q-memory in the form of three necessary and jointly sufficient conditions. For him, I accurately q-remember having an experience iff: "(1) I

seem to remember having an experience, (2) *someone* did have this experience, and (3) my apparent memory is causally dependent, in the right kind of way, on that past experience." (Parfit 1987, 220) Parfit suggests that if we employ the concept of q-memory, then we can avoid Butler's objection, because q-memory, unlike memory, does not presuppose identity. However, while the possibility that one only q-remembers an experience that one seems to remember implies that one may not be the person who had the experience, whether or not one survives or persists as that person still requires more than remembering or q-remembering. Parfit seems to admit this point when he says that "we should not claim that, if I have an accurate quasi-memory of some past experience, this makes me the person who had this experience," and not just because of the possibility of fission, but because "one person's mental life may include a few quasi-memories of experiences in some other person's life." (Parfit 1987, 222) So if there need be something that makes the difference between really remembering something that one did and only quasi-remembering what someone else did, then Butler seems to be right after all to insist on a separate fact that makes an accurate memory the consciousness of having had a past experience of one's own. Parfit has a potential explanation of this needed fact, but discussion of it requires that we first introduce some objections to the necessity of memory for the persistence of persons.

One such objection is that we can think of countless cases where it is intuitively plausible that memory of having performed an action is not necessary for being the person who performed it. I have forgotten many of my previous actions, but that does

not make it the case that they were not my actions. I don't remember having lunch last Wednesday, but that does not mean that I didn't have lunch that day or that I am not the person who ate that particular meal in the place and time that my body was consuming it. I cannot remember my first day of high school, but that does not mean I did not attend. Parfit attempts to answer this objection by extending Locke's criterion so that "direct memory connections" are not required for persons at different times to be one and the same but only "overlapping chains of direct connections." A person X at t_1 and another person Z at t_2 are directly connected by memory or q-memory when the person Z remembers or q-remembers having done something at t_1 that X in fact did. If Z does not bear any such direct connections to X but does bear at least one such connection to a person Y at some time who in turn bears at least one direct connection to X, then Z is related to X by an overlapping chain of direct connections.

This allowance also enables one who endorses a memory criterion (or any other kind of psychological continuity theory) to answer the objection of Reid (1785), illustrated by the example of "the brave officer" that the memory criterion would, absurdly, allow for the possibility that two persons could both be identical and non-identical to one another:

> Suppose a brave officer to have been flogged when a boy at school for robbing an orchard, to have taken a standard from the enemy in his first campaign, and to have been made a general in advanced life; suppose, also, which must be admitted to be possible, that, when he took the standard, he was conscious of his being flogged at school, and that, when made a general, he was conscious of his taking the standard, but had absolutely lost the consciousness of his flogging.
>
> These things being supposed, it follows, from Mr. Locke's account, that he who was flogged at school is the same person who took the standard, and that

> he who took the standard is the same person who was made a general. whence, it follows, if there be any truth in logic, that the general is the same person with him who was flogged at school. But the general's consciousness does not reach back so far as his flogging; therefore, according to Mr. Locke's doctrine, he is not the person who was flogged. Therefore the general is, and at the same time is not, the same person with him who was flogged at school. (Reid 1785, 248-249)

Put generally, the point is that if some person Z remembers an experience of some person Y, at an earlier time, and Y remembers an experience of a person X, at an even earlier time, but Z has no memory of the experience of X that Y remembers, then Z = Y and Y = X, but Z does not = X. However, because of the transitivity of identity, Z does = X. Therefore, for Reid, the memory criterion is unsatisfactory as a condition of identity as it would imply the possibility of such a contradictory situation. However, Parfit points out that where there are no direct memory connections between Z and X, there could be overlapping chains of such connections that lead from Z to X through Y. If identity (or persistence in one-to-one cases) only requires a relation of the overlapping sort then Z = X after all and there is no problem for transitivity. In the brave officer example, while there are only direct connections between the flogged schoolboy and the standard-taker and between the standard-taker and the general, those connections form an overlapping chain that leads from the flogged schoolboy to the general. If one only requires the overlapping chain, and does not insist on a direct connection, then one can hold that the general is indeed identical to the flogged schoolboy without contradiction.

This same distinction between direct memory connections and overlapping chains of direct memory connections can help respond to the earlier objection to the

necessity of memory that it is counterintuitive to insist that someone must remember having done something in order to be the person who did it. One could say, that even if one does not have a direct memory connection to the experience of acting, so long as there is an overlapping chain leading to the person who had the experience, then one is that person (provided there are no other similarly continuous contemporaries). For example, even if I no longer remember blowing out the candles on my 10th birthday, it can still have been me who did so, so long as I, for example, remember blowing out the candles on my 16th birthday and at the time I was doing that, I did then remember doing so on my 10th birthday.

However, the above revision of Locke's memory criterion together with the concept of q-memory, fails to defeat Butler's objection. To understand why, Parfit's (1984) particular way of trying to evade Butler's objection must be further explained. Parfit distinguishes between psychological connectedness and continuity. The latter is what is necessary and sufficient (in one-to-one cases) for identity, but it depends on more than just some overlapping chains of direct q-memory connections. Psychological continuity requires overlapping chains of *strong connectedness*, which in turn requires that there be a sufficient number (over half) of direct connections that there are from moment to moment in most actual lives:

> My mental life consists of a series of very varied experiences. These include countless quasi-memories of earlier experiences. The connections between these quasi-memories and these earlier experiences overlap like strands in a rope. There is *strong connectedness* if, over each day, the number of direct quasi-memory connections is at least half the number in most actual lives. Overlapping strands of strong connectedness provide *continuity of quasi-*

> *memory*. Revising Locke, we claim that the unity of a person's life is in part created by this continuity. (Parfit 1984, 222)

Parfit claims that continuity of q-memory does not presuppose identity, and therefore can (at least in part) constitute it. He seems to suggest, when he allows that one could q-remember something that some other person did, that the difference between a person at t_2 merely q-remembering an experience had by a person at t_1 and genuinely remembering it is accounted for by the *amount* of direct connections. But it seems implausible that if having some q-memories alone is insufficient for being identical to an earlier person, that adding more q-memories would make a difference. Why can't I q-remember most of another person's life without being that person? If one q-memory doesn't imply identity, then why should a lot of them together do the trick?

Schechtman (1990) correctly diagnoses the confusion inherent in the q-memory approach, along the lines of the point above, by showing that non-delusional memory is not as separable from identity as Parfit supposes, so that the idea that one could q-remember something someone else did is incoherent. She thinks that the q-memory theorist goes wrong by misunderstanding the relationship between the "non-delusionality of a memory" and "its relevance to the constitution of personal identity." (Schechtman 1990, 77) Q-memory cannot capture what in memory is relevant to personal identity, such that q-memory might constitute personal identity, without presupposing the identity of the q-rememberer with the person who had the experience, because the qualities essential to experience-memories that make them important to our sense of ourselves necessarily refer to us as individuals due to their dependence on

many of our other mental states. By way of illustration, Schechtman provides an excerpt from a book called *Remembering: A Phenomenological Study* by Edward Casey, in which Casey recalls in detail, some vivid, some murky, going to see the movie *Small Change* with his children. Here's a small sample:

> Anticipating a large crowd, we arrived early and were among the first to purchase tickets. There ensued a wait that seemed much longer than the ten or fifteen minutes it actually was. The children were especially restive and had difficulty staying in the line that had formed - Erin attempting some gymnastic tricks on the guardrail by the entrance, Eric looking at the posted list of coming attractions… Once inside, we sought seats approximately in the middle of the theater, settled there, and interchanged positions a couple of times to adjust to the height of those sitting in front of us. The lights dimmed, and *Small Change* began directly. (Or was there not a short feature first? -- I cannot say for sure.) The film was in French, with English subtitles. I have only a vague recollection of the spoken words; in fact, I cannot remember any single word or phrase, though I certainly remember the characters *as speaking*. The same indefiniteness applies to the subtitles, at which I furtively glanced when unable to follow the French. Of the music in the film I have no memory at all -- indeed, not just of *what* it was but *whether* there was any music at all. In contrast with this, I retain a very vivid visual image of the opening scene, in which a stream of schoolchildren are viewed rushing home, seemingly in a downhill direction all the way. (quoted in Schechtman 1990, 80)

According to Shechtman, if such a q-memory qualitatively identical to Casey's memory were implanted in another person, Jane, "the amount of personal detail… makes it difficult to imagine Jane receiving it as a quasi-memory," for the memory "contains a good many elements that make reference other parts of [Casey's] life and his personality." (Schechtman 1990, 81) Features such as Casey's familiarity with the theater, his knowledge of French, and most importantly, his relationship with and feelings about, his children "are going to be very alien to Jane," who, we can stipulate, has none of that knowledge or those feelings. (Schechtman 1990, 81) Memories get

their character at least partly from their relations to other mental states which imply reference to the identity of the rememberer. Schechtman thinks we have to imagine Jane's q-memory experience in one of two possible ways.

> The first is that Jane will reproduce all of the visual [and I don't see why not to include auditory, olfactory, but perhaps not tactile or kinesthetic?] content of the memory without interpreting it as Casey does. That is, upon awakening from the quasi-memory implant surgery, Jane will have images of being in a… theater, with a woman and two children who she does not recognize, and she will also have images of seeing a movie with these people. The second alternative is that she will reproduce the memory exactly as it occurs in Casey, with all of the same personal elements and associations. (Schechtman 1990, 82)

As Schechtman sees it, the former experience would not capture what "is relevant to personal identity in genuine memory connections," (Schechtman 1990, 82) so that q-memories of that sort could not be constitutive of personal identity, no matter how many of them were shared between persons at different times. It wouldn't even be accurate to say that such a q-memory is qualitatively identical to Casey's memory, since it would be missing much of the context that gives the memory its qualitative character. However, if Jane's q-memory did include all of the personal elements of Casey's memory, then its accuracy or non-delusionality would not be separable from the assumption of identity with Casey. "If… we really wanted to reproduce the qualitative content of Casey's memory in Jane, we would not only have to recreate a great many of Casey's states in Jane, but suppress a great many of Jane's as well, and this begins to looks suspiciously like replacing Jane's psychology with Casey's." (Schechtman 1990, 84) In such a case, Jane would have to, delusionally, believe herself to be Casey. So Schechtman concludes, "The fact, then, that presuppositions about who has a memory

are inseparable from its content means that one cannot, as Parfit claims, specify non-delusionality impersonally by keeping the content of a memory and simply deleting propositions about whose memory it is." (Schechtman 1990, 84) It is difficult to imagine how Jane's q-memory could be true to Casey's memory without presupposing that Jane becomes Casey whenever she q-remembers it.

Parfit could reply that even if Jane must believe she is Casey, and be mentally exactly similar to him while recalling the q-memory, that does not imply that she must *be* Casey. But that would rule out the idea that identity depends on the number of direct q-memory connections, as Jane would need to have at least as many connections as Casey had with his theater-going self in order to non-delusionally experience a single one of his q-memories.

A further problem for using memory as a criterion for identity comes from recent plausible theorizing in neuroscience, which has been put to use in attempting to reduce negative emotional consequences of traumatic experiences, that memories are re-encoded with new information while missing some of the old, every time that they are recalled, so that the content of a memory is never identical to the content of an experience nor to the content of any subsequent remembering. (Schiller et. al 2010 and Hall 2013) Different features of a past experience will be salient and others drop into the background or even out of recollection entirely as one's beliefs and desires change over time and one acquires memories of new experiences. This would mean that direct connections cannot be understood, as Locke seemed to hold, in terms of sameness of

content.

Furthermore it seems that the continuity of many psychological attributes which are central to our sense of ourselves as individuals does not depend on our experience-memories. There are cases of full episodic amnesia where the amnesic retains many other psychological attributes such as beliefs and desires and other propensities. For example, as mentioned in chapter one, Craver (2012) has reported that victims of Korsakoff's disease, some of whom have no capacity for episodic memory still have strong senses of themselves, what they are like, what they want from life, etc.

> Perhaps surprisingly, individuals with episodic amnesia often show considerable constancy of character. KC, for example [who lost all of his episodic memories as well as his capacity to form new ones in a motorcycle crash], prefers the Price is Right and M*A*S*H to other television shows, Black Label to other beers, and the Toronto Maple Leafs to other hockey teams. He is courteous and quiet, but lethargic and forgetful. He has a sense of humor and a pleasant smile. He is a bit flat, but this facilitates a subtle charm. KC has a personality. The persistence of this personality requires or is constituted by a rich set of causal connections between earlier and later mental states. Such connections contribute no less than episodic memory to his continuity over time in the neo-Lockean view.
>
> The main point is this: The simple neo-Lockean formulation of N [the "episodic necessity hypothesis": that episodic memory is necessary for one to be, have, or maintain a self in some significant sense] holds that episodic memory is necessary to connect conscious experiences at different times, which connections constitute the diachronic identity of the self [Craver uses 'self' more or less interchangeably with 'person']. This hypothesis has evolved in response to the threat of circularity such that episodic memory no longer plays a necessary role in the identity of persons over time. If episodic memory's contribution to diachronic identity is as thin as the contribution made by q-memories, it is a contribution that, as a matter of empirical fact, many other cognitive and bodily systems make as well. The most viable surviving relative of the Lockean formulation of N thus fails to support the view that individuals with episodic amnesia lack identities over time. (Craver, 458)

Individuals like KC provide evidence that memory (of the episodic sort, which is

what proponents of the memory criterion seem to have in mind), besides not being constitutive of personal persistence, is also unnecessary for it. An individual with no capacity for episodic memory or episodic thoughts of the future, may still evidence a definitive set of character traits and preferences and may still value some future possibilities over others. This seems to undercut the idea that continuity of consciousness depends on the capacity for episodic memory. Most of the features we care about persisting over time are independent of it. One might instead claim that it is memory of the semantic sort, i.e. memory of facts, that is necessary for personhood, but it is difficult to see how such memory is distinguishable from a capacity to form beliefs in general.

III. The expanded psychological criterion

Memory won't serve as the sole criterion of psychological continuity, and therefore the persistence of a person over time, but one can, as Parfit does, buttress the continuity and connectedness relations by appealing to various other relations between temporally disparate psychological states.

Parfit's (expanded) Psychological Criterion holds that continuity of distinctive beliefs, desires, character traits or behavioral dispositions is necessary and sufficient for a person to persist over time, though the degree of a person's responsibility for a previous action depends on the number of direct psychological connections between the person now and the person who performed the action in the past. Direct

psychological connections require sameness of distinctive psychological states between persons at different times, whereas psychological continuity only requires overlapping chains of such connections, such that two persons may have no distinctive psychological states in common and yet be continuous by virtue of both sharing different sets of distinctive psychological states with a person at a third time.

The Psychological Criterion essentially consists of the claim, (D), that continuity of distinctive psychological states is necessary for personal persistence, which can be formulated more precisely as follows:

(D) If a person X at t_1 persists as a person Y at t_2, then Y's distinctive psychological states are continuous with the distinctive psychological features of X.

Following Greenwood (1994), I take it that the states in question must be dispositional rather than occurrent. Occurrent states are ones of near instantaneous duration, generally conscious, which come to one's mind in the flow of mental life, when one thinks such and such to oneself. Locke rightly rejects these sorts of thoughts as accounting for persistence, because of their very transitory nature, and as Unger (1990) puts it "There is no single, occurrent mental phenomenon, such as a conscious, self-referential thought, that any of us has at every moment of her existence. (1990, 206) However, Locke is mistaken in thinking that all mental states are of this occurrent, hence momentary, nature. Most mental phenomena are not occurrent, but dispositional. The vast majority of my beliefs, desires, doubts, judgments, etc. do not occur to me at every particular moment, but nonetheless persist as relatively stable aspects of my

psychological makeup. If any distinctive mental phenomena are to serve as criteria of personal persistence, they will be those of the dispositional sort.

The insistence on continuity or connectedness of distinctive psychology is natural given the close connection between our sense of ourselves as unique individuals and our sense of ourselves as being primarily psychological individuals. When asked for an account of oneself, one's natural tendency is to offer a description of the qualities that differentiate one from others. Among these distinctive qualities, the psychological ones appear to be of the greatest importance. One feels that most of the features of one's body are relatively cosmetic or accidental, when compared with one's beliefs, desires, goals, wishes, values, aspirations, moral commitments, emotional dispositions, etc.

Given an account of personhood that requires for being a person possession of higher-order desires and beliefs, it might seem as if continuity of those sorts of psychological attributes would be of greatest importance for personal persistence. If my ability to form preferences as to what I should desire and believe, i.e. my capacity for evaluation of my own first-order psychology, is what makes me a person, then it is tempting to infer that the distinctive preferences I have and evaluations I've made make me the specific person that I am. Along those lines, one could regard a person as persisting so long as his or her higher order beliefs and desires remain relatively stable, forming a character with a specific set of values. If those values were to radically change, we might say, and often do, that it is no longer the same person we are dealing with. For instance if an individual who had been an honest, generous, pacifistic liberal

were to wake up one day as a lying, greedy, war mongering conservative, a friend might think that this was not the same person who had gone to sleep the night before. Empirical studies by Strohminger and Nichols (2014) suggest that moral character is the most central of all distinctive psychological features to people's sense of their own persistence and their sense of the persistence of others.

However, there is a countervailing idea about persons, which forms part of the initial basis for worrying about persistence over time in the first place. It is characteristic of persons that they are highly flexible in the distinctive psychological characteristics they possess over the course of their lives, including the higher-order ones. While it is not necessary to being a person that one's distinctive psychological attributes change over time, one might reasonably claim that most persons vary, sometimes smoothly and gradually, sometimes radically and suddenly (e.g. in response to a traumatic or serendipitous event), psychologically throughout the course of their lives. Indeed, one of the primary virtues of a Reductionist approach to persons, one that does not require the existence of a transcendent, immutable soul, is that it allows for persons to persist through such changes. The intuitive appeal of the idea that continuity of distinctive psychological characteristics is necessary for personal persistence is lessened if it is true that such characteristics are more variable than one supposes.

Furthermore, an account of personhood in terms of the possession of capacities lends itself to the principle that so long as a being maintains the capacities constitutive of being a person, i.e. remains a person, then that being must be the same person all

along. Our treatment of watches, trees, and frogs seems to follow this principle, as did Locke's account of all organized beings other than persons. Locke thinks that persistence conditions for a thing should be given by the sort of thing it is. However, in the case of persons, Locke abandons the principle that so long as a being remains the same sort of being it remains the same individual being. On Locke's account, an individual may continue being a person without remaining the same person, because e.g. if some individual person loses all of her memories and gains new ones, there is an organized being in a single place throughout the process with the capacities necessary and sufficient for being a person, but for Locke, there is a different person before and after the change in memories.

Bernard Williams (1970) offers a pair of thought experiments that, so framed, are supposed to elicit the "Lockean" judgment that two individuals have switched bodies, and the "non-Lockean" judgment that no switch has occurred, respectively. The following is a presentation of the first thought experiment, though slightly altered so that it tests Parfit's expanded Psychological Criterion, rather than Locke's Memory Criterion, by referring to all distinctive psychological features rather than memories alone and speaks of feature-swapping, instead of brain-switching, to better highlight the differences between the Psychological Criterion and the Core Psychological Capacities criterion:

If a person, X, were to swap all distinctive psychological features with those realized in a different person, Y, so that X-body would end up with the psychological

features previously realized in the Y-body and vice-versa, the common intuition seems to be that person would then have the Y-body and person Y the X-body, such that X should, post-swap, fear torture administered to the Y-body and Y fear torture administered to the X-body. This way of telling the story clearly motivates the view that continuity of distinctive psychological attributes is necessary (as well as sufficient) for personal persistence.

Here is the second situation imagined by Williams, altered in a similar manner to the first, by making reference to the erasure and replacement of distinctive psychological attributes in general, rather than just memories, and assuming that the mechanism of erasure and replacement leaves core psychological capacities uninterrupted: Imagine that a scientist tells you that while you sleep she will remove all traces of your distinctive psychology, leaving a being with the capacity for self-consciousness and concern but without any of your distinctive beliefs, desires, etc. of any order, that you currently possess. The resulting person will have instead completely different distinctive mental states, perhaps even diametrically opposed to your own before the operation. The person who remains will be tortured. Your distinctive psychological states will be reproduced in the brain of a person far away, replacing the characteristics that person already possesses. If it is rational for you to fear the torture, then you must persist as the person who will be tortured. However, one who endorses (D), which holds that continuity of distinctive psychology is necessary for personal persistence, should not fear the torture in the above scenario.

Deciding what is intuitive and what isn't is, of course, a precarious enterprise. Some work has been done, e.g. by Bruno and Nichols (2010), trying to assess the ways people respond to both ways of framing the "future pain" scenario. The results seem to confirm Williams' assumption that when presented with the *Lockean* frame in which two persons' bodies swap brains, or swap distinctive psychological characteristics in some other way, most respondents conclude that the two persons have really switched bodies, while when presented with the *pain* frame cited above they conclude that one should fear future torture even when one's distinctive psychological characteristics have been completely erased and replaced (whether or not they are implanted in another body somewhere else). Bruno and Nichols conclude that the responses to the latter scenario, the *pain* frame, are due to unfair demands that it places on respondents:

> In that frame, there seems to be a demand to respond that I would feel the pain. After all, if I am not going to feel it then who is?... There is plausibly pressure here to give a persistence response. If this is right, then if we remove or decrease any thought experimenter demand, we should find less inclination to give the persistence response. (Bruno and Nichols 2010, 17)

However, this conclusion is unwarranted. The respondents might think that the person who feels the pain is a brand new one. That they do not only confirms that it is counter-intuitive to think that a complete change in distinctive psychology yields a numerically different person. Furthermore, even when the gathered data suggests that philosophical laypeople accept (D), that acceptance seems to have a limit, and that limit, as Williams predicts, is the anticipation of future pain or death. For example, while studies (Bartels and Urminsky 2011, Bartels et al. 2013) have found that people tend to

discount future rewards for individuals physically continuous but distinctively psychologically dissimilar to themselves, they have also found (Bartels et al. 2013) that anticipated change in distinctive psychology yields no reduction in death anxiety. There is much more that experimental philosophy can do on this subject. However, the work so far shows that folk intuitions indeed are indeed divided over the two frames.

Sider (2001), Shoemaker (2007) and others have claimed that the intuitions for and against (D) are intractably opposed, such that "there exist two candidate meanings for talk of persisting persons, one corresponding to each criterion, and there is simply no fact of the matter which candidate we mean." (Sider 2001, 1) However I am not so pessimistic about the situation and will endeavor to show that the latter should be more compelling than the former once the issue is duly clarified.

IV. What really matters in persistence

Unger (1990), has the intuition that one should fear the torture in the pain frame scenario, because so long as core psychology is maintained in an individual person, she is numerically the same person no matter how radical the change in distinctive psychology. Unger follows Parfit and others in considering the question of "what matters in survival (i.e. persistence)" in the thought experiments under consideration, but urges that there is ambiguity in that phrase. First of all there is the "desirability use" which Unger associates with (D) proponents such as Parfit and Lewis, but which he himself thinks is "not highly relevant to questions of our survival" (Unger 1990, 196). According

to Unger the desirability use amounts to the question of

> what it is that one gets out of survival that makes continued survival a desirable thing for one, a better thing, at least, than utter cessation. On the desirability use, if one has what matters in survival, then, from a self-interested perspective, one has reason to continue rather than opt for sudden, painless, termination. (Unger 1990, 196)

The reason for thinking the desirability use irrelevant to the question of persistence is that it seems possible that one might continue to exist even if one found that existence wholly undesirable. It is likely that most people would find a sudden, radical change in their psychology, particularly as concerns their values, to be highly undesirable, because they would not wish themselves to be a very different sort of person, not only for the sake of others, but for their own sake. I hope I don't grow to be a miserly curmudgeon in my old age, but it would be irrational to believe it impossible that I might become like that given various circumstances such as the sudden accumulation of massive wealth and power. I would hope that my current character would survive, but if it does not, it will be I who is corrupted. Even if that change came about not through natural circumstances but by the unnatural intervention of the scientist, it is rational to worry that it may be I who wakes up with undesirable distinctive psychological attributes, just as I may be worried that I suddenly wake up with highly undesirable physical qualities, such as disfigurement or dismemberment or even that someone else I know might change psychologically or physically.

Next, Unger considers the *prudential* use of "what matters in survival," which he glosses as follows:

> From the perspective of a person's concern for herself, or from a slight and rational extension of that perspective, what future being there is or, possibly, which future beings there are, for whom the person rationally should be 'intrinsically' concerned. Saying that this rational concern is 'intrinsic' means, roughly, that, even apart from questions of whether or not he [sic] might advance the person's present projects, there is the rational concern for the welfare of the future being. (Unger 1990, 196)

Unger thinks that it is only this latter use of "what matters in survival" that has bearing on the metaphysical question of personal survival. This is because, while the desirability use only pertains to what we want out of life, the prudential use also pertains to what is undesirable yet no less real. "Very roughly, the desirability use aims at just those situations that we should most like to encounter, while the prudential use aims at all those that, somewhere or other in logical space, must be faced." (Unger 1990, 197) In other words, the desirability use is about what we positively value in our survival, whereas the prudential use includes attributes that we might positively or negatively value. A third use of "what matters in survival" is purely *constitutive*, such that "we focus on what *counts toward* the case being one that involves a person who does survive," and are not directly concerned with "the evaluative, or the motivational matters that surround the topic of survival," and therefore "this use has no direct connection with questions of rational concern for oneself in the future." (Unger 1990, 197-8) In other words, the constitutive use has nothing to do with what matters to the person who survives, but rather, only what matters to us, the metaphysicians. Unger thinks that there is an important connection between the prudential use and the constitutive use, the former being the latter's "motivationally relevant counterpart" (Unger 1990, 199), in the sense that it describes the features constitutive of our persistence that we, the

persisting, might be concerned about, but that there is no such connection between the desirability and constitutive uses. Unger diagnoses the intuition that psychological continuity of distinctive psychological attributes is necessary for personal persistence as a mistake that results from employing the desirability use of “what matters in survival” when seeking insight into the constitutional facts of persistence, when we should be employing only the prudential use.

For Unger, Williams succeeds in revealing this mistake when he changes perspective from the earlier third-personal *Lockean* frame to the first-personal *pain* frame. In place of (D), Unger offers what he calls the “Physical Criterion”, the main component of which is the claim that only the maintenance of core psychological capacities is necessary for personal persistence (C).

To say that a capacity is maintained, for Unger, is to say something about the nature of the physical properties which underlie it in a physical being or succession of such beings, namely that there is a suitable structure such that the capacity is realized:

> My basic mental capacities will exist from now until a future time only if, from now until then, they are continuously realized in some physical entity or, at least in an appropriate succession of physical entities. In largest measure, this is just a brute fact about the relations between myself, mentality, and the objective world order. Now, while both of us are similarly objective physical beings, and while both of us have precisely similar basic mental capacities, you and I are different people. So, at least during some of the time that you exist, and perhaps during all of it, your mental capacities must be realized in one physical entity, or one succession of them, while my capacities are realized in another. (Unger 1990, 206)

So (C) can be stated more precisely as follows:

> (C) If a person X at t_1 persists as a person Y at t_2, then Y contains the physical realizer of X's core psychological capacities or a physical realizer continuous with the physical realizer of X's core psychological capacities such that the realization of those capacities has been uninterruptedly maintained from t_1 to t_2.

If one's core psychological capacities are maintained over time, then, according to Unger, what distinctive psychological qualities one has from moment to moment is irrelevant. This view allows that persons are psychologically flexible, and can vacillate gradually and conservatively or dramatically and radically in what they believe and value from one moment to the next. Consistency might be a virtue in a person, but is not a requirement, so that continuity of distinctive psychology is not necessary for personal persistence. Therefore, in a situation like Williams' *pain* frame, the fact that the person tortured will not share any of your distinctive psychological characteristics should not make you fear the torture any less. The intuition generated by the *Lockean* frame, then, rests on not only mistaking the desirability use of "what matters" for the prudential use, but also from thinking that a sudden change in distinctive psychological features yields a new person, despite the continued maintenance of core psychological capacities throughout the change. If swapping distinctive psychological features required that the structure which instantiates core psychological capacities be dissolved and reconstituted, then there would be a new person, because the old one would have perished.

If a proponent of (D) wishes to embrace the premise about the psychological variability of persons in ordinary, gradual cases of change, but deny that persistence occurs in the extreme cases of sudden and radical change, she must draw a principled line where a sufficient qualitative difference yields a numerical difference between one person and another. Parfit's distinction between psychological connectedness and continuity is already an attempt to draw such a line by saying that a person X persists as person Y, so long as Y has at least half the number of direct psychological connections with X as a person has from moment to moment in most actual lives or else there is an overlapping chain of at least so many connections between persons at any pair of times leading from X to Y. However, in the first place, any specified necessary amount of connections seems arbitrary. A single connection yielding 50% connectedness, rather than 49%, should not make such an important ontological difference. Secondly, the move toward requiring for personal persistence only continuity instead of direct connectedness seems to be a rather ad hoc fix, which gives up the very intuition that made the Psychological Criterion plausible in the first place. If my distinctive psychological features are essential to who I am, I shouldn't think that overlapping chains of connected features should secure my persistence as some future person with whom I have no distinctive psychological features in common. Such thinking is at least not implied by the intuition that my distinctive psychology is essential to me. Furthermore, in the cases where an individual's distinctive psychology changes all at once, in a moment of total revelation, for instance, there would not even be the

overlapping chains of connections necessary, on Parfit's view, to secure continuity. And in a case where my psychology changes completely at t_1, but temporarily so, so that it is later restored at t_2, there would be psychological continuity between the person before t_1 and after t_2, but not between the person between those times and the person before and after, so the proponent of (D) would have to claim that the person before t_1 ceases to exist at t_1 but then comes back into existence at t_2. Here, I will only say that I find this possibility *prima facie* absurd, but in the next section I will provide an argument for why such intermittent existence is in principle impossible.

Besides the considerations already discussed, a further reason for accepting (C) and rejecting (D), is that (C) is compatible with ordinary judgments of continued responsibility in a way that (D) is not. According to (D), someone who wakes up tomorrow with my body which has uninterruptedly maintained its core psychological capacities, but who has radically different distinctive psychological attributes from those I have today, would be responsible for the things I have done. My intuition is that this is the right view about continued responsibility. Others may not share it, but as with the memory criterion, I don't see why my radically changing my mental nature should change whether or not I self-consciously performed some action in the past. Moreover, actual legal practice does not take sufferers of retrograde amnesia or individuals who go through personality overhauls to no longer be responsible for past actions. Imagine if that were the practice, and there were a procedure for criminals to have their memories of their crimes wiped along with their desire to steal. They would then have an easy way

to be acquitted of their deeds, arguing that they were no longer the persons who committed them and therefore no longer responsible for the crime.

Furthermore, consider what happens when an individual recognizes his or her responsibility for an action and takes responsibility for it in a way made explicit through words and actions, which, in cases where one regrets those actions, leads the way to atoning for them. Such atonement is what leads us to forgive a person, though that atonement is made possible where it was previously not, because the person's deeply held beliefs, desires, and values change in a way which allow him or her to reconsider the ethical nature of the action performed (even if he or she does not remember having performed the action). For instance, I might come to realize that my setting fire to a trash can outside my school was not an admirable act of righteous rebellion, but a childish cry for attention that was thoughtless and potentially extremely harmful. My change in values did not make it such that I was no longer responsible for the action, rather it only allowed me to consider my responsibility in a more sophisticated, ethically transformed way. Such an improved ethical understanding of my previous actions requires that I recognize them as my actions, whether the improvement happens gradually over time or all at once in a character transforming moment of insight. Therefore, I maintain that continuity of distinctive psychological attributes is not necessary for the persistence of persons in a way relevant to continued responsibility. In fact, I take it as the great liberating insight of reductive approaches to personhood, such as those maintained by many Buddhists, that persons are highly mutable and that it is

irrational to think of oneself as bound to a distinctive psychological essence.

Parfit's views on desert and responsibility are a bit difficult to sort out, but are, nevertheless, important to consider. He holds that while persistence requires only psychological continuity, desert is a matter of the degree of direct psychological connections between a past and future person: "When some convict is now less closely connected to himself at the time of his crime, he deserves less punishment. If the connections are very weak, he may deserve none." (Parfit 1984, 326) However, he seems to distinguish this claim from one of 'diminished responsibility,' though he does not quite make clear what that distinction amounts to. He continues:

> This claim should be distinguished from the idea of diminished responsibility. It does not appeal to mental illness, but instead treats a criminal's later self as like a sane accomplice. Just as someone's deserts correspond to the degree of his complicity with some criminal, so his deserts now, for some past crime, correspond to the degree of psychological connectedness between himself now and himself when committing the crime. (Parfit 1984, 326)

I am tempted to interpret this passage in a way that puts Parfit in agreement with my own view about the relationship between desert, responsibility, and distinctive psychology; namely, that desert might diminish with distinctive psychological changes, but not responsibility. However, this does not seem to be what Parfit has in mind, for he goes on to say:

> We may be tempted to protest, 'But it was just as much *his* crime.' This is true. And this truth would be a good objection if we were not Reductionists. But on the Reductionist [I take both of our views to be 'Reductionist' in the sense that we both explain personal persistence entirely in terms of more basic relations] view this truth is too trivial to refute my claim about reduced responsibility. It is like the claim, 'Every accomplice is just as much an accomplice.' Such a claim cannot show that complicity has no degrees. (Parfit 1984, 326)

So Parfit does seem to think that desert and responsibility go together, but not responsibility and persistence/identity. Still, on this point, I agree with Parfit in spirit, if not in letter. For various reasons, I may deserve less punishment over time for something that was, nevertheless, done by me. I depart from Parfit, however, in thinking that we may and do forgive persons for things they have done, despite their being no less responsible for having done them.

V. Persistence and Causation.

So far I have established my agreement with Unger that it is maintenance of core psychological capacities, rather than continuity of distinctive psychological attributes that matters, in the metaphysically relevant sense, to personal persistence over time. In other words, continuity of distinctive psychology is not necessary for personal persistence, but maintenance of core psychological capacities is. Unger argues that physical continuity is also necessary for personal persistence, because as a matter of empirical fact, maintenance of core psychological capacities, at least in the case of human persons, requires the continued existence and functioning of a physical object, namely a brain or central nervous system. To allow for some other imaginable kinds of persons, Unger offers the following general formulation of the 'Physical Criterion' for persistence of persons:

> The person X now is one and the same as the person Y at some time in the future if, and only if, (1) there is sufficiently continuous physical realization of a core

> psychology between the physical realizer of X's core psychology and the physical realizer of Y's core psychology...) (Unger 1990, 202)

The above criterion reveals the second way in which Unger disagrees with proponents of the Psychological Criterion. According to most versions of the Psychological Criterion, there need be no continuous physical *realizer* of the relevant psychological qualities, but only continuity of the qualities realized. The science fictional thought experiment of teletransportation, as introduced by Parfit (1984), can be used to illustrate the version of the Psychological Criterion which holds that connectedness of distinctive psychology requires no special causal connection between the distinctive psychological features of a person at one time and another, but only similarity : If I were to be "teletransported" in the following way: the arrangement of the parts composing the physical realizer of my psychology (e.g. my brain), down to the atomic level were recorded and then destroyed, but then an individual were created in another location whose physical realizer of his or her psychology were constructed out of entirely different matter according to the exact arrangement recorded from my physical realizer in such a way that psychological qualities exactly similar to my own were instantiated in this individual, then according to Parfit, the resultant individual would be me. However, for Unger, since the physical realizer of the resultant individual's psychology would not be continuous with my own, the necessary continuity of core psychological capacities would not be maintained and therefore I would not persist as the resulting person, no matter how similar in distinctive psychology. Unger, therefore, disagrees with the Psychological Criterion, not just about which sorts of psychological features must be

continued or maintained, but about what that maintenance must consist in. Specifically, they differ over what the causal connections must be between persons at different times such that one may persist as the other. For Unger, the disagreement over causal requirements follows from the disagreement over essential features.

Unger endorses the following principle: that persons, like other concrete particulars, do not admit of intermittent existence. In other words, a person cannot cease to exist and then come back into existence no matter how instantaneous the duration of the interval between. This would imply that resurrection is, in principle, impossible. Unger calls this "the condition of no interruption," (Unger 1990, 205) which I will from here on refer to as (N) and define below:

(N) If X persists from t_1 to t_2, X must exist at all times between t_1 and t_2

As explained above, Unger rejects the view that continuity of distinctive psychology is necessary for personal persistence, in favor of the view that all that is necessary, psychologically speaking, for personal persistence, is the maintenance of core psychological capacities. He thinks this requires the capacities' uninterrupted continuance in a physical realizer or succession of them. The second disjunct is meant to allow for the possibility that one physical realizer might gradually replace its components with ones of a different material, e.g. organic neurons to artificial nodes, without interrupting the processes instantiating the capacities. In such a case the person would persist, even though the physical realizer does not. Unger also allows that capacities might be maintained even when they have been suspended for an indefinite

period of time, such as in the case of super-freezing and thawing. (Unger 1990, 192)

One way to put this point, is to say that a capacity need not be *poised* for use in order to be maintained, but may be temporarily *suspended* such as in the case of someone in a deep, but reversible coma (whose capacity for self-consciousness may be suspended), or one with temporary akinetic mutism (whose capacity for concern seems to be suspended - see Prinz 2012 for discussion of this condition). This is a further reason for rejecting (D), since it is not clear that distinctive psychological features, such as beliefs, desires, memories, and values are continuous through such conditions, while core capacities are. In a coma I may not believe or desire anything at all, but so long as the coma is reversible, my capacity for self-consciousness may be maintained, assuming that the necessary cerebral structure remains intact.

Unger does not claim that there is a conceptual or logical connection between maintenance of core psychological capacities and physical continuity. He merely makes an empirical claim about the sorts of processes that underlie the maintenance of persons' core psychological capacities and therefore persistence, i.e. physical processes of the brain and central nervous system, and then generalizes that claim to allow for imaginable variations that are admissible so long as they do not depart from his general assumption of the "correctness of a certain view of reality as a whole… as being, at the least, very largely a physical world..." which

> is reasonably stable, regular, and well-behaved: For example, like rocks, trees, and cats, people do not, along with their matter, pop out of existence, or pop into existence. Rather, people begin, continue, or end, as a consequence of the arrangements of certain comparatively simple physical things. (Unger 1990, 203)

For Unger, (N), is if not an absolute condition of personal persistence or survival, "then, at least... provides a strong guideline for any adequate account of our survival." (Unger 1990, 205) Endorsing (N) and (C) as well as the view of reality quoted above, Unger derives his Physical Criterion of personal persistence. Unger sums up his reasoning as follows:

> For you to exist at a future time, you must exist, continuously, from now until then. For that to be so, there must be the continuous existence, from now until then, of your particular basic mental capacities. For there to be the continuous existence of just those capacities, there must be, in this wholly or largely physical world of ours, the continuous physical realization of them in a physically continuous realizer, or at the least, in a physically continuous succession of physical realizers. Consequently, for you to exist at a future time, there must be appropriate physical continuity. (Unger 1990, 207)

However, I think there is a more powerful argument for (N) than Unger has provided, which is that it follows from (C). Maintenance of the core psychological capacities of persons implies that they be uninterrupted. (C) implies (N) because one can only individuate tokens of a capacity in terms of its uninterrupted maintenance over time. Since self-consciousness and concern are capacities shared by all persons, one token instantiation of self-consciousness or concern can only be distinguished from another by its continuous maintenance by the concrete particular or continuous series of concrete particulars that realizes it. Saying that a person persists by maintaining her core psychological capacities already implies that the person must exist uninterruptedly. By accepting (C), one implicitly commits oneself to (N). Since, as argued above, there are good reasons for accepting (C), one should, therefore, also accept (N).

Like causal continuity in general, physical continuity of a thing can be understood in a narrow sense where we only "think in terms of its constituting matter through ordinary space with respect to time," (Unger 1990, 203) or in a "wide" sense that is meant to allow for various more exotic physical possibilities including the possibility that there "be other physical dimensions in which, during a certain interval, the individual's matter exists." (Unger 1990, 203) The wide sense involves a similarly liberal conception of 'matter' according which "some matter will be any portion of physical reality, regardless of state, that is suitable for constituting (wholly or largely) physical individuals." (Unger 1990, 203) However, Unger is only interested in cases that have "some basis in reality," the reality presupposed by the general worldview cited above, which would rule out cases of interrupted existence.

The primary implication of Unger's Physical Criterion is that it rules out as cases of persistence those scenarios where a person's matter is replaced all at once. In ordinary cases a person's matter is gradually replaced over time as cells die and new cells are generated, such that over a person's lifetime their cells will be completely replaced several times. However, the gradualness of the replacement is what allows the individual to persist. To be clear, the point is not about rapidity or slowness of the replacement but the gradualness or abruptness. "As long as they are relevantly even and gradual rather than uneven and abrupt, I can survive the most rapid of complete serial replacements." (Unger 1990, 211) His wide Physical Criterion is in essence a version of (N) that presupposes the generally physicalist worldview.

The rejection of (D) in favor of (C) leads Unger to understand the necessary causal relations that allow persons to persist over time to be equivalent to the relations involved in physical continuity, because the maintenance of core capacities is best understood, given the widest interpretation of the empirical evidence at our disposal, on the model of physical continuity with its assumption of (N). This view would rule teletransportation, but would allow for gradual replacements of organic by cybernetic parts, and even gradual assumption of one's core psychological capacities by a computer, so long as that hard drive were capable of supporting a self-conscious and concerned being (assuming those are the capacities necessary and sufficient for personhood) with no interruption to the maintenance of its capacities. It would also support most common intuitions about brain switching cases (that the person goes with the brain). If the brain is the realizer of a person's core psychological capacities, and it is kept in a suitable condition throughout the process, then one person could exchange bodies with another through a brain transplant. Similarly, if such thing as a soul atom existed, a single particle that could by itself realize the core psychological capacities of a person, then a person could exchange bodies by transferring that particle from one to the next, so long as the soul particle continued to realize those capacities.

VI. Abstracting from the Physical

I agree with the spirit and most of the matter of Unger's empirically hedged criterion, though it is misleading to call Unger's view the "Physical Criterion" as the maintenance of core psychological capacities is what is most crucial for personal persistence. Furthermore, although there is no conceptual connection between core psychological and physical continuity, there is a conceptual connection between core psychological continuity and whatever realizes core psychology. I take the primary reason for commitment to (N) to be independent of the physicalist worldview Unger presupposes. That is because, as I argued in the previous section, (N) is implied by the conception of the persistence of persons in terms of the maintenance of capacities and by the nature of concrete particulars in general, not only by the nature of persons as physical objects. So to be more metaphysically neutral, and therefore ascend to something more like a conceptual truth, which I think is desirable, one should leave open the possibility that there be some non-physical element of a person, in the sense that it does not obey physical laws or cannot be detected by physical senses or instruments, that is the realizer of a person's core psychological capacities In such a case, a person would persist so long as the realizer or series of realizers of core psychological capacities continue(s) to maintain those capacities. One can thereby insist on just (N) and not the physical criterion to hedge one's metaphysical bets. That would be in keeping with the spirit of Locke's decision to look at the sort of psychological features that underlie personhood and the persistence of persons rather

than the sort of substance. One would then say that for a person to persist, such a non-physical element would have to obey (N) and be causally continuous in a way equivalent to physical continuity.

Locke accounted for the persistence of most complex bodies, such as oak trees, frogs and watches, in terms of the maintenance of an organized structure that yields the capacities essential to the being in question. However, he refused to extend this account to persons, because, as his example of the Prince and the Cobbler shows, he insisted, without argument, that distinctive characteristics, memories in particular, are necessary and sufficient for persistence. However, if he had seen core psychological capacities instead, as criteria of persistence, he could have extended his analysis of other organized bodies to persons.

Therefore, I think Locke should have said about persons what Reid suggested, which was that as long as a being uninterruptedly meets the conditions for being a person, it remains the same person. (Reid 1785, 113) However, Reid assumed that such persistence requires the continued existence of a simple, unchanging, immaterial soul that is distinct from any of the person's thoughts, actions, or body parts, whereas I do not.

For Reid, X is a person at t_1 and persists as a person Y at t_2 iff:

1. X has a soul with the capacities constitutive of personhood at t_1

2. Y has a soul with the capacities constitutive of personhood at t_2

3. Y has the same soul at t_2 as X has at t_1, which means that that the soul in question has continuously existed, maintaining the capacities constitutive of personhood, at all times between t_1 and t_2

My account of persistence is what Locke should have said about the persistence of persons if he had extended his account of the persistence of other organized beings to them.

For Locke, a frog (one example of an organized being) X at t_1 persists as (or is identical with) a frog Y at t_2 iff:

1. X at t_1 possesses all the capacities constitutive of being a frog,

2. Y at t_2 also possesses all the capacities constitutive of being a frog

3. Y is physically continuous with X

So it should follow given Locke's definition of persons in terms of the constitutive capacities, which in the case of persons are psychological, that a person X at t_1 persists as a person Y at t_2 iff:

1. X possesses all the capacities constitutive of being a person at t_1

2. Y at t_2 also possesses all those capacities

3. Y is physically continuous (or if X and Y are non-physical, causally continuous in a way equivalent to physical continuity) with X

Locke's rejection of this application of the principle of persistence for organisms to the case of persons led him to the mistaken conception of personal persistence as constituted by continuity of distinctive psychology. My view accepts the application of Locke's principle of persistence for complex objects, in general, to the case of personal persistence. Personal persistence over time consists in the continued existence of a self-consciously concerned being, distinguished by the uninterrupted maintenance of those capacities for self-consciousness and concern, which need not be stable and unchanging in the states they produce, but rather may be, and usually are, dynamic and variable.

VII. The Phenomenal Criterion

There is a recent approach to personal persistence that should be discussed because it has much intuitive appeal. It is a reconsideration of Locke's criterion of continuity of consciousness which appeals to neither memory nor causal relations between distinctive psychological features, but instead to "phenomenal relations between experiences." (Dainton and Bayne 2005, 549) The "phenomenal approach" as its advocates call it, is supposed to account for the anti-Lockean intuitions spurred by the Williams' *pain* frame. The idea is that in the sort of thought experiment where my distinctive psychological features are wiped and replaced and the resulting person is tortured, I have the intuition that it would be me who suffers the torture, not because it would still be my body, but because my phenomenal continuity, the continuity of my

experiential states would be preserved throughout the procedure.

Analogous to the relation between psychological continuity and connectedness as defined by Parfit, phenomenal continuity consists of overlapping chains of phenomenal connections. Phenomenal connectedness is the diachronic version of the unity of conscious experience. In its synchronic form “phenomenal connectedness is simply that relationship of experienced togetherness that holds between all the diverse contents of a state of consciousness at a given time…” (Dainton and Bayne 2005, 554) Diachronically, phenomenal connectedness involves the seamless merging of one experience into another. Overlapping chains of such phenomenal connections form a fluid *stream* of consciousness. According to Dainton and Bayne the connectedness and continuity of experience are conceptually independent of beliefs, desires or memories. On the other hand, they adhere to the inseparability thesis, which is the claim that individual persons are inseparable from the continuity of their experiences. Where the stream of consciousness goes, so goes the person. “Self and phenomenal continuity cannot come apart: all the experiences in a single (non-branching) stream of consciousness are co-personal.” (Dainton and Bayne 2005, 557) So as long as one’s stream of consciousness is unbroken, no psychological or physiological changes will result in one’s failing to persist. The pain sensations felt by the post-op person in the Williams case will be connected to the pre-op person “by an unbroken chain of directly experienced transitions.” (Dainton and Bayne 2005, 557)

The intuition that the continuity of phenomenal consciousness need not be

affected by psychological changes is supported by the view of conscious experience espoused by those philosophers who are impressed by zombification thought experiments (Chalmers 1995/1996) in which one imagine an individual who is stripped of phenomenal consciousness while retaining all of that individual's physical and behavioral characteristics. If such a zombie is conceivable, then we can conceptually divorce conscious experience from other psychological facts. Furthermore, the idea is that a person would not survive the process of zombification, which would make phenomenal consciousness both sufficient and necessary for personal persistence.

The phenomenal approach has similar intuitive appeal to the zombie thought experiment, by exploiting our sense that consciousness is something irreducible to any other phenomena, whereas other psychological features, such as beliefs, desires and memories are so reducible (into e.g. physical, behavioral, or functional states). It seems a zombie could still have beliefs and desires without experience because intentional states admit of, for instance, functional analyses in ways that phenomenal experience cannot. This is not the place to engage with the former claim, but it is enough to say that it is not obviously true. The reducibility of experience is one of the most hotly contested topics in philosophy of mind, so a theory of personal identity that depended on taking a position on this debate would be terribly restricted.

In any case, despite its intuitive appeal there are a good many ways to challenge the phenomenal approach. I will consider a few in what follows. I will argue that a weighing of the various considerations raised by these objections will militate in favor of

my own view rather than the phenomenal approach.

To begin with, it is not clear that the continuity of experience is independent of other psychological features or processes. Dainton and Bayne reject the idea that such continuity can be accounted for in terms of working memory, which is one popular view. (Horwich 1987, 35 and Mellor 1998, 122) For them such accounts are “from a phenomenological perspective… highly unrealistic. Although we can certainly remember experiencing change and persistence, we can also *experience* change and persistence -- and we do so all through our waking hours.” (Dainton and Bayne 2005, 554) However, this comment conflates the various kinds of memory recognized by most philosophers and psychologists. My memory of having experienced continuity of experience is an example of experiential memory. My memory that my experience was continuous is an example of semantic or propositional memory. However, working memory is neither of these things. It is a process that is generally, perhaps necessarily unconscious which allows us attend to our tasks for periods of time. It is not something which is an object of phenomenal experience and so is a perfectly good candidate for what gives rise to such experience. Furthermore, just because memory doesn’t seem to underlie phenomenal continuity doesn’t mean that it in fact doesn’t. Rather, as just suggested, it seems more likely that the basis of phenomenal continuity would be something not available to introspection.

A second objection to the phenomenal approach concerns the assumption that a person’s stream of consciousness is necessarily unified *synchronically*. In a mundane

sense, our attention is often divided as we perform various tasks at once. Still, even when I am listening to music as I write this, there is still a sense in which I have one experience of performing both activities. Either that or I may be switching back and forth between tasks so quickly that it only seems as though I'm engaged in them simultaneously. More problematic are the rare, but real split-brain cases, where the corpus callosum is severed. In such cases, subjects seem to have two streams of consciousness since some information is available to one side of the brain and not the other. (Sperry 1968, Nagel 1971, Puccetti 1973, Moor 1982, Parfit 1987, Bayne 2008). For example:

> What is flashed to the right half of the visual field, or felt unseen by the right hand, can be reported verbally. What is flashed to the left half field or felt by the left hand cannot be reported, though if the word 'hat' is flashed on the left, the left hand will retrieve a hat from a group of concealed objects if the person is told to pick out what he has seen. At the same time he will insist verbally that he saw nothing. (Nagel 1971, 400)

Despite phenomena like that reported above, most philosophers are hesitant to say that there are two persons in such cases. Dainton and Bayne, so long as they agree that there is more than one stream of consciousness, or even less radically, that there is a single, but partially disunified stream, in such cases, would have to say that there is also more than one person.

In a 2008 paper, Bayne argues against the interpretation of the split brain cases that appeals to two simultaneous streams of consciousness as well as the one that posits a partially disunified stream. In their place he proposes an alternative interpretation he calls "the switch model" according to which split brain patients do not

experience two simultaneous streams of consciousness but switch their attention back and forth between streams such that they experience only one stream at a time.

> As the name suggests, the switch model holds that consciousness in the split-brain switches between the patient's two hemispheres. The hemispheres contribute in succession to the contents of the patient's consciousness, but, for the most part at least, consciousness does not occur in both hemispheres simultaneously. The switch model paints the split-brain patient as suffering from a kind of fluctuating perceptual extinction: when the left hemisphere is activated stimuli in the RVF win the competition for entry into consciousness at the expense of LVF stimuli, and the converse happens when the right hemisphere is activated. In general, inter-hemispheric activation will march in step with changes in the subject's attentional focus. (Bayne 2008, 294)

According to Bayne, this model is supported by one aspect of the early split brain findings by Levy et. al. that was hitherto unexplained, namely that one hemisphere never disagreed with a perception of the other.

> ...patients gave one response on the vast majority of competitive trials. Further, the nonresponding hemisphere gave no evidence that it had any perception at all. Thus, if the right hemisphere responded there was no indication, by words or facial expression, that the left hemisphere had any argument with the choice made, and, similarly, if the left hemisphere responded, no behavior on the part of the patient suggested a disagreement by the right hemisphere. (Levy 1990, 235)

The idea, according to the switch model is that the faculty of attention, which results in conscious experience, vacillates between attending to perceptions from the right hemisphere and those from the left, resulting in only one stream of consciousness at a time. Suppose, for example, that the word 'key' is flashed in the left hemisphere's field of vision and the word 'ring' is flashed in the right hemisphere's field of vision, with the left reporting that it saw 'key' and the right reporting that it saw 'ring' but neither reporting that it saw 'key ring', and neither hemisphere protesting the other's testimony

while it's being made. The switch model would explain this event by saying that the hemispheres perceived the information unconsciously and only attended to each hemisphere's perceptions when called upon to use that hemisphere's method of reporting (e.g. speech for the left and pointing or drawing with the left hand for the right.)

This is not an implausible interpretation of what's going on in the split brain cases, although there do seem to be some examples of conflict. For instance, Nagel (1971) reports the following:

> A pipe is placed out of sight in the patient's left hand, and he is then asked to write with his left hand what he was holding. Very laboriously and heavily, the left hand writes the letters P and I. Then suddenly the writing speeds up and becomes lighter, the I is converted to an E, and the word is completed as PENCIL. Evidently the left hemisphere has made a guess based on the appearance of the first two letters, and has interfered, with ipsilateral control. But then the right hemisphere takes over control of the hand again, heavily crosses out the letters ENCIL, and draws a crude picture of a pipe. (Nagel 1971, 400)

Even in this case, however, the conflict involves vacillation of control between the hemispheres, and therefore, is well accommodated by the switch model. Perhaps more troubling is the case Nagel mentions of a man whose "left hand appeared to be somewhat hostile to the patient's wife" (Nagel 1971 401), though perhaps further details would reveal that this example also conforms to the switch model. However, even if the switch model is the right explanation of the split-brain, Dainton and Bayne's view does not by itself save the phenomenal approach to persistence. One would have to give an account of how when one switches from one set of perceptions to the other and back, the two sets are continuous with one another, such that a single person persists through the switching. If each hemisphere can be thought of as a separate experience producer,

then there would be two persons in a single human being. Philosophers have been reluctant to embrace such a consequence with good reason. Split brain patients generally function as single, unified individuals in their everyday lives, and the inconsistencies in their behavior are usually only observable under experimental conditions. This fits better with the idea that what unifies them is not the continuity of their experiences but of their higher-order capacities for self-consciousness and concern, both of which seem undivided by callosotomy.

The most obvious objection to the phenomenal approach is that experience is not, generally continuous. We normally think of persons as persisting through regular gaps in phenomenal continuity, in dreamless sleep or, more dramatically, blacked out or fugue states. Overlapping chains of phenomenal connections, whatever those are supposed to be, do not account for these discontinuities. Dainton and Bayne recognize this problem of how interrupted streams of consciousness can be continuous, calling it the "bridge problem," and offer three options in reply that I are interesting enough to consider. The first invokes James' idea that there is a particular qualitative feel to one's experience, such that "even when there is a time-gap the consciousness after it feels as if it belonged together with the consciousness before it, as another part of the same self." (1952, 154) Dainton and Bayne dismiss this suggestion as dubious, though it has some intuitive appeal. However, if there is a special me-ness to my experiences I'm hard pressed to put my finger on what it is or provide evidence for its existence, and even if I could, what rules out the possibility that someone else could have a me-ish

experience - an experience which feels like one of mine? Perhaps I could misremember whether or not an experience was me-ish or q-remember the quality of someone else's experience, without remembering.

The second option is to treat as continuous streams of consciousness that are not strictly so, as long as they *could* have been continuous given their qualitative similarity. Dainton and Bayne rightly object that the character of one's experiences before and after interruptions is generally quite different, and actually it is this very divergence that is often evidence of the fact of interruption. For example, say I am crossing the street, thinking about this problem. The next thing I experience is being in a hospital in extreme pain. It is the fact that my current experience is so different from the immediately previous one that I infer an interruption in my stream of consciousness. Furthermore, as Dainton and Bayne also point out, if experiential subjects need not be strictly continuous, then they would enjoy intermittent existence. Such an account would then require abandoning IIE.

Dainton and Bayne's most promising solution to the bridge problem appeals to the notion of a capacity in much the same way that Unger's view and my own both do. "When a person becomes unconscious, none of their experiential capacities are active, but the capacities nonetheless remain in existence: the irretrievable loss of the capacity for consciousness is what differentiates being merely unconscious from being dead." (Dainton and Bayne 2005, 565) So it seems phenomenal continuity does not require actual connections between experiential states at all, but only the maintenance of the

capacity for such states by an “experience producer” (EP). Even if an EP is not continuously producing experiences, “we can appeal to experiences they would produce if they were active. EPs that would produce phenomenally connected experiences if they were active should also be regarded as co-personal. Since phenomenal connectedness, actual or potential, obtains diachronically as well as synchronically, we have all the ingredients we need to solve the bridge problem.” (Ibid) So a person, as EP, persists after the wipe and replacement of his or her psychological features, so long as he or she maintains the capacity to produce experiences throughout the procedure, even if he or she is completely unconscious during it. In this way, their view becomes nearly indistinguishable from my own, except that the only capacity they regard as essential is the capacity for a unified stream of consciousness. However, while such a capacity might be sufficient for a conscious being to persist, it is not sufficient for a person to do so, because while perhaps necessary (though perhaps not -- see chapter four), it is not sufficient for something to be a person that it has unified phenomenal consciousness. This doesn’t seem to bother Dainton and Bayne as they “incline to the view that *no* cognitive sophistication is necessary for our survival, and that we could survive with a consciousness of the simplest of forms, e.g., a few basic bodily feelings (it is arguable that we all enjoyed a consciousness of this form prior to birth.)” (Dainton and Bayne 2005, 561) For them we are essentially “phenomenal things.” (566) If one accepts both my account of personhood and their account of persistence one is forced into the position that we are not essentially persons. This may or may not be a terrible

result. Was I a fetus who became a person? If not, then I must suppose that I only came into being when my capacities for self-consciousness and concern were fully developed, in which case I might have to say that I was never an infant, which seems absurd. On the other hand, if I were to lose my capacities for self-consciousness and concern, and cease to be a person, would I still exist, provided I still had the capacity for experience? Could I gradually turn into a frog and still exist? Settling that question is beyond the scope of this study. What we are looking for is an account of *personal* persistence, such that responsibility is maintained over time. Phenomenal continuity, even if it ensures *my* persistence, does not ensure that I persist as a being responsible for my past actions.

VII. Summing up the Core Psychological Criterion and objections

The view developed so far in this chapter can be summarized as follows:

a person X at t_1 persists as a person Y at t_2 iff

1. X possesses all the capacities constitutive of being a person (self-consciousness and concern) at t_1,

2. Y at t_2 also possesses all those capacities

3. Y is physically continuous (or if X and Y are non-physical, causally continuous in a way equivalent to physical continuity) with A

Now consider the following case[21]:

Mar-Vell dies of cancer. His wife is a brilliant geneticist and figures out how to

[21] Adapted from Reed (2008)

alter the body of another person Kl'rt, on the genetic level in such a way that Kl'rt's body comes to exactly resemble Mar-Vell's body just before he died, minus the cancer, but with a brain structured in such a way that he is psychologically exactly similar to Mar-Vell before his death. Before the procedure, Kl'rt was Mar-Vell's arch-nemesis, with diametrically opposed moral beliefs and values. The person who wakes up after the procedure at first believes he is Mar-Vell, as does anyone he encounters, however upon discovering Mar-Vell's wife's notes, he comes to believe that he is not Mar-Vell but an imposter.

i. Objection 1 - mechanical reproduction

My account of persistence supports the intuition that this is not Mar-Vell we're dealing with, that it is still Kl'rt and he is right to think of himself as a fake as far as being Mar-Vell is concerned. However one might worry that we are being misled by the biological natures of Mar-Vell and Kl'rt. What if Mar-Vell were a mechanical person and didn't die of cancer but of corrosion of his mechanical organs. If Kl'rt's body was nearly exactly similar to Mar-Vell's but had been encoded with a different psychology which is now altered to exactly resemble the psychology of Mar-Vell before he died, would we still be inclined to think of the resulting persons as a Mar-Vell imposter? What significant difference would there be between this individual and Mar-Vell had his mechanical organs not become corroded?

My answer to this objection is that it's not that we have been misled by the biological nature of the individuals in the first case, but by the mechanical nature of

those in the second. We don't think of mechanical objects as surviving or perishing in the same way that we do organic ones, because they don't. Furthermore, mechanical objects, unlike organic ones can generally persist despite being completely taken apart so long as their parts remain in working order and do not become parts of something else. Also, they tend to be mass produced in a way that makes them generally interchangeable. However, if I buy two watches that are qualitatively identical except that one is black and the other is white, the black one is destroyed beyond repair (it no longer realizes the criterial capacities of a watch) and I paint the white one black, this will not constitute the resurrection of the original black one, despite its now being exactly similar to it. I think we should say the same thing of mechanical persons. Mechanical Kl'rt has been made to be just like Mar-Vell was, but Mar-Vell has not persisted in this new body, because he did not persist at all, he was destroyed.

Furthermore, while up to this point I have assumed for the sake of argument that it is possible to transfer one's distinctive psychological features from one body to another without an actual transfer of brain matter, there is reason to think this assumption faulty. The contrasting intuitions elicited by the version of Williams' differently framed thought experiments that I have considered depend on this assumption. The Lockean frame suggests that such transfer would constitute body-switching and the pain frame suggests that it would not. However, if this assumption is faulty, as I think it is, the conflict in intuitions cannot even get off the ground. The assumption seems to involve a conflation of type and token as regards psychological

features. In the watch example, after painting the white watch black, I might say that the new watch is the same color as the old one. However, I cannot mean that it has the very same *token* of black color as the old one, but only the same color *type*. Similarly, were I to produce in someone else's brain a belief in the nature of personhood qualitatively exactly similar to the belief that I currently possess, it would not be the case that this person now possesses the same belief token that I possess, but only a belief of the same type. This would be the case even if first annihilated my own instantiation of the belief. It is manifestly impossible for me to transfer the tokens of my psychological attributes to another person. That is, unless I transferred the physical realizer of those psychological tokens, i.e. my brain. Therefore the dichotomy between psychological and physical approaches to identity is a false one.

Furthermore, even if it were possible to transfer tokens of psychological attributes from one individual to another, if one wants to claim that the identity of the person who originally had the attributes would be assumed by the second individual, given transfer of a sufficient number of attributes, Butler's criticism of the memory criterion can be extended to the expanded psychological criterion which would require that the attributes be q-attributes, such as q-beliefs and q-desires. However, the problems Schechtman raises for the account of identity in terms of q-memory, discussed above, apply just as well to q-beliefs and q-desires, because "the content of these psychological states, too, cannot be defined without presuppositions about who it is who has them." (Schechtman 1990, 86) For example, even if I have a desire to win at tennis, at which I am an expert

(I'm not really, but it's Schechtman's example), and my friend who is a novice comes to have a similar desire with the same surface content, i.e. that 'I' win the game, "her experience will be different from mine because it will not include the anticipation of the play of the game, the feeling that she, the desirer, had been beaten by my opponent too many times, or the anticipation of standing in front of the spectators exalted if I win and humiliated if I lose." (Schechtman 1990, 86) Schechtman's diagnosis of the confusion generated by the "psychological approach" of Parfit and others, is that features of a first-personal nature, such as memories, beliefs, and desires, are appealed to to give an account of persistence that is similar to accounts of the persistence of material objects taken from a third-person perspective.

> Taking the fact that psychology is what turns out to be important… psychological-continuity theorists thus make the unwarranted assumption that sameness of psychology can be used to provide a noncircular criterion of identity of the sort which is given for objects. But such a criterion cannot focus on subjectivity; it is, by definition, to be objective, and must be capable of being spelled out without including the first-person perspective of a given individual. The pieces that make up a person's psychology, must, to fulfil this purpose, be viewed to be as discrete and detachable as the planks of a ship or the grains of sand in a heap… atomic, isolable, and in principle independent of the subject who experiences them -- a view that I have argued is highly implausible. (Schechtman 1990, 89)

Schechtman's solution (in the 1990 paper) is to abandon the project of establishing objective conditions in favor of focusing on a person's own self-formed narrative as constituting their persistence over time. I have, following Unger, made the opposite move, by grounding personal persistence in the objective conditions for the maintenance of a person's psychological capacities. I do, however, think that something like Schechtman's view is the right account of the formation of a person's *self*, which I

take to be distinct from the persistence of the person, and which will be the topic of chapter four.

ii. Objection 2 - change of type = change of token

A second potential problem for my view which comes out of the Kl'rt/Mar-vell story involves the question of, assuming that Kl'rt has not become Mar-vell, whether or not it is still Kl'rt we are dealing with after the transformation. According to my view so long as the core psychological capacities of Kl'rt have been maintained throughout metamorphosis, the same person, Kl'rt, should remain at the end of the process. However, in keeping with the mechanical analogy above, there may be times when we want to say that a certain object, a car for instance, cannot remain the numerically the same object, if it transforms so radically as to become a token of a radically different type than it previously was. For example if a VW bug gradually had its parts replaced until there was no longer a bug in the garage, but a porsche, it would be strange to say the new car is numerically identical to the original. Or to reverse the example, if I were to lend someone my porsche and they were to return to me a bug, I would insist that the returned car is not the same one that I lent them. Similarly, one's intuition might be that the transformation makes Kl'rt into such a radically different type of person that it cannot be one and the same person we are dealing with before and after metamorphosis. My response to this objection is that there is no clear way to distinguish types of person in the same way as we can distinguish types of car, for the very reason that I rejected continuity of distinctive psychology as necessary for personal persistence in the earlier

section. Persons do not always maintain a steady enough set of characteristics. Their beliefs and desires change gradually, and even vacillate radically from one moment to the next. Furthermore, even when there is constancy of character, the Myers-Briggs personality inventory notwithstanding, persons do not divide neatly into types the way automobiles do.

A more difficult analogy to contend with would be works of art, which are like persons in their uniqueness. If I painted my own face over an Albrecht Dürer self-portrait, it might remain a painting, even the same type of painting (a self-portrait), but would clearly be a numerically different painting from Dürer's. By analogy, one might think that a sufficient change in distinctive psychology would make one person into a numerically different person, even if it was a person all along. Where I think the analogy fails, is that a necessary condition (though perhaps not sufficient) for a work of art remaining the same work, is that it has been wrought by the same artist or group of artists. Once an alien hand has affected it, it loses its numerical identity. Persons, however, have no clear set of authors, and therefore cannot have that condition of their persistence. With paintings, the original artist may go back to a work and alter it, thereby changing what remains numerically the same work. If they were to completely cover the painting with gesso and begin painting anew, there would not be a continuous work -- the original would be destroyed. This would be analogous to killing a human being and then reanimating the body with a new brain. If, however, Dürer were to paint my face over his, one would be inclined to say that he had created a new painting, but I think

that might be a mistake. Why consider it a new painting and not a revised one (though not for the better)? Perhaps the intention of the author can by fiat change the identity of a work, but there is no such analogue in the case of persons, because persons have no author distinct from the “work” itself.

A related issue is whether or not an individual who ceases to be a person at all can or must remain the same individual or whether something that becomes a person was the same individual before becoming so that it is after having begun to instantiate personhood. For instance, infants, lacking fully developed self-consciousness, are not yet persons. Was I, who am currently a person, the infant who issued, second after my sister, from my mother’s womb? This issue involves difficult questions about the nature of individuals and whether or not there is an absolute, identity-simpliciter relation or only category specific ones, i.e. identical-person, identical-infant, etc. (Geach 1962, Perry 1970) that are beyond the scope of this study, because my interest here is only with the identity of persons, insofar as that identity accounts for continued responsibility. An individual who ceases to be a person cannot be responsible for anything, whether or not that particular individual survives as a non-person.

IX. Persistence and identity

It is now time to give some scrutiny to the working assumptions made at the beginning of this chapter. The first was that, following Parfit, the relation of identity is merely a special case of persistence or survival. Making this assumption allowed me to

put to the side worries about branching scenarios which make the notion of identity problematic. If it is possible for me to “split like an amoeba” into two individuals who are equally continuous with me in all the relevant psychological and physical respects, then I might worry about which one of them will be me. Parfit’s answer is that neither of them will be, but both will bear the same relation to me as if there had only been one contender. The second assumption was that that we need not take identity in a strict, but rather only a loose sense, because strict identity would require exact mereological and qualitative similarity over time. Loose identity, would only require a kind of continuity between somewhat dissimilar instantaneous entities.

If one is uncomfortable with either assumption, there are metaphysical resources available to bring such concerns in line with the view of persistence so far developed. These resources come from discussions about the metaphysics of time. One metaphysical conception of time, 4-dimensionalism, which draws some support from Einsteinian physics treats time as a dimension akin to the three dimensions of space. Since objects have parts distributed throughout different points in space, on this conception they would also have parts, temporal parts, distributed over different points in time. This is often invoked to defend the identity of persons and other transitory objects from the charge that since identity implies Leibniz’s law (if x=y then x and y have all their properties in common) something cannot change over time while remaining the same thing. If as an 8 year old I am 4 feet tall and 6 feet tall as an 18 year old, then I seem to possess contrary properties. Which one am I: 4 feet tall or 6? The temporal

parts theorist replies that I am strictly neither and that my possessing different properties at different times yields no contradiction. This is because the whole of me does not possess any of these properties, but only temporal parts of me do. The part of me at 1992 is 4 feet tall and the part of me at 2002 is 6 feet tall. What I am is the 4-dimensional space-time worm composed of these distinct parts. I do not persist by *enduring*, continuously existing as a complete person from one moment to the next, but by *perduring*, having temporal parts that are related to one another in a particular way (e.g. constituting the kinds of processes that maintain the capacities for self-consciousness and concern).

If one is committed to the idea that persisting individuals must be strictly identical over time in some way, one can say that while the relation between person-stages at different times is only a loose kind of identity, strict identity would apply to 4-dimensional objects taken as wholes. Furthermore, if one is uncomfortable with Parfit's view that in instances of branching, a person persists as two distinct persons, neither of which are identical to the original, then one can appeal to pairs of 4-D space-time worms that share some temporal parts in common, having two heads or two tails and say that there are two distinct persons who share some temporal parts in common or only one person who has some temporal parts that are spatially discontinuous.[22] This is not the place to debate the merits of this appeal to the identity of 4-D worms taken as wholes or the theory of temporal parts in general, and I think many of the disagreements involved are

[22] See Moyer (2008) for a thorough discussion of these ideas

largely verbal. However, if one is uncomfortable with abandoning strict identity, and settling on the looser sense of the term or with identity as a merely special case of persistence, then some comfort may be achieved by appealing to temporal parts. However, the discomfort seems to me, in the first place, to be the remnant of attachment to the idea of an enduring, substantial self, which should be eschewed.

Now, while 4-dimensionalism is compatible both with my view and DPN, if one is already a proponent of 4-dimensionalism, then one has a further reason for accepting my view over the alternative. This is because of the difficulty for 4-d theorists of providing a criterion for distinguishing between space-time worms, which can be overcome if one accepts my view of the persistence of persons and other complex, organized beings. What makes one space-time region a part of one worm and not another? On my view, we can answer this question in terms of which physical processes co-occurring in space over time yield a being with certain capacities. The physical parts which serially make up the processes that ground the maintenance of the capacities of an individual, in the case of persons these are self-consciousness and concern, occupy the space-time regions of which an individual worm consists.

An alternative to 4-dimensionalism for trying to resolve the paradoxes of identity discussed above would be to deny some of the logical principles traditionally thought to be built into the notion of identity, e.g. substitutivity and transitivity. Priest (2014) does just that, arguing that such a move dissolves problems of fission, fusion, colocation, and vagueness among others. For the purposes of this study I have chosen to leave out

discussion of these problems concerning the notion of identity (with the exception of the issue concerning overdetermination discussed in chapter three which I understand to be a special challenge to the position of reductionism about persons), so as to focus on the issues that seem to make the logic of persons depart from that of other kinds of objects, though I grant that any problem concerning the identity of things in general will also be a problem for the identity of persons.

Chapter 3: The Ontology of Persons

I. Reductionism and Persons

Some form of Reductionism, either of theories, facts, or entities, is for many philosophers, a crucial tenet of naturalism. In metaphysics, complex objects are believed to be reducible to their parts and relations among those parts, which may then be further reduced, so that they bottom out at a fundamental level of simples or else are infinitely reducible to infinitely lower levels. In philosophy of science, the facts of biology are taken by Reductionists to be reducible to those of chemistry and those of chemistry to those of physics. Reducibility in metaphysics and science is desirable because it allows that the world has a kind of hierarchical unity which would make possible a complete understanding of the inner workings of all things. If the assumptions of Reductionism are correct and one wishes to include persons in a naturalistic worldview, then it seems that facts about persons must also be reducible to facts on a lower level of description. This is not the place to evaluate those general Reductionist assumptions, but even without them, Reductionism about persons may possess various virtues of its own.

Reductionism about persons, in general, holds that persons are nothing over and above their psychophysical components and the relations between them. These components may be mental states such as beliefs and desires; physical particles such as molecules, atoms or electrons; or more occult elements such as Nietzschean drives. The phrase “nothing over and above” is a bit vague as it stands, and intentionally so, for

it will mean something slightly different depending on how stringent a Reductionism one endorses. The particular sort of Reductionism I am arguing for, and hence the more precise meaning of "nothing over and above" I take to be true of persons, will become clear in what follows. In any case, I understand Reductionism to be a moderate position between the two extremes of Inflationism (or Non-Reductionism) and Eliminativism. Inflationists believe that persons are irreducible, either because they are, or contain as essential parts, irreducible entities that are distinct from and independent of their psycho-physical components; or because there are facts about them that for either metaphysical or merely linguistic reasons cannot be reduced to concatenations of facts about their components or relations.

Eliminativists, however, either agree that persons are reducible to their psychophysical components, but hold that, in general, fully reducible composite objects do not really exist, or else, in the spirit of Churchland's (1975) Eliminativism about folk psychological concepts, think that the concept of a person necessarily refers to something with irreducible properties, so that if all the things we are tempted to call persons turn out to be fully reducible, there are no genuine persons after all. This conclusion is what is known as the "Extreme Claim" (Parfit 1984, Siderits 2003, Schechtman 2014) with respect to persons.

The "Extreme Claim" which, as Parfit (1987) puts it, is the claim that "we have *no* reason to be concerned about our own futures" (Parfit 1987, 307) or as Siderits (2002) states it, more expansively, "that four central features of our present person-regarding

practices cannot be rationally justified: interest in one's own survival, egoistic concern for one's future states, holding persons responsible for their past deeds, and compensation for one's past burdens." (Siderits 2002, 37) Siderits' version is of greater interest in the context of the present study, because Parfit's version, as stated[23], would not by itself pose a serious challenge to the concept of a person developed herein. While concern, in the sense of emotional investment in the satisfaction of one's desires and truth of one's beliefs, is necessary for responsibility and therefore personhood, it is not necessary that one's desires target one's own future conditions or states of affairs. I may be concerned only for the happiness or good fortune of others when I act responsibly. I may anticipate reward or punishment in the form of benefit or harm to my loved ones that could come at a time after I have ceased to exist. Only a particularly radical psychological egoist would hold that concern is necessarily concern for one's own benefit or harm, because such a position assumes that such self-concern is the only ultimate concern that is possible to have. The part of Siderits' formulation of the Extreme Claim that is threatening to the conception of personhood developed in the present study, is the idea that holding people responsible for past deeds or compensating them for past burdens is irrational for a Reductionist.

In the interest of showing that the conception of persons developed in the earlier chapters of this study is consistent with the kind of naturalistic Reductionism described above, I will endeavor to defuse some of the arguments for both Inflationism and

[23] Parfit does seem to think that the Extreme Claim is relevant to responsibility as well, but does not mention it in that formulation.

Eliminativism. The specific form of Reductionism that I endorse may not be sufficiently strong to satisfy all Reductionists, particularly not those who insist on what Parfit (who is one of them) calls the "Impersonal Description (ID) Thesis," which holds that all the facts about persons can be otherwise stated, without remainder, in an impersonal language, one that does not refer to persons or their identity. I take the ID thesis to be untenable, but for reasons that do not threaten the general outlook of naturalistic Reductionism that I am endorsing.'Person', on my view, is not just a convention, a merely convenient designator that could be eschewed at the sole costs of time and verbosity, but is truly required for a complete description of the world.

On my view persons are complex, organized, composite objects that have features, i.e. the capacities for self-consciousness and concern, which distinguish them from other such objects. Being the sorts of objects that have such features is due to the organization of their constituents and nothing more, so that persons are reducible to the organization of those constituents. Therefore, if arguments for the elimination from strict ontology of composite objects, such as baseballs, tables, chariots, etc. on the basis of those purported objects' reducibility to their constituents succeed, then they must also succeed in demonstrating that persons should be so eliminated.[24] Therefore, the first part of this chapter will be devoted to defusing the arguments for the elimination of

[24] Most Eliminativists make an exception for living beings when arguing for the otherwise wholesale elimination of composite objects. Furthermore, complex, though abstract objects such as Clubs and nations may be exempt from elimination. However at this stage, such exemptions are unnecessary because the argument against composite objects isn't sound in the first place. I will, however, discuss those reasons in the context of arguments against the ID thesis as well as those in favor of Inflationism.

composite objects in general. Those arguments generally appeal to the idea that if composite objects exist then they are causally redundant given that the events they are supposed to cause can be explained entirely as the effect of the composite objects' components, and so if both constituents and composites exist, then events caused are overdetermined. Since events are not overdetermined, composite objects must not really exist. My rebuttal to this line of argument will appeal to the idea that persons and other composite objects are contingently identical to the organized bodies which can survive gradual changes in their components so long as the organization of components is maintained, and so are not identical to any specific group of components themselves, and therefore do not compete with their components for causal relevance. This part of the chapter will have the added benefit of situating my account of persons within the general contemporary metaphysical landscape of discourse.

The second part of the chapter will be devoted to assessing the ID thesis and its relation to naturalistic Reductionism. Some of the Eliminativist arguments take the ID thesis to be the part of Reductionism that entails the Extreme Claim. Parfit sees the ID thesis as essential to the Reductionist position, but thinks it is compatible with our forensic judgments about persons. I will argue that the ID thesis is not essential to Reductionism after all, for the considerations against it do not conflict with the general naturalistic worldview that otherwise favors Reductionism. A Reductionist need not endorse the ID thesis and therefore need not worry about its entailing the Extreme Claim. Like some other composites, such as clubs and nations, there are facts that

cannot be stated without using the word ‘person’ but that is not because persons are something distinct from or independent of their constituents and the relations among those constituents. They’re distinct from mere collections of constituents, but not from constituents so related or organized as to yield a being with the capacities for self-consciousness and concern. The nature of the constituents and their relations may only be describable given the assumption that they are constituents of a person, therefore some reference to the person as *owner* of the constituents and *author* of some of her actions may be required, but that doesn’t make the person something over and above the constituents in a sense that would conflict with naturalistic Reductionism.

In the third section I will consider reasons that Inflationists have for thinking persons irreducible. The reasons depend on the assumption that persons are or require for their existence irreducible souls, or else have irreducible properties and powers, namely indeterministic free will or “top down” causal powers. It is no surprise, given my endorsement of Reductionism, that I reject the Inflationist’s conception of persons as having such irreducible features. However, some Eliminativists have argued, on the grounds of such rejection, that Reductionism necessarily slides into Eliminativism, because if persons are beings capable of being responsible for their actions, then they must have the features that the Inflationists insist they have. If no being has such features, the Eliminativist argues, then no being is capable of responsibility and therefore, no being is a person, i.e. that the Extreme Claim is true. I will endeavor to demonstrate that contrary to the argument just sketched, Reductionism about persons

does not entail the Extreme Claim, and therefore, does not slide into Eliminativism. Persons either need not have all the sorts of features Inflationists require of them, and the features they do require, namely responsibility, are compatible with their reducibility.

Schechtman (2014) understands the Extreme Claim in terms of a difference of opinion between Locke's view and later Reductionist neo-Lockean psychological continuity theories such as Parfit's concerning whether or not a person can be a 'forensic unit,' "a kind of entity that can sometimes be rightly rewarded or punished for its actions."[25] (Schechtman 2014, 15) According to Schechtman, Reductionist views, such as Parfit's depart from Locke's view in denying that a person is a forensic unit, though they still want to hold that judgments of responsibility can be rationally made of persons reductively construed. She claims, and reads Locke as suggesting, that

> individual judgments about responsibility and like concerns depend upon the existence of a more basic forensic unit for their legitimacy. Reductionist psychological theories do away with any kind of meaningful forensic unit, and so cannot provide that legitimacy. Relations that would justify the ascription of moral responsibility if they held within a forensic unit are not by themselves enough for such an ascription if the existence of such a unit is not presupposed. (Schechtman 2014, 35)

Parfit sometimes seems to think judgments of responsibility do not require forensic units, i.e. judgements of responsibility can be made about individual actions or mental states in the absence of there being someone who performs those actions or possesses those states. However, a Reductionist need not reject the notion of persons as forensic units. Units in general admit of analysis into smaller units and their relations.

[25] Schechtman's distinction between person as forensic unit and person as moral self will be discussed in the next chapter.

The kinds of relations Schechtman is concerned cannot ground responsibility without being part of or 'within' a forensic unit are the ones usually emphasized by Reductionist accounts of persons, i.e. memories, beliefs, desires, and values, but the point may apply just as well to my view, which takes the maintenance of the capacities for self-consciousness and concern to be necessary and sufficient for responsibility. Those capacities are capacities of persons, beings that persist over time, and who therefore can be responsible for actions performed in their pasts.

While I agree that a forensic unit is required to make judgments of responsibility, I don't think that the existence of such an object is incompatible with Reductionism. Insistence on such incompatibilism requires that Reductionism entail the Extreme Claim, or otherwise collapse into Eliminativism concerning persons as forensic units. I will argue that Reductionism does not in fact entail the Extreme Claim nor need it collapse into Eliminativism for any other reason. I hold that there is a conception of responsibility that is compatible with Reductionism. This conception of responsibility does not support all of our pre-reflective attitudes about praise, blame, revenge, and punishment, but it does account for and justify our sense of ourselves as purposive agents who, at least sometimes, act because of reasons that we are able to reflect on, and therefore can be responsible for some of those actions.

II. Eliminativism and Composite Objects

One reason why someone might think that Reductionism rules out the existence of persons as forensic units is prior commitment to a general Eliminativism about composite objects, i.e. objects that are not indivisible simples, but are composed of parts (which may or may not themselves be indivisible simples.) If one holds such a commitment, then it would follow that there are no persons that are "units," forensic or not, unless persons are indivisible simples or are not fully analyzable into relatively simpler components, which would go against the claims of Reductionism.

Merricks (2001) defends the general Eliminativist stance regarding composite objects. His main argument in favor of it can be reconstructed as follows: If baseballs, for example, are composed of particles-arranged-baseball-wise, then every baseball is co-located with the simple particles which, so arranged, compose it. But then, when some event one would normally call 'a baseball breaking a window,' occurs, that event would be causally overdetermined by, on the one hand, the baseball, and on the other, the particles so arranged. However, events cannot be causally overdetermined. Therefore, if the particles-arranged-baseball-wise broke the window, then the baseball could not *also* have broken it. Merricks calls the argument for this last claim, the "Overdetermination Argument," which he states as follows:

> The baseball--if it exists--is causally irrelevant to whether its constituent atoms, acting in concert, cause the shattering of the window. (2) The shattering of the window is caused by those atoms, acting in concert. (3) The shattering of the window is not overdetermined. Therefore, (4) If the baseball exists, it does not cause the shattering of the window. (Merricks 2001, 56)

Generalized, this argument is meant to show that baseballs do not cause any events at all, because every event they are purported to cause can be wholly causally accounted for by the action of particles-arranged-baseball-wise. And since "every macrophysical object causes something," Merricks (2001, 82), concludes, baseballs must not really exist. Baseballs are "causally irrelevant" to the breaking of the window or anything else, and only causally relevant objects truly exist. To speak of baseballs as existing alongside the particles that compose them would be metaphysically redundant. The term 'baseball' or 'apple' might be useful as a shorthand for 'particles-arranged-baseball-wise' or 'particles-arranged-apple-wise,' respectively, but strictly speaking, or "in the philosophy room," as another Eliminativist about most composite objects, Van Inwagen, puts it (borrowing the phrase from David Lewis), "'There are apples'... may well express a proposition whose falsity is consistent with the truth of the proposition expressed by typical utterances of 'There are apples on the sideboard if you want one,'" (Van Inwagen 1993, 178) where that latter sentence is shorthand for the proposition that in the philosophy room one would most accurately express as "There are particles arranged apple-wise on the particles arranged sideboard-wise."

This line of reasoning seems absurd, especially if one is familiar with Gilbert Ryle's (1949) notion of the 'category mistake.' One makes a category mistake, according to Ryle, when one takes objects of two different 'logical categories' and treats them as if they are of the same logical category by counting them as separate items in a group or on a list. For example, one commits a category mistake when one counts a left

hand glove, a right hand glove, and the pair comprised by the two gloves as 3 distinct objects. A pair of gloves, after all, *is* just a left hand glove and a right hand glove. Similarly, it seems intuitively obvious that a baseball *is* just particles arranged baseball-wise. One may invoke the relation of identity to explain away the seeming redundancy: a baseball is identical to a bunch of particles arranged baseball-wise, and therefore whatever the particles do, the baseball does, but these are not distinct doings, because there aren't distinct things doing them. This would be akin to the move made by Kim (1998), who, when discussing the mind-body problem, first invoked worries about causal redundancy, but thought such concerns could be assuaged if one were to embrace the Reductive Materialist claim of mind-body identity.

However, the Eliminativist about composite objects rejects such identity claims on the grounds that supposed baseballs would have different persistence conditions from the particles that are arranged baseball-wise. For instance, baseballs, if they exist, constantly lose small portions of their particles, but in such a way that the baseball's persistence and window-breaking powers appear unaffected. Furthermore, a baseball, if one were to exist, could be destroyed without destroying the particles that, when suitably arranged, composed it. Therefore, baseballs cannot be identical to any particular bunch of particles, because they do not have the same persistence conditions.

Before I explain what's wrong with the above line of reasoning, I should point out that neither Merricks, nor Van Inwagen, actually extend their Eliminativist arguments to

the case of persons (in Van Inwagen's case, insofar as persons are organisms they survive the culling). Rather, they offer reasons for excluding persons, along with other complex organisms from the wholesale elimination of composite objects, thereby embracing them in their 'sparse ontology'.[26] For Merricks, persons evade the Overdetermination Argument because they, being identical with human organisms (at least paradigmatically), have causal properties that cannot be fully analyzed in terms of the causal properties of their parts. "For material objects," he writes,

> *to be is to have non-redundant causal powers*…. Human organisms have non-redundant causal powers or exercise downward causal control over their parts. This deep, fundamental difference between the powers of human organisms and the powers of alleged baseballs (and statues and rocks and so on) makes all the difference with respect to the Overdetermination Argument. (Merricks 2001, 115-16)

I will not go into the details of the argument for the claim that humans have non-redundant causal powers at this point, because to assert the claim in question is to abandon Reductionism about persons, and my purpose here is to explore the implications of endorsing Reductionism. An Inflationist might claim that in order to be suitable targets for judgments of responsibility, persons must be capable of exerting such "downward causal control" over their parts in a way that is inconsistent with Reductionism. I will argue against this claim later on. For one thing, other sorts of things such as clubs, nations, and even some machines of human invention can exert a kind of downward causal control. A club can dismiss a member, a nation can send some of its

[26] A term employed (though not invented) by Schechtman 2014 (176) to describe the ontology of Van Inwagen and also Eric Olson.

people to war, a computer can turn itself off after a period of disuse. However, there does not seem to be any good reason to think that the actions of these things are not explicable in terms of the actions of some of their constituents (majority vote, executive branch, energy saving protocol, respectively). So Merricks must have something else in mind when he talks of persons and animals, likely something which a naturalist, and therefore Reductionist, about persons should claim is not required for personhood.

However it takes no appeal to special causal powers to defend ordinary, non-living, composite objects from the Overdetermination Argument. One strategy for doing so has been developed by Thomasson (2007), in the spirit of Ryle, which provides a clearer definition of what counts as a logical category by explaining that objects of different logical categories bear relations of 'analytic entailment' to one another. She uses the expression 'analytically entail'

> to mean 'entail in virtue of the meanings of the expressions involved and rules of inference', so that a sentence (or set of sentences) φ analytically entails a sentence Ψ just in case, given only logical principles and the meanings of the terms involved, the truth of φ guarantees the truth of Ψ. Thus where φ analytically entails Ψ, given knowledge of the truth of φ, as well as grasp of the meanings of the terms and reasoning abilities, a competent speaker may legitimately infer the truth of Ψ on that basis alone. (Thomasson 2007, 16)

Thomasson then employs this analysis to explain away problems of causal and ontological overdetermination. According to her, if claims about particles arranged baseball-wise causing windows to shatter analytically entail statements about baseballs causing windows to shatter, it "does not require more of the world" or "any *extra* causal action" (Thomasson 2007, 16) to make both of the two statements true than it does to

make either one true individually. Therefore, there is no overdetermination, “no doubling or competition between the claims” (Thomasson 2007, 16). Claims of existence, on her view, “are to be resolved by determining whether the applications [for the sort of thing in question] are fulfilled, and that conditions for those ordinary terms are established by ordinary, competent speakers.” (Thomasson 2011, 157) In the case of baseballs, the application conditions are that particles have been assembled by an artisan according to the official standards of professional baseball in such a shape that they are collectively capable of being thrown, hit, and caught in the ways required for playing a game of baseball. That is just what it is, analytically, to be a baseball and “if the serious ontologist disregards the application conditions standardly accepted by competent speakers in favor of higher metaphysical conditions, then her denial that these conditions are met tells us nothing about whether or not there are any [baseballs], for if she shifts the application conditions she shifts the terms of discourse and is not denying the existence of our familiar [baseballs].” (Thomasson 2007, 157) A ‘baseball’ that has causal powers beyond those of the particles that compose it, would not be a baseball at all, but some kind of super-baseball. To borrow a term from Paul Edwards (1949), demanding that an object meets such a condition in order to exist is to “highly redefine” the term ‘object,’ to add necessary conditions to its application that are not part of the ordinary meaning of the term.

However, this strategy for rejecting Eliminativism carries with it the cumbersome baggage associated with the notion of analyticity and so may be objectionable to many

philosophers influenced by Quine (1951). Furthermore, as stressed by Bennett (2009), Thomasson's view makes the identification of the components and object composed too strong, i.e. necessary, whereas one might wish to allow for possible worlds where particles arranged baseball-wise would not yield baseballs. Finally, her view does not adequately address the problem of identifying a composite object with a collection of constituents that changes over time.

The strategy for countering the Overdetermination Argument which I prefer requires claiming only contingent identities between composite objects and the arrangements of objects that compose them. In arguing for such an identity, one need only show that the Eliminativist's reasons for rejecting the identity between, e.g. baseballs and particles arranged baseball-wise are unfounded. To begin with, notice the second reason above for rejecting the identity of composite objects with their parts. There the problem was that all the same parts could remain in existence, though the composite be destroyed. However, this possibility betrays the fact that it is not the parts themselves that matter for the identity of the object, but rather the way they are organized or arranged. As discussed in chapter 2, Locke's account of the persistence of organized, composite objects over time does not appeal sameness of parts, but rather their organization or arrangement. Similarly, baseballs are not identical to the *particles* which are arranged in such and such a way, but the *arrangement* of particles, which may include some particles at some times and different ones at others. What makes it the same arrangement, and hence the same baseball, is the maintenance of the

capacities constitutive of baseballs, just as maintenance of the capacities constitutive of persons is what accounts for an individual person's persistence over time.

Goldwater (2014b)[27] argues along these same lines that both sides of the debate over composite objects have mistakenly identified those objects with the "mereological" sum of their parts, whereas it is not the sum, but the arrangement of those parts which matters. He asks the reader to consider the following question: "if tablewise arrangements play the role of tables in perception and discourse, whereas composites of simples do not, might this suggest the table just *is* the tablewise arrangement, rather than the composite?" (Goldwater 2014b, 3) And he concludes:

> My answer is yes. That is, I argue that a table just *is* a tablewise arrangement, and a chair just *is* a chair-wise arrangement. More generally, I argue that all ordinary material objects (the inanimate ones, at least) just are *arrangements* (of simples, most likely). Correlatively, I deny that ordinary objects are *composites* of simples (in the way the nihilist and universalist conceive of them); instead, they have a different nature.
>
> Moreover, an existence claim is not far behind. For if there are tablewise arrangements, and tablewise arrangements are identical to tables, then there are tables. Thus, by showing (or reaffirming) there are such arrangements, I defend the existence of ordinary objects- whatever the fate of mereological sums. (Goldwater 2014b, 3)

So if what I have so far been calling composite objects are not merely sums of parts, but arrangements of them, and the Elminativists (or 'nihilists', as Goldwater calls them) admits that there are such arrangements, then it turns out that they believe in so-called composite objects after all. As Goldwater understands it, arrangements are multigrade relations "expressible by variably polyadic predicates such as 'arranged

[27]Goldwater, also has his own analysis of the notion of logical category (Goldwater 2014a)

tablewise'" or else by "other linguistic forms such as names," (Goldwater 2014b, 10) though some nominalists might wish to resist such an option. Similarly, Goldwater assumes there can be both tokens and types of arrangements, though the latter may reek offensively of Platonism to some noses. In general, nominalist convictions might lead some philosophers to scoff and the introduction of one class of objectionable ontological entities in exchange for an abandoned other. I will not address such concerns here except to reiterate that as Goldwater contends, the defender of arrangements need not suppose any ontological claims that have not already been assented to by the Eliminativists, who agree that there are particles arranged in such and such a way.[28]

As I see it, arrangements should not be understood as existing independently of particles that are so arranged. However, arrangements of particles, like composite objects as ordinarily conceived, admit of changes in the actual particles so arranged, so is there is no trouble identifying the arrangements with the objects. Furthermore, the

[28] He does also offer the following argument against a "nihilist-cum-nominalist", i.e. one who believes in simples arranged table-wise, but not in tables as arrangements: "consider specifically a mereological nihilist-cum-nominalist account. On this view, there are simples arranged tablewise, but there are no tables, and no tablewise arrangements either ('arranged tablewise' being a predicate applicable without incurring its own commitment to tablewise arrangements). One consequence of this view is that a person (assuming one exists) cannot *perceive* a table- since a table does not exist to be perceived (obviously, I'm taking 'perceive' to be factive here). So what does the person perceive? It can't be the tablewise arrangement, since that doesn't exist either on the nihilist-cum-nominalist view. The only remaining answer, then, is that it is the simples which are perceived. But simples can't be perceived- they're too small. Tablewise arrangements, however, are perceptible (and they're just the right size and shape). So even the mereological nihilist – i.e. she who denies there are composite objects such as tables – should at least accept the tablewise arrangement. Or else it is hard to see just what someone is seeing when they look at an alleged table." (Goldwater 2014b, 15)

causal efficacy of objects such as baseballs, should not be attributed to the mereological sums of particles, but to the arrangements they participate in. A bunch of atoms cannot break a window unless they are appropriately arranged. It is the arrangement, i.e. the baseball, that does the breaking. As Goldwater puts it: "scattered atoms do not have the same causal powers that those same atoms would have if arranged more densely. As the only difference between these scenarios is their arrangement, the difference in causal power is attributable to that arrangement." (Goldwater 2014b, 13) So the baseball does have causal powers that its parts don't have after all. But that doesn't mean the baseball isn't fully reducible. It is nothing over and above its parts and the relations between them, specifically those relations, or that single multigrade relation, which is their table-wise arrangement.

The identification of composite objects with arrangements shows the way toward diagnosing second major error in Merricks' Overdetermination Argument.[29] (The first error was thinking that composite objects are supposed to be identified with some specific group of particles) That argument depended on the idea that if, e.g. baseballs, exist, then both they and their parts arranged baseball-wise have the power to break windows. However, there is sleight of hand concealed in the power attributed to the particles. The particles are on a lower level of explanation than is the window. The window is on the level of the baseball. If baseballs don't exist, then neither do windows,

[29] This argument is also effective at defusing the related Eliminativist argument from co-location involving, e.g., statues and clay. The statue is not to be identified with the lump of clay, but with the clay arranged in a statue-shape.

so if Merricks is right, then nothing has the power to break windows, particles nor baseballs. He should have said that the particles arranged baseball-wise have the power to scatter the particles arranged window-wise. But if he had said that then there would be no redundancy - baseballs break windows, particles arranged baseball-wise scatter particles arranged window-wise. Or given identification of composite objects with arrangements, the situation can be stated more accurately as follows: baseball-wise arrangements of particles, i.e. baseballs, destroy window-wise arrangements of particles, i.e. windows.

One potential problem for identifying composite objects with arrangements lies in the individuation and persistence of arrangement tokens. What makes one table-wise arrangement distinct from another, assuming that tables can be moved from one spatial location to another, and even dismantled and reassembled? Goldwater himself declines to make any claims about the persistence or individuation conditions of arrangements in general, which, I think, is just as well, because there probably are none except that whatever capacities are constitutive of that sort of arrangement function separately (individuation) and are maintained (persistence). One must attend to the particular characteristic properties of each arrangement, if not arrangement type, to know what counts as separately functioning, and what sorts of changes the arrangement can persist through. Living, organic beings do not seem capable of surviving certain kinds of dismantling. In the case of human persons, dismantling of the brain or disconnecting it from natural or artificial life support is sufficiently disruptive to the capacities for self-

consciousness and concern as to entail annihilation, though an android person could potentially be more resilient, should the parts of the android brain have such properties that if they were reassembled would allow the android to resume the use of its person-constituting capacities. In the case of a human brain, disassembly does not yield parts that have such properties.

III. Parfit and the ID thesis

Parfit (1984) argues for a kind of Reductionist position, influenced by Hume and the Buddhist tradition, that he calls the "bundle theory" of persons or "constitutive reductionism", which holds that persons are nothing more than series of bundles of constituent psychophysical states. These constituents change from moments to moment, so that each bundle has only a temporary existence. These bundles are strung together from moment to moment by the relations of psychological continuity and connectedness. Parfit contrasts the bundle theory primarily with the Inflationist "ego theory", the view that there is a persisting, perhaps immortally so, 'self', 'soul' or 'ego' which is separate from the fleeting psychophysical elements of which a person's mind-body are composed and which thereby accounts for a person's continued existence over time despite the transience of the psychophysical elements.

Setting aside for the moment the specifics of Parfit's account of personal identity (which I have already largely disagreed with in the previous chapter), Parfit sums up the most general form of the Reductionist position with regard to personal identity in the

following two claims:

> (1) that the fact of a person's identity over time just consists in the holding of certain more particular facts…
> (2) that these facts can be described without either presupposing the identity of this person, or explicitly claiming that the experiences of this person's life are had by this person, or even explicitly claiming that this person exists. These facts can be described in an *impersonal* way. (Parfit 1984, 210)

The second claim is often referred to as the Impersonal Description (ID) thesis, which holds that all the facts about persons can be otherwise stated, without remainder, in an impersonal language, one that does not refer to persons or their identity. This claim is roughly equivalent to the Buddhist view that persons (among most, sometimes all, other things) have a merely "conventional" existence. This "Buddhist Reductionism" rests on a distinction in Buddhist philosophy between what is 'ultimately real' versus what is merely 'conventionally real', i.e. between what is real independently of the perspectives, purposes and concerns of persons and what is only real relative to those perspectives, purposes and concerns, insofar as marking something out as significant and distinct from other things has some utility for the purposes of survival, experience, communication, discourse, etc.[30] According to Buddhist Reductionism, as Siderits (2003) calls it, persons, like other composite objects are not ultimately real, because what is conventionally referred to as a person is only a series of distinct, momentary collections of psychophysical elements. Nevertheless, persons are real in the conventional sense, because grouping some such momentary psychophysical elements

[30] Though Carpenter (2014) argues that the Buddha's teaching should be interpreted as saying that any positive view about the nature of persons or selves should be avoided because it is the source of ego-clinging and therefore, suffering.

into collections and those collections together into some temporally extended series rather than others has a certain utility. This utility is grounded synchronically in the spatiotemporal contiguity between the (physical) elements and diachronically in the causal connections that obtain between the momentary collections of elements. Buddhist Reductionism is, therefore, meant to be generally Reductionist, not Eliminitavist.

The distinction between Reductionism, Non-reductionism and Eliminativism is illustrated in the Buddhist literature through the example of the chariot. As Siderits explains it:

> ...'chariot is a convenient designator for a set of parts assembled in a certain way. Thus while there are ultimately no chariots, there are those wholly 'im-chariotal' facts into which all chariot-talk may be reductively analyzed; it is these facts that explain the utility of our talk of the fiction… Given this utility we may say that while the chariot is ultimately unreal, it is conventionally real. This will be the reductionist view of chariots. The non-reductionist will claim that chariots are both conventionally and ultimately real -- that in addition to the parts of which chariots are composed, ultimate reality also contains some sort of separately existing chariot-essence. And the eliminativist will claim that chariots are both ultimately and conventionally unreal -- that our talk of chariots is misleading and should be replaced by some entirely new way of conceptualizing collections of chariot parts. (Siderits 2003, 7)

Similarly, persons, are seen by the Buddhist Reductionist as conceptual constructions that are only conventionally, not ultimately real. This is the Buddha's self-described 'middle path', developed by the Abidharma schools, between the Eternalism of the non-reductionist Nyaya and Samkhya schools, who believe in a transcendent self, called, respectively, *atman* and *perusha*, and the annihilationism of the Eliminativists, for whom the denial of an eternally and separately existing self entails "that the person

goes out of existence after a relatively brief duration." (Siderits 2003, 13)[31] So if one adopts the ID thesis, then like the Buddhists, one may use the term person as a "convenient designator" or a bit of time saving shorthand, but there would be no facts about persons that cannot be expressed as facts about their constituents.

Parfit contrasts his Reductionism with non-Reductionism (what I call Inflationism), which comes in two major forms that agree in their denial of the two Reductionist claims stated above, though for different reasons. The first form of Inflationism holds that "A person is a separately existing entity, distinct from his brain and body, and his experiences," either "a *purely mental* entity: a Cartesian Pure Ego, or spiritual substance" or "a separately existing *physical* entity, of a kind that is not yet recognized in the theories of contemporary physics." (Parfit 1987, 210) I reject this form of Inflationism about identity/survival for the same reasons I objected to it as an account of personhood in chapter one. If the separately existing ego is a non-physical entity and hence unobservable and unexplainable by observable processes then making its existence a necessary condition of personhood would violate the present study's commitment to naturalism and if the ego is supposed to be a separately existing physical entity, then there is so far no empirical evidence that such a thing exists. Furthermore, if the ego is independent of and distinct from any particular psychophysical functions, then its continued existence would not guarantee the

[31] This sort of Eliminativism, of course, only rules out persons as diachronic, not instantaneous beings. However, instantaneous persons could never be responsible for actions performed in the past and so considering them persons would be of no utility.

continuance of any of the features relevant to personhood, i.e. self-consciousness and concern.

The second form of Inflationism is the so-called “further-fact” view, which Parfit states as follows: “though we are not separately existing entities, personal identity *is* a further fact” beyond any enumerable impersonally construable psychophysical facts. (Parfit 1987, 210) This view can be interpreted as a genuine ontological thesis or else as a merely linguistic one. The former amounts to the claim that facts about persons cannot be fully analyzed into concatenations of psychophysical facts. The latter, linguistic, reading of the further-fact view, is simply a rejection of the ID thesis, not of Parfit’s first Reductionist claim, insisting that facts about persons are not fully translatable into an impersonal language. In other words, the linguistic version of the further fact view holds only that we can’t say everything true about persons without using the word ‘person’. The ontological interpretation of the claim is anathema to naturalistic Reductionism, for it would imply a disunity between personal and subpersonal levels of explanation. If facts about persons are something more than concatenations of psychophysical facts, so that no good explanation can in principle be given of how the personal facts arise from the subpersonal facts (including physical, chemical, biological, and psychological facts, etc.), then there is a genuine gap in the scientific worldview. However, interpreted linguistically, the further-fact view by itself poses no such threat. This is because there are some reasons why we cannot do without the term ‘person’ that do not imply a disunity between explanatory levels.

Sorabji (2006) attacks the ID thesis, arguing that the ownership and authorship of one's mental states and actions constitute facts about persons that cannot be fully described by referring to non-personal elements. In other words, Sorabji takes the facts about ownership and authorship to be what are left out if one uses only impersonal language, and he thinks that Parfit concedes this in a 1999 paper where the latter revisits the Reductionist position. "The idea that thoughts, acts, or experiences are *had by*, or *performed by*, something..." (Sorabji 2006, 266) is not something that can be rephrased in purely impersonal terms. Therefore, "ownership, which is part of the concept of a person, is no longer included as something deducible. This seems to be a weaker form of reductionism, [than Parfit's earlier view] in that the account is not of a person, but only of a person's *components*. And it might be added that the components themselves are under-described, in that it is omitted that the mental processes and events are *owned*." (Sorabji 2006, 266) However, for Parfit, elimination of facts about ownership is not elimination of anything of great importance. The difference between the personal and impersonal schemes

> ...is not metaphysically deep… is in part merely grammatical. In our [personal] scheme, all thoughts, experiences, and acts are claimed to be *had by* or *done by* either some persisting body or embodied brain, or some distinct entity that has this body and brain. In my imagined scheme, these thoughts, experiences, and acts might instead be claimed to *occur in* this persisting body or embodied brain… I do not see the importance of this distinction…. my imagined beings… would be missing certain truths, since it is true that all thoughts have thinkers, and that all experiences have subjects. But this is like the truth that, for every continuous flowing of water in a certain pattern, there is a river which does the flowing. And that truth does not have to be understood in any adequate understanding of such flowings of water. The same may apply to the truth that, for every stream of thoughts or experiences, there is an entity that thinks these

> thoughts and has these experiences. This metaphysical scheme… is no worse than ours. (Parfit 1999, 260-62)

Parfit takes the facts about ownership to be irreducible but trivial, apparently in the sense that our forensic and ethical practices do not depend on them. Sorabji disagrees, and attempts to show that various everyday statements of great importance about persons are inexpressible without referring to persons. Most of Sorabji's arguments, however, target Parfit's own specific variety of Reductionism, and which appeals to streams of consciousness formed from links between atomistic mental states. Such an appeal, Parfit believes, allows him to say, in contrast to the 'hyper-reductive' views of Williams (1970), Thomson, and Nagel (1986), that persons are logically distinct from their bodies or brains (such that they could persist in different bodies or brains over time) though they are not separately existing entities. (Parfit 1999, 218) This prevents Parfit from saying that the body or brain is the subject of experiences. Instead, Parfit wants to say, along with his imagined Reductionist beings, that experiences occur "in some persisting body…" without the body "or any other entity..." being "the subject of these experiences, the thinker of these thoughts, or the agent of these acts." (Parfit 1999, 228) Thoughts and decisions, under this impersonal conceptual scheme are mere co-located happenings, not acts or properties of individuals, so that no reference to the individuals that have them is needed in order to describe such 'happenings'. That there are facts about individuals who have such thoughts and make such decisions, is, for Parfit, true, but trivial. None of our usual practices concerning persons depend on them.

I think Parfit is wrong here on all counts. The facts of ownership are reducible, they are not trivial, yet they cannot be described impersonally. To begin with the latter two points, Sorabji offers persuasive reasons why Parfit's version of Reductionism, by eliminating facts about ownership, cannot do justice to crucial aspects of agency and ethics. Most relevant to the current study are his arguments to the effect that judgments of responsibility, and the attendant practices of praising and blaming, are inexpressible or incoherent if we are restricted to an impersonal language. He says:

> First, what about deserving credit or blame? We are not now being allowed, except as a way of talking, to think of a *person* as deserving credit or blame. Rather it would be the *act* that deserved credit or blame, and the resulting stream. But this would have to be in the different sense that it would be more admirable, or less so, just a sunset may be admirable, without anybody *deserving* credit or blame… Could we... substitute for the idea of deserving punishment the idea of using punishment to deter? Deterrence would be difficult to effect if, in the absence of owners, there is no one who would suffer from deterrent measures, and no one who would benefit from their being applied. (Sorabji 2006, 275)

Sorabji is right that actions, experiences, and streams of consciousness cannot be responsible for themselves. However, as I argued in chapter two, the relations of psychological continuity and connectedness that form Parfit's streams of consciousness are not the best candidates for what constitutes persons and their persistence over time. Rather, endorsing the alternative view I have developed, which sees persons as persisting due to the maintenance of the capacities constitutive of personhood, allows one to reduce facts about persons to facts about the things which are organized in such a way that they instantiate the capacities. According to this view the person is identified not with a stream of consciousness or series of psychological events, nor with any

particular brain or body, but with whatever arrangement of components continues to maintain uninterrupted instantiations of the constitutive capacities. Some fact about a person at a time can then be reduced to some fact about whatever parts of the person are currently instantiating the relevant capacities, usually the brain and central nervous system, which can in turn be reduced to the organization of neurons or other elements. However, reference to the psychophysical constituents of a person presupposes reference to the person herself. After all, we describe their organization by saying they are arranged person-wise. Alternatively, we might say that they are arranged in such a way that they instantiate the capacities of self-consciousness and concern, but psychological capacities are necessarily capacities of beings who have them. They don't exist independently of those beings like Platonic Forms, but are immanent to them, like Aristotelian formal properties. An individual capacity for self-consciousness or concern must be the capacity of some individual. If an individual possesses both such capacities then that individual is a person. So long as those capacities are maintained uninterruptedly, it is the same person who has them the whole time.

The crucial point here is that capacities don't float around by themselves but are what they are, in part because they are owned. This is analogous to a club or a nation. A member of a club is only a member is we presuppose that there is a club. However, that doesn't mean that the club is not reducible into its members and the relations between them. The same goes for a nation and its citizens. If, as Parfit supposes, persons are like clubs or nations, then it is no surprise that certain facts about them

cannot be described impersonally, even though they are fully reducible. Of course nations don't own their citizens in the usual sense of property but in a looser sense of 'own' which just implies having them as constituents and that their identity as citizens is dependent on them being constituents of the nation. If one prefers, 'bearers' could be substituted for 'owners'. In any case, components of an organized whole, whether concrete parts or abstract properties are what they are at least partially in virtue of belonging to that organization. A heart is not a heart unless it's pumping blood through a body. A capacity for brittleness depends on there being an object that is easily shattered. However, the conceptual priority of the organization does not imply its irreducibility.

Organized entities with the capacity for self-consciousness and concern can be responsible for their actions and are the appropriate objects of praise and blame. They can be reduced to their components, but the components themselves cannot be held responsible, so that responsibility is an emergent property of organized beings. But that doesn't mean that persons are irreducible to their components, nor that facts about ownership cannot be reduced to facts about the relations between components. My actions belong to me as the author of them because they are performed by my brain and body, which are mine because they are, perhaps only temporary, components of me. They are indescribable in isolation from their role as components of me, but that doesn't make me something over and above them in the inflationist sense.

IV. Shifting Coalitions and the Extreme Claim

The version of Reductionism I endorse asserts only the first of Parfit's two theses, which was the following:

(1) that the fact of a person's identity over time just consists in the holding of certain more particular fact.

For my purposes it will help to broaden and sharpen that claim - broadening it to include all facts about persons, not just those about identity (with which I would expect Parfit to be in agreement); and sharpening it by specifying just what more particular facts, the facts about persons consist in. For the latter, I have already explained that facts about persons consist of facts about the organization or arrangement of psychophysical constituents, such that they instantiate and maintain the capacities for self-consciousness and concern.

Inflationists, on the other hand, besides denying the ID thesis, insist that there is more to persons than the above. Persons possess properties or powers that are not fully explicable in terms of the organization of their components and without which they would not be the sorts of beings that are responsible for any of their actions. In other words, the Inflationists hold that Reductionism about persons, of the sort I have endorsed, implies the Extreme Claim as regards responsibility and hence slides into Eliminativism. If Reductionism is true of all the beings we ordinarily call persons, then the Inflationist claims, those beings are not actually persons, since they are incapable of being responsible for their actions.

I have already given reasons for rejecting the sort of inflationism that rests on the claim that the existence of persons depends on some element that is distinct and independent from any of the person's psychophysical constituents. Presently I will only consider that version of Inflationism which appeals only to "further facts" about persons in the genuinely metaphysical sense above distinguished from the merely linguistic sense.

Why there should be irreducible facts in the absence of an irreducible soul or ego is a mystery to me, but it is just this mysteriousness that is the core of the further fact view, according to which persons are possessed of properties and powers whose underlying causes no scientific investigation will ever reveal, because there are no such causes. This view goes hand in hand with libertarian views about free will, which hold that persons are capable of initiating actions that are not completely determined by previous events. I have already rejected such a capacity in chapter one as a necessary condition of personhood because it is in conflict with a commitment to naturalism. However, proponents of libertarian free will claim that it is necessary for responsibility and therefore personhood, so to reject it would entail the Extreme Claim. In response I have offered a compatibilist conception of responsibility following Frankfurt. Still, there are some objections to the idea that a coherent account of responsibility can be built upon Reductionist premises, so that Reductionism will entail the Extreme Claim after all.

One such objection that I find worthy of defusing is addressed by Siderits (2003) as part of his general defense of Buddhist Reductionism, though this particular

argument is threatening to any Reductionist, even one who does not accept the Buddhist variety with its claim that person is a mere conventional designation.

Siderits is concerned to show how persons could be capable of self-scrutiny, self-control and self-revision (examples of the kind of downward causal powers appealed to by Inflationists) without violating what he calls the "anti-reflexivity" principle which states that "an entity cannot operate on itself." (Siderits 2003, 27) Self-scrutiny, in particular, is the sort of Frankfurtian higher-order judgment of one's beliefs and desires which requires the capacities self-consciousness and concern, and that I argued in chapter one, is necessary and sufficient for personhood. According to proponents of the argument that Reductionism implies the Extreme Claim, self-scrutiny requires a separate self as subject and chief executive with one's particular mental states comprising its object, because if the mental states that play the role of the subject that scrutinizes and are ever themselves objects of scrutiny, which any state seems potentially capable of being, then some mental states would have to serve as both subject and object, and that would violate the anti-reflexivity principle. "For if each of them is a potential object of the executive function, and an entity cannot operate on itself, then it seems that none of them could be the one enduring subject that performs this function." (Siderits 2003, 26) This is why it seems that responsibility requires a distinct self, something that scrutinizes all.

Siderits invokes the "shifting coalitions" conception of self-revision as Reductionist alternative to the Inflationist's distinct self. He rightly points out that the

anti-reflexivity principle is only violated if the same mental states are subject and object *simultaneously.* A Reductionist can endorse an account of self-scrutiny that appeals to shifting coalitions of mental states playing the role of chief executive at different times, such that each coalition can be the object of scrutiny at the times when it is not the subject. According to the Reductionist, the temptation that leads to positing a distinct self and that should be resisted is to take a particular set of mental states that play the subject role relatively frequently and hypostatize them into an enduring subject. As Siderits puts it: “Thus arises the notion that a person has an essence -- that some constituents are more central to the existence of the person than others.” (Siderits 2003, 27) Holding this view would require one to deny that the set of mental states taken to be the subject is itself subject to scrutiny. But if one has no way of revising the chief executive, then one can’t be responsible for the way that executive scrutinizes and potentially controls and revises one’s other mental states and behavior. Like any good commonwealth, there need to be checks and balances on executive authority. “For instance, when I decide to curb my bedtime snacking I may be employing a particular standard of acceptable body shape, which I may subsequently decide is politically problematic and morally questionable.” (Siderits 2003, 26) The Inflationist claims that the self, being independent from the psychophysical elements, is, like an absolute monarch, the sole source of independent valuation in a person (and may be propped up by a conception, usually religious, of an infallible conscience or divine mandate). But if one, for good reason (namely, that we have no evidence for such a thing and the

concept of it may be internally incoherent in various respects) denies that such a distinct self exists and says that the executive function is played by some mental/brain states, those mental/brain states must also be subject to scrutiny by the rest of the person/brain at some times. Siderits offers the shifting coalitions view as a solution to this problem:

> If I am to be capable of revising [or at least scrutinizing] my own character, then I require a stock of beliefs and desires on the basis of which I may critically evaluate and seek to reform various of my dispositions and tendencies I am called upon to monitor. It may now seem as if, were they to constitute a part of the 'I' that performs self-revision [or scrutiny], then the anti-reflexivity principle would be violated. But what this picture omits is the possibility that a given stock of beliefs and desires might serve as a basis for a particular bout of self-criticism, yet some among these stand under subsequent scrutiny on the basis of a distinct, (though perhaps overlapping) stock of beliefs and desires… On one occasion my anal-compulsive disposition might lead to extirpation of the desire to smoke. Yet, subsequently a wish to be more accommodating to others might lead to an effort to curb my anality. At one time the anal disposition belongs to the coalition making up the 'executive', later it falls out of this shifting coalition. (Siderits 2003, 65)

The shifting coalitions approach posits a kind of feedback loop between mental/brain states, which allows one to have a sense of self-determination that depends on nothing that is undermined by Reductionism. Each coalition that at one time plays the role of executive can be at another time the object of a different coalition's scrutiny as well as control and revision. Even if the activity of each coalition is causally determined, the fact that there are internal checks and balances and that I can't be aware of all the facets of my psychology at once yields a rationally tenable sense of self-determination.[32]

[32] Nietzsche seems to have something like this in mind in his analysis of the phenomenon of willing: That which is termed "freedom of the will" is essentially the affect of superiority in relation

Siderits' view is vulnerable to an immediate objection: if Reductionism is interpreted in a particularly strong sense. If persons are reduced to collections of simples of minimal magnitude or reduce infinitely into elements of infinitely smaller magnitudes (gunk), a kind of mental punctualism seems to follow. If mental states are fleeting, momentary things, then a particular set of beliefs and desires would not have the temporal duration required to be at one time the subject and at another time the object of scrutiny. So for the shifting coalitions view to work, Reductionism cannot be conceived in such a way that it implies mental punctualism or atomism. Reductionists should not make the mistake of denying temporal extension to mental states. Beliefs and desires must supervene on physical processes in such a way that they exist for some duration, long enough to be both subject and object of revision. Furthermore, our current understanding of how mental states are realized in the brain suggests that this is the case. As Brown (2006 and 2013) argues, thoughts should be identified, not with static configurations of neurons, but with patterns of synchronous neural firing. When philosophers claim that mental states are identical to or supervene on brain states, the

---

to him who must obey… A man who *wills* commands something within himself that renders obedience, or that he believes renders obedience… We are at the same time the commanding *and* the obeying parties… "Freedom of the will" -- that is the expression for the complex state of delight of the person exercising volition, who commands and at the same time identifies himself with the executor of the order -- who, as such, enjoys also the triumph over obstacles, but thinks within himself it was really his will itself that overcame them. In this way the person exercising volition adds the feelings of delight of his successful executive instruments, the useful "under-wills" or under-souls -- indeed, our body is but a social structure composed of many souls -- to his feelings of delight as a commander. *L'effet c'est moi.* what happens here is what happens in every well-constructed and happy commonwealth; namely, the governing class identifies itself with the successes of the commonwealth. In all willing it is absolutely a question of commanding and obeying, on the basis, as already said, of a social structure composed of many "souls." (Nietzsche 1886, I.19)

term must, somewhat misleadingly, refer to such patterns of firing if those claims have any chance of being true. I don't mean to rule out the possibility that mental states can be realized in some other way, but if we see that in the case of human beings, the neural foundations of the mental are extended in time, then it may be less difficult to understand how the mental states themselves could be so extended (perhaps indefinitely).

The shifting coalitions strategy will not satisfy such thinkers as Strawson (1986) or others who think that Reductionism undermines self-determination because it implies causal determinism. Strawson thinks that to see oneself as self-determining, one must think of oneself as ultimately responsible for one's character as well as one's actions. Even if one's character is formed internally by the feedback mechanism of shifting coalitions, Strawson would argue that the way in which this system functions is determined by factors before one's birth. Siderits offers the following example of the kind of self-determination an agent must be responsible for: "So my miserable childhood resulted in a predisposition to behavior that causes trouble for myself and others? Others tell me to stop kvetching. I agree, and set about trying to reform and improve my character." (Siderits 2003, 64) The shifting coalitions model shows how, for a Reductionist, such an example of self-revision is possible. However, Strawson would object that in such a case whether I am or am not "the sort of person" who would respond that way to the criticism of others, or who can find the strength within myself to push back against the forces of my upbringing, is not really up to me. In other words, it

is still a matter of deterministic luck whether or not I have the right coalitions, with the necessary strength to bring about a particular act of self-revision.

The shifting coalitions view allows us to accurately distinguish between two types of phenomena, i.e. cases where one's actions are compulsive or automatic, and ones where one's actions are caused by inner states that have been subjected to self-scrutiny, those which one may be responsible for. This is not the kind of full, ultimate responsibility that Strawson is interested in. Nothing short of a genuinely transcendent agent acting outside of the deterministic causal matrix could fit that bill. However, it is also not the weak sort of responsibility that Siderits rightly rejects,

> according to which it is enough that the action 'come from within' the agent, regardless of how the agent came to have the particular beliefs, desires, dispositions etc. from which the action flowed. But this temptation should be resisted, since we do expect agents to take responsibility not just for their actions but also for their own character… Being responsible for my actions means being responsible for being the sort of person who would perform those actions. Any account of freedom that omits this is justly criticized as too weak. (Siderits 2003, 64)

The sort of responsibility made possible by the shifting coalitions strategy is distinguished by the recognition that actions 'come from within' in different ways. The ones that result from a process of dynamic self-scrutiny are the ones that we may be responsible for. As long as that is the only sort of responsibility required for personhood, then Reductionism about persons need not imply the Extreme Claim as regards responsibility, and therefore, need not slide into Eliminativism.

Chapter Four: ‘Person’ and ‘Self’

In several places throughout the preceding chapters I have left promissory notes regarding the term ‘self’, saying that I am reserving it for special purposes which I am now prepared to explain. There is an ordinary use of the word ‘self’ (and related words where it is conjoined with something else - such as ‘myself’, ‘yourself’, ‘himself’, ‘herself’ etc.) according to which the word has a merely indexical use, referring to a person. Locke had this sort of use in mind when he says “Person, as I take it, is the name for this self,” apparently jabbing a thumb in his own direction, and continues: “wherever a man finds what he calls himself, there, I think, another may say is the same person.” (Locke 1690, ii.xxvii.26) And a bit before that: “so far reaches the Identity of that *Person*; it is the same *Self* now it was then; and tis by the same *Self* with this present one that now reflects on it, that the Action was done.” (Locke 1690, ii.xxvii.25) However, there is also a tradition of using ‘self’ to refer to something that is not identical to an entire person, but is only part, if an essential part, of one. In this latter usage, ‘self’ is often taken to be synonymous with ‘soul’ and understood to be a part that is separate from any of a person’s mental and physical components, though it is also thought to account for the identity of a person over time, in both the metaphysical (as discussed in chapter two) and socio-psychological (which I will explain below) senses. Additionally, ‘self’ has been thought to account for subjective experience and the unity of consciousness. So far I have rejected accounts of personhood and personal persistence that have appealed to an enduring, separate or distinct, self on naturalistic grounds. However, that

does not mean that there are not good naturalistic candidates for what 'self' might refer to when not referring to an entire person. Furthermore, while my account of personal identity does not appeal to such a 'self', at the same time it also does not appeal to continuity of distinctive psychological states. For that reason, people interested in 'identity' in what I call the socio-psychological sense, will at this point think that my view of persons is seriously impoverished. Social psychology is largely interested in people's 'identities', meaning what characteristics they take to be most central to their own sense of the sorts of persons that they are. Social psychological theories of identity, appeal to individuals' perception of their own personhood, how they feel about their bodies, their membership in social groups, their relations with other individuals, the music, television shows and other art and media they enjoy, their moral codes, and their styles of dress, among many other factors that contribute to how these individuals self-identify. For this reason, this conception of identity, or what I will call 'the self,' as opposed to 'the person,' is a fundamentally subjectively constituted sort of thing. It is constituted by the ways in which individuals are aware or at least think they are aware of the persons they are. I don't mean to suggest that such identities are formed in isolation from other people. Our selves can be partially or even largely socially constituted in the sense of 'social' distinguished by Greenwood (1994): that we hold them because we believe they are held by members of the social groups to which we belong, as well as the broader sense of the term, which it is better to call 'interpersonal': that our particular ideas about our selves are influenced by other people and our interactions with them. However, in

the end, each self is a mental product of the bodily experiences and neural activity, or analogous implementation, of one individual person and what that person takes her or himself to be.

I understand the issue concerning that social psychological sense of self to be distinct from the one about metaphysical identity or persistence as well as the issue about continued responsibility, but I do not find it uninteresting, unimportant or delusory. Therefore, I propose a conceptual division of labor. Talk of persons and their identity and persistence should cover the metaphysical question which I addressed in chapter two. The term ‘self’ will be reserved for talking about the social psychological issue of identity, but also questions about subjectivity and the unity of consciousness, because I take them all to be related. They all have to do with the way in which individuals (some of which are persons) experience themselves as distinct from, but in various ways, related to the rest of the world. With this division of labor, seemingly paradoxical statements about persons and selves can be translated in ways that make them coherent. For example “He’s not himself today” can be taken to mean, “this person has a largely different self today from the one he usually has” so as to avoid the paradox that one and the same person can be a different person at one time than at another. The self changes though the person remains the same. More controversially, “she’s not the person she used to be” can be translated as “this person has a radically different self than the one she used to have.” The latter translation is more controversial, because there the word ‘person’ in the sentence is actually being replaced by the word

‘self’, while ‘person’ takes the place of the pronoun ‘she’. However, as I will endeavor to show in this chapter, I believe such a revisionary move is theoretically justified (though it might not be convenient or pleasant to actually talk that way in everyday life). The self, in my sense of the term is what one is conscious of when one is self-conscious, i.e. the object of self-consciousness, and hence having one or more selves is necessary for being a person. As I hope to make clear in what follows, this use of ‘self’ is not entirely unrelated to the indexical use, because ‘selves’ are formed by subjectively indexing mental states to individuals. In pathological cases, where a person indexes their states to more than one distinct individual (what the person might consider distinct persons, but which are really distinct selves), multiple selves arise, which may appear to the person so affected as different persons though that is not objectively the case.

The first part of the chapter will be an explication of what I mean by ‘self’ and why I think it should be treated as a distinct concept from ‘person’. This will lead to a discussion of whether or not there is a self as the subject of experience, taking on arguments such as those by Hume (1748) and more recently, Prinz (2012), to the effect that we have no experience of such a thing. I will argue that the critique of self-experience is correct in saying that there is no experience of self as distinct from mental particulars, but that the experience of self is the experience of a mental object, which is itself a complex and dynamic mental particular. Insofar as there is a subject of experience, it is not itself experienced, nor is it properly called ‘self’, but is only the perspective from which a person experiences the world and her self as object or else

merely the thoughts a person has about her self. Next I will offer my positive account of the self as dynamic internal representation of an organized being. To illustrate the sort of thing I have in mind, I will engage with neurocognitive accounts developed by Damasio (1994 and 2011) and Metzinger (2004), but my view is compatible and potentially continuous with various theories of the self as biologically and socially constructed. My view is that the 'self' is a dynamic internal representation of an organized being's psychological states (which themselves, on the first order, are representations of the individual's bodily states and actions, as well as of properties of the outside world and the being's relations to it) of which, in some cases, usually of persons, the being is capable of being aware, via even higher order representation. Since that awareness is only partial and there is the potential for inaccuracy at every order of representation, a person can be wrong about her self. Her self can misrepresent her, the person's, first order mental states as well as bodily states and properties of the world/relations to it, and she can also misrepresent her self in introspective self-consciousness. Discussion of Damasio's idea of the 'autobiographical self' will lead to consideration of whether or not the selves of persons must be narratives. I will argue that they need not be explicit or conscious narratives, but just having what Damasio calls a 'core self' in the first place requires a kind of minimal, implicit narrativity. The last part of the chapter will deal with specific characteristics of the self, particularly what kind of entity it is, ontologically speaking, as well as whether it is necessarily unified or stable. I will argue that it can be understood either as a mental

object, i.e. a representation or as a series of (usually neural) processes, because the neural processes are identical with the representation. Furthermore, selves are necessarily unified and stable if only from the perspective of the being that has them. In self-conscious beings there is a possibility of multiple selves in a single being due to the subjective indexing of mental and physical states to distinct representational units or individuals. The paradigmatic examples of such multiplicity of selves are cases of dissociative identity disorder (DID).

I. Self vs. Person

The first reason to distinguish between 'self' and person is that, given the account of personal persistence discussed in chapter two, the fact that a person persists over time says nothing about what that person is like other than that he or she continues to be a person. However, when one thinks about the 'identity' of a person, in various social and psychological contexts, one is interested in what that person is like, e.g. what her distinctive character and values are, what groups she sees herself as affiliated with, who her family members are, what her occupation is, what her artistic preferences are, etc. The main intuition that leads me to the view of persistence defended in chapter two is that all these things could change radically and yet the same individual person would remain. However, I do not deny that there is a real sense in which a person may shed an 'identity' and assume a new one or in which I may not 'be myself' some days. To talk about the kind of identity that is built up out of these

contingent, distinctive factors, that may come and go throughout the life of a person, it is appropriate to talk about the 'self' as a feature of a person, conceptually distinct (but not ontologically separate or independent) from the person.

A second reason for distinguishing between 'person' and 'self' is that the distinction corresponds to a distinction between the objects of inquiry concerning responsibility for an action versus the amount of praise or blame/reward or punishment appropriate for that action. Recall Schechtman's (2014) distinction between a person as a "forensic unit" and as a "moral self," according to which the former notion marks out which beings are appropriate targets for any forensic inquiry, whereas the latter is the object of particular forensic inquiries into whether or not a person is responsible for a particular action performed in the past. I have in mind a similar distinction, except that I don't think that a change in moral character has any bearing on continued responsibility for a past action. As I see it, if a person at one time persists as a person at another time (the earlier forensic unit persists as the later), then the latter is responsible for anything for which the former is responsible. However, inquiry into the moral self of the latter, as compared to the former, is required in order to decide the degree to which praise or blame/reward or punishment is appropriate. The idea is that while a person is always responsible for, e.g. the bad things she has done, if she has changed morally over time, so that she would not repeat such an action or has come to feel remorse for the action, then some of the work that blame or punishment would have done has already been accomplished, making it inappropriate for the same degree of reprobation to be meted

out as would be appropriate if no such change had occurred. However, it is harder to make the same case for morally good actions, i.e., that if a person has changed morally for the worse since the time praiseworthy actions are performed, the same degree of praise/reward is no longer appropriate. There may be an asymmetry between the cases of moral change for the better and for the worse. A person in line for a promotion who later does something disgraceful, would not still deserve the promotion, but that may have more to do with the fact that a promotion is based on what we expect that individual to accomplish in the future, not just what that individual has done in the past. A better example might be someone who is supposed to receive a humanitarian award, but goes on a killing spree the day before the ceremony. However, even in that case our reluctance to give that individual the award might have more to do with not wanting to look as if we're promoting the recent negative behavior than a change in what the individual deserves.

Regardless of whether or not the cases of positive and negative moral change are symmetrical, a useful way to employ my distinction between 'person' and 'self' in these kinds of situations is to say that one might be the same *person* as an earlier one yet have a radically different *self*, such that while one is still responsible for the things one did in the past, one may no longer be deserving of reward or punishment for those actions. This is why in criminal law there is usually a judgment of guilt or innocence by a jury and then a separate decision about punishment determined by the judge, the statutes and precedent.

A third reason for the distinction between ‘person’ and ‘self’ is the sense of ‘self’ as a *subject* of experience, the experiencing, knowing, owning and acting inner agency within the person. This is the sort of thing that Descartes takes to be indubitable along with each act of thought, while Hume denies that he has any experience of it. The conception of ‘self’ in the previous two paragraphs was of a kind of *object* (though perhaps one that is essentially subjective, in the sense of being experienced privately by the one who has it, as well as subjectively constituted). Prinz (2012) has recently defended Hume’s skepticism about the self as subject of experience, or more precisely, has cast doubt upon the idea that we have any experience of such a thing, while Damasio (2011) and Strawson (1999, 2011) have argued in favor of a ‘phenomenal I’ (to use Prinz’s phrase). This debate over whether or not there is an experienced self as subject of experience is not, for the disputants, a debate about the existence of persons. Therefore, it is clear that there is already a conceptual distinction between ‘self’ and ‘person’ in play, and assuming the debate about the self is a real disagreement and not a pseudo-issue (though it might turn out to be), that may be further reason for observing and maintaining the distinction. At this point, it is necessary to delve further into this particular issue of whether or not there is a self as subject of experiences and whether or not we have any experience of such a thing, where I will argue that we in fact do not.

II. Is there a self-as-subject and can it be experienced?

Most of what I have to say about the self in this chapter is about the self as it is an object of experience. However, the self is also sometimes thought of as a subject of experience, an observer of events both mental and physical, a doer of deeds, and a constant presence distinct from those experiences, events, and deeds. This has been the primary role of the self in traditional theories from Hinduism's *Atman* to Plato's *Psyche, to* Descartes' *Cogito* that identified the self with an immortal, immutable soul that lies behind and apprehends one's particular mental states and experiences, and could survive death, persisting into the afterlife/next life. More recently, Strawson (1999) and Damasio (2010) have posited a self as subject of experiences. For Damasio, 'self' is something generated by sufficiently complex nervous systems and, in its most sophisticated form, appears to an individual who has one as two different things, depending on the perspective one takes. On the one hand there is the self as object, which Damasio describes as "a dynamic collection of integrated neural processes, centered on the representation of the living body, that finds expression in a dynamic collection of integrated mental processes," (Damasio, 9) which comes close to the idea of self that I mean to develop in this chapter. Briefly, this conception of the self-as-object is of an internal representation of the states of an individual in relation to its environment, which the individual utilizes to guide its interactions with that environment. On the other hand, according to Damasio, there is also the self-as-knower, the self that apprehends the self-as-object.

Hume's (1738) denial of the self was aimed at the self-as-subject standing behind one's individual perceptions and his argument was that when he introspects, he has no experience of such a thing. He says: "[W]hen I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch myself at any time without a perception, and can never observe any thing but the perception." (Hume 1738, 252)

Prinz (2012) argues at length in favor of Hume's view, by showing that experiences that are believed to be of the self-as-subject are not of it, but rather one or several of Hume's stumbling blocks. First of all, Prinz challenges Damasio's conception of self as knowing subject. He specifically objects to Damasio's account because of its overemphasis on continuity of bodily feelings as the basis of self-experience. Damasio follows William James in taking the "core of sameness running through the ingredients of the Self" (James 1890, 350-52) to be the experience of similarity despite changes in bodily feelings, so that one experiences oneself as being the same subject despite such changes, as opposed to Hume's view that all one experiences is a continuous series of changing impressions. Prinz offers three specific arguments against the Damasio-James view. To begin with, he argues that bodily experience is not necessary for self-experience. Many highly intellectual tasks that are paradigmatic examples of when the self is most present, such as doing philosophy or solving crossword puzzles, are ones where one's body is largely absent from consciousness. Secondly, Prinz argues that

bodily awareness is not by itself sufficient for the sense of self, citing instances when one notices changes in another person's body and infers emotional change from those behaviors, but is therefore focused on the other person and not on one's self. The third argument appeals to the general lack of correlation between bodily feelings and a sense or experience of self. More intense emotions are not accompanied by a greater sense of self, but rather shift one's focus to external matters, e.g. as your feeling of "terror after hearing an intruder enter your house… makes you forget yourself for a moment and focus intensely on the sounds coming from the intruder." (Prinz 2012, 227) Now it seems to me that these arguments do not clearly distinguish self-as-subject from self-as-object, but I agree with the basic idea that awareness of bodily changes or sensations is not necessary for having a self. Changing conditions of the body and the feelings associated with them are some of the things that are often the content of the complex representation of the self as object, e.g. the nagging sensation in the pit of my stomach ever since I learned that a loved one has fallen ill. Damasio may even be right that the selves of the simplest organisms are mostly or wholly composed of somatic representations, but that truth is not a necessary one, nor need it even be contingently true of human beings and other organisms whose selves involve more abstract or intellectual features. I will have more to say about Damasio's conception of the self-as-object in the following section.

While in his discussion of Damasio's view, Prinz does not clearly distinguish between self-as-subject and self-as-object, sometimes he does correctly make the

distinction, as in his critique of Goldberg, Harel, and Malach's (2006) defense of the self as subject. Those authors point to conditions in which one seems to "lose oneself" as exceptions that prove the rule. If there is a self that is recognized to be absent in some experiences, they argue, it must be present in others. For example, Goldberg et al used fMRI scans to compare the brains of people asked to read a list of words and decide whether or not those words were true of themselves versus people asked to read a list of words and decide whether each was a noun or a verb. They found that the first group had greater activity in the superior frontal gyrus (SFG) region of the brain. This increased activity was also found in other self-directed tasks, leading the authors to infer that "the SFG is the neural correlate of self-awareness" (Prinz 2012, 221) and that tasks in which it is active are those where one has an experience of the self-as-subject. Prinz correctly objects to this reasoning by pointing out that

> the Goldberg study can best be regarded as an investigation of the self as object, rather than the self as subject. In their tasks, we report things about ourselves, but in so doing, we are treating the self as just another thing in the world with certain describable features. We are not experiencing ourselves acting as the subject of thought or experience. This is not the elusive self as I. They do not establish that some thoughts have a qualitative component that occupies the same position that the word *I* occupies in self-ascriptions, such as 'I like this music.' (Prinz 2012, 223)

I agree with Prinz here and think that he could have extended this point to cover many of the other supposed pieces of evidence in favor of an experienced self-as-subject. It just seems to me incoherent to claim that the self-as-subject is experienced. To make the incoherence plain, we can paraphrase the claim as "the self-as-subject is the object of some experience." How can one thing be both subject and object at once?

This was the paradox of self-scrutiny addressed in chapter three, but there I advocated the shifting coalitions view which claims that the subject of self-scrutiny is never the same element of a person as the object being scrutinized. Any belief or desire can potentially be the object of self-scrutiny, but when it is the object, it cannot at the same time be the subject.

At the end of his discussion of this issue, Prinz switches gears and suggests that, despite there being no experience of the self-as-subject, we may have reason to think it is there "*by virtue of* its absence." (Prinz 2012, 240) In other words, we know there is a self-as-subject implicitly or inferentially, despite our inability to directly experience it. He gives three reasons to think there is a self-as-subject. First of all, there is "the fact that we always perceive the world from a perspective… conscious states are presented from a point of view." (Prinz 2012, 240) One way to think about how this reveals the self-as-subject is to reflect on Prinz's comment about the self-as-subject being analogous to the 'I' in 'I like this music'. If we always view the world from a perspective, then 'I' can act as an indexical which refers to a particular point of view. Secondly, Prinz claims that the self-as-subject is shown in the boundaries of our experience, that we only ever perceive a portion of our surroundings, e.g. I cannot see what is directly behind me or things far away. Finally, he makes a suggestion that he says echoes Wittgenstein echoing Schopenhauer (who was echoing Kant), that the self-as-subject is inferable from the fact that "the qualities of our experience are dependent on our sensory apparatus… the senses do not simply pick up the world as it is; the

impose order on it" such that "the self is the limit." (Prinz 2012, 240) So the self as subject is the locus of the perspective from which we view the world, which is bounded and imposes a structure on our experience colored by our sensory apparatus, beliefs, desires and values. My only objection to this idea is that I don't see why the self should play this role rather than the person. The line I will take in what follows, drawing from Damasio, Metzinger and Rosenthal's ideas, is that the self is a dynamic representation of the states of the person. By generating thoughts about that representation a person is able to scrutinize it. Those introspective thoughts may become part of the self if they are themselves introspected. However, when they are not themselves scrutinized, those introspective thoughts are not re-represented and therefore not indexed to the self. So to say that they are the self as subject is somewhat unwarranted. Rather, whatever thoughts are scrutinizing a person's thoughts and perceptions at any time are generated by and therefore belong to the person as a whole. It is, therefore, the person that is the subject of experiences, not the self.

Nietzsche writes in *Beyond Good and Evil* §54:

> Formerly, one believed in 'the soul' as one believed in grammar and the grammatical subject: one said, 'I' is the condition, 'think' is the predicate and conditioned – thinking is an activity to which thought *must* supply a subject as cause. Then one tried with admirable perseverance and cunning to get out of this net – and asked whether the opposite might not be the case: 'think' the condition, 'I' the conditioned; 'I' in that case only a synthesis which is *made* by thinking. (Nietzsche 1886, 67)

On my view there are two things 'I' might refer to, neither of which is a soul in the traditional sense: 1. the person, who is the thinker of the thoughts (though not

necessarily a cause or the cause of them) and 2. the self, in the sense I will explicate in the next section, which is a result of, i.e. conditioned by, thinking. The first is the subject that has thoughts and experiences, the second is a thought of and experienced object formed from more basic thoughts and experiences.

III. Self as dynamic internal representation

Damasio (2010) offers a biologically grounded account of the self, conceiving of it as a representation the states of an organism and its relations to the world, which he sees as anchored in William James' views on the subject. According to Damasio, 'selves' are generated by organisms to help regulate their states in response to changing environmental circumstances, distinguish them from external objects and in some cases mark them out as having relatively stable characteristics over time.

Damasio distinguishes between three stages in the development of 'self' demonstrating the differences as well as the continuity between 'selves' of simple organisms and those of human beings. First, there is the 'protoself,' which is "an integrated collection of separate neural patterns that map, moment by moment, the most stable aspects of the organism's physical structure," (Damasio 2011, 190) generating interoceptive 'primordial feelings' that allow the organism to monitor and preserve its well-being through changing environmental conditions. This emphasis on 'feelings' is what Prinz takes issue with, as noted above. However, 'feelings' can be substituted with 'first order states' to yield a version of Damasio's view that does not fall

prey to Prinz's critique. An organism is capable of balancing its internal states in reaction to changes in its surroundings by registering those changes against its represented current state and modifying itself accordingly. In humans this is accomplished via "master interoceptive maps" in the upper brain stem nuclei and insular cortex. (Damasio 2011, 191)

The level of core self introduces a 'protagonist' into the internal representation of the organism in order to more fully represent the distinction between the organism and what it encounters in the world. This level is bit more obscure than the other two, but Damasio claims that there must be "some intermediate self process placed between the protoself and its primordial feelings, on the one hand, and the autobiographical selves that give us our sense of personhood and identity, on the other." (Damasio 2011, 202) The core self state, for Damasio, involves representing objects in the world that are the cause of the protoself's modifications but are distinct from it, resulting in a

> 'feeling of knowing the object,' a feeling that differentiates the object from other objects of the moment… The core self, then, is created by linking the modified protoself to the object that caused the modification, and object that has now been hallmarked by feeling and enhanced by attention. (Damasio 2011, 203)

While I'm not sure Damasio would agree, I think the core self should be understood as not a distinct self from the protoself, but merely one which is more complex and clearly marked off from its surroundings. The core self is generated through 'pulses' of images, telling a non-verbal narrative of the relations between the organism, external objects, and the feelings caused by their interactions. Also, I don't think the core self should be understood as representing objects in the world directly,

but as representing first -order states of the organism which in turn represent the world and the organism's relations to it. I take having a core self, in this sense, to be necessary for being a person.

Metzinger (2004), whose account of self and consciousness is partly influenced by Damasio's work, comes close to this idea, positing a theoretical entity he calls the Phenomenal Self Model (PSM), the content of which is:

> ...your current bodily sensations, your present emotional situation, plus all the contents of your phenomenally experienced cognitive processing… one could even say that you *are* the content of your PSM. All those properties of yourself, to which you can now direct your attention, form the content of your current PSM. Your self-directed thoughts operate on the current contents of your PSM: they *cannot* operate on anything else. When you form thoughts about your "unconscious self" (i.e., the contents of your *mental* self-model), these thoughts are always about a *conscious* representation of this "unconscious self," one that has just been integrated into your currently active PSM. If you want to initiate a goal-directed action aimed at some aspect of yourself -- for example, brushing your hair or shaving yourself -- you need a conscious self-model to *deliberately* initiate these actions. (Metzinger 2004, 299)

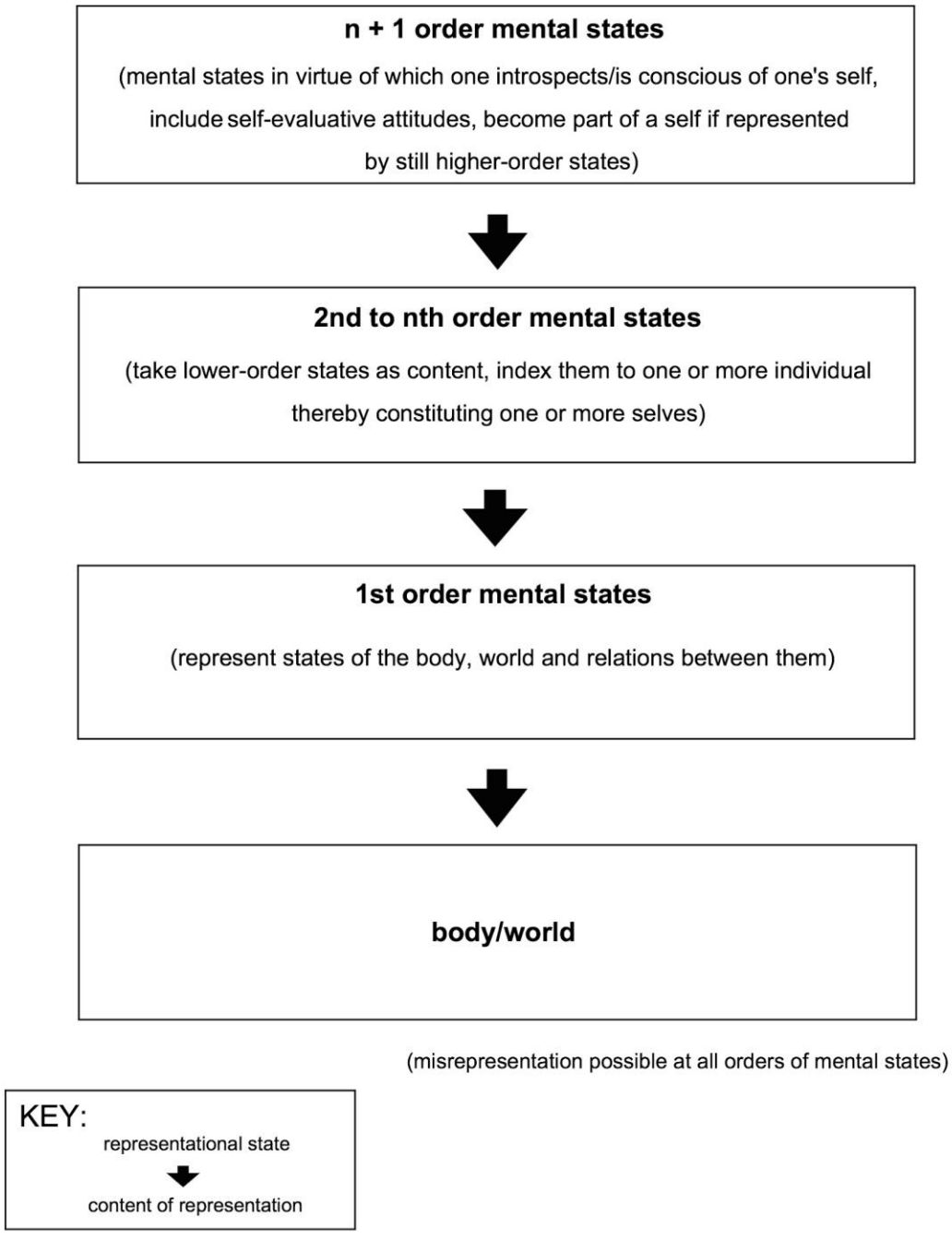
The first order states of you, the person -- affective, intentional and perceptual -- whose contents are states and actions of your body, of the world, and of the relations between your body and the world, are in turn the contents of your self which represents them. Introspection and deliberation involves thinking about one's self and therefore, according to higher order views of consciousness, results in the representations that constitute one's self becoming conscious. (See Figure 1 below) This claim is also compatible with first-order views of consciousness, except, for those views, introspection does not involve thinking about or re-representing the self, but attending to it, making it globally available, etc. In any case, the point is that when one introspects,

one becomes conscious of one's self. I take the difference between persons and animals to be in this ability to consciously scrutinize the self. Having a self of a certain complexity, at least at the level of Damasio's core self or Metzinger's PSM already implies being conscious in the basic phenomenal sense, but the kind of *self*-consciousness that distinguishes persons from other concerned beings requires the even higher-order capacity to introspect the represented features that constitute the self, so as to be able to take evaluative attitudes toward one's conscious desires and beliefs. Beings that lack self-consciousness can have selves, but they can't be conscious of those selves. Among beings with the capacity for concern, only persons can be conscious of their selves.

However, this consciousness need not be infallible, either because one's self inaccurately represents one's first order states or else because one misrepresents oneself in introspection. To begin with, one's self-representation might be incomplete or false. Persons can be largely deluded as to what they are actually like. For example, a person might think him or herself open-minded or charitable and yet really behave narrow-mindedly or miserly. That is not to say that in all or even most cases people do not accurately self-represent, but only that it is sometimes the case. Caruso (2013), drawing on Rosenthal's HOT theory, has argued that it is the incompleteness of one's consciousness of one's own first order states, i.e. the internal causes of one's actions, that leads to the subjective illusion of indeterministic free will. Whether or not such a strong claim is true, a somewhat weaker one, that we are often unconscious of the true

inner causes or motivations for our actions, is plausible and may have some empirical evidence in its favor (e.g. Nisbett and Wilson 1977, Haidt 2001), though such purported evidence is still up for debate. Even if we are usually aware of the causes of our actions, we may be only aware of the proximate causes and not the distal ones, e.g. I may be aware that I bought a particular brand of chewing gum because I prefer that brand, but unaware that my preference was caused by subliminal advertising. It may still be the case that we are mistaken about or unaware of some of our mental states that are not directly related to action. Practitioners and proponents of Buddhist mindfulness meditation claim that such practice can increase the accuracy and completeness of awareness of one's self. Mindfulness practice seems to have had some effectiveness when combined with cognitive behavioral therapy in treating depression and other mood disorders which may result from a negative mis-characterization of one's self (as shown by, e.g., Manicavascar et al. 2011).

Map of Self As Dynamic Internal Representation



[Figure 1]

IV. The narrativity of selves

Damasio describes the process that yields a core self as a non-verbal narrative. According to him, this narrative becomes explicit and conscious in more sophisticated organisms that possess an 'autobiographical self'. With the autobiographical self, an organism is capable of linking together memories into a coherent pattern, generating the 'sense of self' stressed by many thinkers as essential to one's 'identity'. Narrative conceptions of personal identity as well as social constructivist psychological views are often motivated by this autobiographical ability, though as I have already argued, there are persons who lack this ability because of deficits in episodic memory. However, even individuals with Korsakov's disease have some conception of themselves as extended through time and as having relatively stable psychological features. They are just unable to recall past events or anticipate future ones. So I think that while persons must have core selves, they need not have autobiographical ones, they need not construct explicit narratives out of their past and anticipated experiences. However, I do think that already on the level of core self there is necessarily a kind of minimal or implicit narrativity. To make this point clear it will be helpful to engage with some of the literature on the narrative view of personal identity.

A position on the issue of personal identity championed by Schechtman (1996/2007), Dennett (1992) and Velleman (2005), is that it is constituted by a narrative which links successive events of a person's life together into a coherent whole. Taking this as a view of the self, rather than the person, would be to identify the self with the

narrative or to insist that selves are necessarily narrative. By way of example, I will focus on Schechtman's formulation of the view, which she calls the Narrative Self-constitution view. (NSC) According, to Schechtman, in her 1996 book, *The Constitution of Selves*:

> the difference between persons and other individuals...lies in how they organize their experience, and hence their lives. At the core of this view is the assertion that individuals constitute themselves as persons by coming to think of themselves as persisting subjects who have had experience in the past and will continue to have experience in the future, taking certain experiences as theirs. Some, but not all, individuals weave stories of their lives, and it is their doing so which makes them persons. (Schechtman 1996, 94)

So according to this view, to be a person and remain the same person one must think of oneself as persisting over time, and narrativize one's experiences into stories. Schechtman originally presented this view as an alternative to reductive psychological theories of personal identity that understood such identity in terms of relations between the psychological features of momentary person-stages, and which, for her, fail to express "the deep diachronic unity of self-consciousness that is taken to underlie the capacity for forensic actions." (Schechtman 2014, 100) However, Schechtman has distanced herself from that view to some extent, because she thinks it does not properly distinguish between the forensic unit and the moral self, or between what she calls the "re-identification" and "characterization" questions appropriate to those objects of inquiry, respectively. (Schechtman 2014, 103) In other words, it is not clear if NSC is an explanation of the continuity of a person's character (moral self), or of the person itself (forensic unit). It seems to me (and Schechtman seems to more or less agree), as

should be no surprise given my account of persistence in chapter two, that ‘self’ is better suited to explaining characterization rather than re-identification. Even if narrativity does not explain how a person persists, or is to be re-identified over time, as an answer to the characterization question we should consider whether or not it is necessarily true of the selves of persons that they are narratives. In the following discussion, all comments about persons should be applied to consideration of selves.

According to NSC, persons must think of themselves as persisting subjects over time and weave their past experiences and anticipated future ones into coherent stories in order to persist over time. Strawson (1999) asserts that he himself does no such thing. Personally, I do tend to think of my life in terms of a narrative, but this may be somewhat of a delusion and I certainly am in no position to make any claims about another person’s mental life on that score. Individuals with Korsakov’s disease seem clearly to be persons and have selves in the sense I have been discussing in the foregoing, and yet do not have the access to past and anticipated experiences in order to construct stories of their lives (or any stories at all). Furthermore, some persons may have fragmented or discontinuous self-narratives as do those who experience frequent fugue states.

Schechtman does, however, have a response to these sorts of objections. She denies that one’s narrativizing need be conscious or explicit. “...’having an autobiographical narrative’”, she writes, “does not amount to consciously retelling one’s life story always (or ever) to oneself or anyone else. The sense in which we have

autobiographical narratives on this account is cashed out mostly in terms of the way in which an implicit understanding of the ongoing course of our lives influences our experience and deliberation." (Schechtman 2014, 101) Now, it's not entirely clear to me what such unconscious or implicit narrativizing might amount to. If it is just that one's present thoughts and actions have an effect on the future and one sees oneself as being formed by what has happened in the past, then that does seem necessary for personhood insofar as it is necessary for self-consciousness and concern, and that is a kind of narrativity of which Korsakov's patients do seem capable. They can remember facts and link them together in logical sequences, though they are unable to describe a scene as if they were present for it in the way necessary for explicit storytelling.

Another sense in which one's self might be implicitly or unconsciously narrative is if it is formed not by the person who has it but by other people. Schechtman considers this idea when she offers an expanded version of the NSC on the way to developing her more recent position on personal identity, the "Person Life View" (PLV). Now clearly, the way I interact with other people and the ways I perceive their perceptions and judgments of me contributes to my self, but I don't see how the way they understand my life narratively should *constitute* my self-narrative. I might not care at all how others think about me and fail to take their stories into account, even implicitly and unconsciously. All in all, the selves of persons are necessarily narrative in only the minimal, implicit sense, noted above. However, I do think it is true that only (but not all)

persons are capable of narrativizing in the more robust, explicit sense and this is why it is characteristic, though not constitutive, of personal selves that they are narrative.

V. Ontology, unity, and stability of selves.

Strawson (1999) offers an analysis of the concept of the self which includes various suggestions about what characteristics belong to it. One issue is about the ontology of selves, specifically, where Strawson claims, rather vaguely, that the self must be understood as a 'thing.' Now there is an obvious and uninteresting way in which it must be true that the self is a 'thing', in the way that everything with a name is a 'thing', but Strawson seems to have something more substantial in mind. He says "the self is not thought of as a state or property of something else, or as an event, or as a mere process or series of events," though not "a thing in the way that a stone or a chair is," but "has the typical causal profile of a thing - as something that can *undergo* things and *do* things… a 'thinking active principle' [in Bishop Berkeley's sense]" (Strawson 1999, 132) From this description, it seems that by 'thing', Strawson means an object, though not a physical object. I'm not sure I agree that objects, as opposed to states, properties and processes, are the only things that can undergo or do things (and in the final analysis I think at the least 'mental state,' 'mental property' and 'mental object' talk is interchangeable), but regardless, I understand the self to be a kind of mental object, i.e. a representation The self is a representation of one's first order states and therefore, a mental object. Mental objects can do things insofar as they play causal roles in an

individual's behavior, e.g. different beliefs about my self may cause me to pursue different career paths, and they can undergo things insofar as they can be re-represented and scrutinized by even higher-order representations in introspection and sometimes may even be altered in light of such scrutiny. The self as mental object is also reducible to physical states or processes, so it is both an object and a set of states or processes. Selves are mental, but, given psychophysical reductionism, they are also physical.

Two characteristics that I take to be, in some sense, essential to selves are stability and unity. By stability, I mean that selves do not change radically over time. Historically, most views of the self have insisted that it must be completely or largely immutable. By unity I mean that the self is experienced by a being who has one and is capable of being aware of having one as a unified whole, at every moment, belonging to a single individual. Selves are individuated only by a person's experience of their sameness and difference, by whom they represent their intentional states as belonging to. Radical instability and disunity of self imply multiplicity of selves though not multiplicity of persons. A consequence of this view is that the correct description of cases of dissociative identity disorder, and perhaps other similar conditions, is to say that in such cases a single person has multiple selves.

There is seeming synchronic unity and diachronic stability in self-experience, while at the same time, there is some evidence that suggests plurality, divisibility or instability. While self as object, like conscious experience, does usually seem

synchronically unified, it may also be experienced as divided, for instance, in conflicts of motives or crises of identity. Diachronic stability of self as object seems to be one of the characteristic features of selves, and the reason why they are often thought to play an essential role in personal persistence. The idea of self is partially the idea of something relatively enduring. However, just as it may become synchronically disunified, one's self may also, become diachronically destabilized, as one feels oneself to be (or have) "a million different people [read: selves] from one day to the next" (as expressed in the song "Bittersweet Symphony" by the Verve (1996)). Strawson (1999) thinks diachronic unity is not a necessary feature of selves, and suggests that they may actually be very short lived and fractured.

The self-representation of a person usually constitutes, for that person, a stable 'identity' constructed out of the person's self-perceived personality traits, affiliations, and preferences. This leads to the idea that personal identity consists in such stability. As has been seen in Korsakov's patients like KC, such a sense of stability is independent of the capacity for episodic memory or future-directed mental time travel. However, even in typical cases, this sense of stability may be largely an illusion as regards the actual characteristics of a person. Social-psychological studies related to the situationist critique of the idea of virtue or stable character traits (e.g. Isen and Levin 1972/1975, Harman 1999/2000, Doris 1998/2002) purport to show that people's behavior displays less regularity than they believe it should, based on their own self-concept. The jury is still out on how serious a threat to idea of stability of character these studies truly pose.

However, even if our character traits aren't genuinely stable, the subjective sense of stability may be essential to the self and this may reflect the way we as organisms must maintain stability on a subpersonal level regarding homeostasis and other processes involved in life regulation. In Damasio's terms, the sense of stability that comes with our awareness of core self states may be parasitic on the conditions regulated by the protoself.

Theories of personal identity that appeal to stability of traits include social psychological accounts that take the relations between individuals and their families, ethnicities, and other social groups to be partially constitutive of who they are. For instance, according to Greenwood (1994) a person's identity is constituted by her "identity projects" or the "moral careers" upon which she embarks. He writes:

> Theoretical descriptions of identity projects are theoretical descriptions employed in the explanation of intentional human behavior, or human actions. According to this form of social psychological theory of identity, a theoretical reference to intrinsically social identity projects -- and their associated emotions, and motives -- provides the *best* explanation of many human actions, and of the similarities and differences in actions to be found between different persons, and the same person in different times and places. Thus for example similarities and differences in the preparation and performance of high school and college students may be explained in terms of differential levels of commitment to the moral career of academia. The disruptive activities of some high school students may best be explained in terms of their commitment to alternative moral careers, such as those provided by teenage gangs. (Greenwood 1994, 112)

I grant that commitment to different identity projects can play the explanatory role that Greenwood thinks it can. On the other hand, I think someone can remain the same person despite radically altering her identity projects. For example, the kid who has had enough of the thug life and decides to give school another shot is the same person who

was previously committed to her gang. As far as the self goes, it does seem as if, subjectively, persons think of themselves largely in moral and social terms. The colloquial expression “I was a different person then” which I think should be revised as “I had a different self then” seems especially appropriate when it is a change of morality or social membership that has provoked it. “I’m not [don’t have] my self today” might also be especially apt if one acts in a way that conflicts with one’s usual moral beliefs. I earlier cited the Nichols and Strohminger (2014) study which showed that participants were more likely to think that a change in identity occurred when an individual’s moral character radically changed than if their memories or their desires and preferences were eliminated, or if they suffered from visual object agnosia. That provides evidence that one’s self is largely or often constituted by one’s moral beliefs. If those beliefs are dependent on one’s social group membership, then such membership is transitively constitutive of the self and a radical change in group memberships might entail a quantitatively different self.

Wilkes (1988) has argued that in conditions such as hypnotism, fugue states, and bouts of epileptic automatism, where individuals behave radically uncharacteristically for periods of time, usually without remembering the period in question, we do not think of them as being different persons during that period, and I agree, but suggest that we should think of them as possessing different selves. Similarly, in cases where treatable neurological ailments cause radical changes in personality, such as the case of Mary Jackson, a patient of neurologist Kenneth Heilman, whose sudden shift from a

monogamous honors student to a promiscuous crack smoker was explained when a tumor was found pressing on her prefrontal cortex (Heilman, 2002), we might want to say that the person's normal self is temporarily replaced by a different one. However it isn't entirely clear that we should say so. Individuation of selves has to do with the degree to which a person thinks of those selves as individuated, and only in extreme cases, where a person genuinely and consistently believes there is more than one self within them and this belief is reflected in behavior and physiology, do I think we should say a multiplicity of selves is present.

An issue at the crossroads of consciousness studies and the topic of the self is the phenomenal unity of experience. Conscious experience seems to be of a seamlessly unified field, our various sensory modalities, thoughts and feelings seem tied together at each moment into a coherent whole, and there are several theories that purport to explain why that is. One view, popular among neuroscience-minded philosophers, is that conscious states are unified due to the synchrony of synaptic firings underlying them, particularly in the 40hz range. The view has been endorsed by Prinz (2012) among others, and has been used to explain not only the unity of consciousness, but also the more basic concept of a brain state (Brown 2006 and 2013). However, while giving an account of the neural correlates of unity, this leaves the issue on the cognitive level of explanation, of exactly how synchrony yields unity, unanswered. A view which does not appeal directly to neurology but which assumes potential reducibility, and which pitches its explanation on the cognitive level, is

championed by Rosenthal (2003), who draws upon the resources of his HOT theory to explain the unity of consciousness. According to his account, higher order representations of mental states include an “essential indexical” which tags the represented states as belonging to the same individual. As Rosenthal puts it:

> each HOT makes one conscious of oneself in a seemingly immediate way, encouraging a sense of unity across HOTs. And the same considerations that make us assume that our first-person thoughts all refer to the same self apply also to HOTs; becoming conscious of our HOTs in introspection thus leads to a sense that our conscious states are unified in a single self. (Rosenthal 2003, 325)

This view fits very nicely with what I have said about the self so far and shows why such a self is necessarily unified. Since having a self is entirely a matter of how one represents one’s own states, then if those states are represented as belonging to distinct individuals, experience would not only be disunified, but there would in fact be separate selves to which the differently indexed states would belong. There would not be more than one person, but subjectively it would appear that way to the person whose experience is disunified. Common experiences of divided attention and conflicting motivations, which may be interpreted as examples of disunity do not involve the states in question as being represented as belonging to distinct subjects and so are not genuine examples. On the other hand, the voices heard by schizophrenics, alien hand syndrome, “hidden observers” in hypnosis, seem to be true disunities, and therefore involve a plurality of selves. As a description of such phenomena, my account of the constitution of selves may seem too voluntaristic. Individuals with DID do not seem to be in conscious control of the number of selves they have. However, my account should

not be construed as voluntaristic. Selves are subjectively constituted, however, such constitution is not always in the control of the person who has them. When one becomes conscious of one's first order mental states, the first order states are represented as belonging to an individual. However, that indexing does not result from consciousness but only occurs as part of it, making it appear to the person that he or she is composed of one or more individuals.

Given the distinction between persons and selves developed in this chapter, I propose that we describe disunities (and discontinuities or instabilities) of self as a single person having more than one self at a time or over time. This phenomenon is most apparent in cases of dissociative identity disorder. Wilkes (1988) describes a number of cases which she thinks call into question the "Lockean" criteria, for being an individual person, of unity and continuity of consciousness. The most extreme of these cases, and the only ones she thinks really compel us to revise our concept of an individual person, are cases of multiple personality disorder, which is now referred to by clinicians as dissociative identity disorder (DID). One case Wilkes describes is of a patient known as "Christine Beauchamp", who appeared to fragment into multiple personalities that vied for control over "Christine's" body. Some of these alters had knowledge of the actions of other alters, and some had control over others' thoughts. Some alters would be present, hidden observers, "listening" in on what was being thought said and done when another alter was in control. In such a case, some mental states are being attributed to one subject while others are attributed to a different one.

For Wilkes, this means that there is reason to say more than one person is contained in the body or else we must revise our concept of an individual person. I take the latter option, that we should revise the concept of an individual person so that it does not require unity and continuity/stability (or recognize that those conditions were not part of the concept in the first place) in the respects that a DID patient lacks them. I think we should say that in such cases there is more than one self present, because selves do require unity and continuity/stability, though those things are constituted by their being perceived as such by the person. There is, in such cases, however, only one person, because there is only one self-consciously concerned being - only one organized structure realizing those capacities, one agent representing the mental states as belonging to multiple individuals. Therefore, when deciding responsibility for an action, a DID patient is a single person who may or may not be responsible for some actions. However, all of his selves would have to be taken into account both when deciding whether or not he is responsible for the particular action as well as the degree of praise or blame/reward or punishment appropriate.

Returning to the case of Mary Jackson, though her character and behavior changed radically due to her tumour, she did not represent her thoughts and actions as belonging to a distinct individual. For that reason I would say that she had a single self from before during and after the period when she had the tumour, though it changed qualitatively quite a bit. One might object at this point, that individuating selves in this way makes it so that the concept of ‘self’ doesn’t do the work I mean for it to do in

interpreting phrases like “I am (have) a different person (self) than I used to be (have),” since such a phrase is not always uttered and believed by someone with a pathology such as DID or schizophrenia. However, even in non-pathological cases, the phrase is sincerely uttered when someone thinks that she has changed so radically that her current thoughts, actions and experiences are not attributable to the individual she was in the past. So in such a case I think it is appropriate to say the person has a quantitatively different self, as their current features are indexed to a different individual than their former features were. Again, this criterion is genuinely subjective, so if Mary genuinely feels that her thoughts, actions and experiences during her tumour period really belonged to someone else, then she did have a different self during that period. If she believes she was a different person than she is mistaken about that, though if she were to come to understand the distinction I have drawn between persons and selves she might come to agree that she was not actually a different person but merely had a different self.

Chapter 5: Metaphysical and Moral Personhood

I began this study by explaining that the concept of a person has been considered not only to be a metaphysical concept, but one with ethical or moral one connotations as well, and there has been some doubt cast by various writers, including Dennett (1978) and Chappell (2011) on the notion that one can fully disentangle the metaphysical criteria from the ethical and moral connotations. Indeed, to some extent it follows from my account of persons that one cannot fully do so because of the forensic implications of 'person', which are at the same time its principal utility as a term. The set of capacities, self-consciousness and concern, possessed jointly by all and only persons, is also the set of capacities required for someone to be responsible for his or her actions such that it would be reasonable to put that individual on trial or have him or her offer testimony, sign a contractual agreement, or give informed consent, provided the person is able to communicate with a sufficient number of the other persons involved in a given case. However, despite being bound up with the notion of responsibility, the metaphysical notion of a person as I have presented it implies nothing about the moral goodness of persons, i.e. that they are concerned for others or otherwise behave in morally obligatory or praiseworthy ways. Furthermore, establishing whether a being is or isn't a person cannot by itself settle bioethical debates about euthanasia, abortion, and animal cruelty, where the question of the rights of individuals, their value as beings, and the reality or significance of their suffering is at issue.

In this chapter, I argue that the metaphysical concept of a person has no moral implications, beyond the issue of responsibility, that follow from it uniquely. I will begin by arguing that nothing about the metaphysical concept of a person or a responsible agent grounds any ideal of moral behavior, because while responsibility, and therefore personhood, requires concern, it does not require that one is concerned for the interests of others, so that even complete psychopaths may be persons. Then I will turn to the role of personhood in bioethics, appealing to writers in that field who have offered reasons for thinking that the concept of a person is unhelpful for resolving bioethical debates and offering my own argument for that claim, to the effect that the set of paradigmatic person constituting capacities are not the same as the capacities morally relevant to those debates and that unlike responsibility, which is a matter of metaphysical fact, rights are social constructs that are conferred upon persons and other beings by persons and so do not have definite metaphysically grounded conditions of application. Finally, I will argue, following Singer (1975/2002), that while our judgments of whose interests are morally significant are partially grounded in the capacity for concern, because being a concerned being is necessary for having interests in the first place, judgments of moral significance are not similarly grounded by other capacities such as self-consciousness and responsibility.

I. Personhood and concern for others

According to the view developed in chapter one, to be a person one must have the capacity for concern, which I define as affective investment in the satisfaction of one's goals (which is what makes a goal a genuine desire) and the truths of one's beliefs, at least insofar as they are related to the satisfaction of one's desires. This is the case, because without concern one cannot fully appreciate the consequences of one's actions and therefore cannot be responsible for any of those actions. One's concerns may be self-directed or directed towards others, in the sense that one may be emotionally invested in the satisfaction of one's own desires or in the satisfaction of someone else's desires. A psychological egoist would claim, however, that all concern is ultimately self-directed. This is not the place to argue that point, but, as I mentioned in chapter two, one can in the present be concerned that things will be one way and not another in the future long after one ceases to exist, so that one's concerns need not be self-directed in the sense of pertaining to one's own anticipated experiences. I think that they need never be so, and in that sense a person might be entirely altruistic. On the other hand, I hold that a person needn't ever be concerned for the satisfaction of the desires of anyone other than him or herself and in this way I depart from many other writers on the subject of personhood. As I see it, normal human beings, the paradigm examples of personhood, as well as many other creatures[33], admit of varying degrees of

[33] Peterson 2011 relates evidence of varying levels of empathy and altruism within different animal species. For instance: "Fifteen rhesus monkeys were taught to acquire food treats by pulling on one of two chains, either of which would deliver the same amount of food….

concern for others. Paragons of altruism, such as Siddhartha Gautama, Jesus and more recently Mother Teresa are at one end of the spectrum, overflowing with their concern for others seemingly without limit. In the middle are people like myself, and probably my reader, who are deeply concerned for a few others, particularly those people one regards as close friends and family members, and have some minimal concern for all humans and perhaps some other sentient beings, but also a substantial degree of selfishness that limits our concern for others. Loving someone might be understood as being as concerned for that individual as much as (or more than) you are for yourself, so that most of us can love one or a few other individuals, but not everyone.

On the opposite end of the spectrum from the saints are the psychopaths, who seem to have no concern for others. The term 'psychopath' and the closely related 'sociopath' have been dropped from recent editions of the DSM, and the symptoms

---

whenever the experimental monkeys pulled one of the two chains [a third monkey within their view] would receive a shock… two thirds of the monkeys quickly developed a significant preference for pulling on the chain that did not shock the other monkey, and of the one third of the group who showed no preference, two actually stopped pulling the chain altogether... [choosing] genuine hunger and even the possibility of starvation rather than cause pain to a fellow monkey." (Peterson 2011, 229) Also: "One day in late June 2000, a young African forest elephant weak from malnutrition, collapsed off to one side of a narrow, sandy trail in a Central African forest… within a few hours died… During... [the following] two days, then, elephants walking along the sandy trail made 129 visits to a fellow elephant in trouble… About 50 percent of them reacted as you might expect: They showed signs of fear and avoidance… One exceptional individual, known as Miss Lonelyheart, visited several times on the second day and reacted aggressively to the body, stabbing it with her tusks and attempting to tear pieces away from it. Miss Lonelyheart was already well-known [to the observing scientists] as a social misfit, and her bizarre behavior was not out of character… the elephants identified in the other half of the sample… included many instances of socially positive reactions to the drama of another elephant in trouble. Some 15 percent of the total visits during those two days involved protective behavior: the visitor seeming to protect or guard the body from others. And in about 18 percent of the cases, the visiting elephants looked as if they were trying to assist or revive the dying elephant, mostly by attempting to push or lift her upright, using their feet, trunks or tusks." (Peterson 2011, 217-218)

associated with such labels are now included under the larger umbrella of ‘antisocial personality disorder’. However ‘psychopath’ still looms large in the psychological literature, so I will continue to use it here. Hare (2003) offers a checklist of psychopathic personality and lifestyle traits which includes the following: glib and superficial charm, grandiose self-worth, need for stimulation or proneness to boredom, pathological lying, cunning and manipulativeness, lack of remorse or guilt, shallow affect, callousness and lack of empathy, parasitic lifestyle, poor behavioral controls, promiscuous sexual behavior, early behavior problems, lack of realistic long term goals, impulsivity, failure to accept responsibility for actions (which implies that they are in fact sometimes responsible), many short-term marital relationships.

Psychopaths are primarily characterized by their lack of remorse and failure to govern their actions according to moral rules. Though long considered a psychological disorder, psychopaths are often very high functioning and successful by many measures. In response to a casual statement by Hare that ‘‘Not all psychopaths are in prison. Some are in the Boardroom”’ (Hare, 2002). Babiak et. al (2010) ran studies whose “results provide evidence that a high level of psychopathic traits does not necessarily impede progress and advancement in corporate organizations.” (Babiak et. al, 192) Therefore, psychopathy does not fit into the general definition of a psychological disorder as something that impedes an individual from achieving her goals or living her life as she chooses. Nevertheless, psychopathy is seen as pathological because of the harmful effects the psychopath’s behavior has on the rest of society. There is

considerable debate over just how to understand the underlying psychological defects that lead to the psychopathic personality and lifestyle. Some comparison has been made with another psychiatric condition, autism, which has been thought to be primarily constituted by a deficit in mindreading or Theory of Mind (ToM), i.e. the ability to infer the beliefs, desires and other intentional states of other people from their behavior. (E.g. Baron-Cohen 2002) However, recent work has shown that psychopaths do not characteristically display deficits in ToM. (Richell et al. 2003) Other research has distinguished cognitive ToM from affective ToM (i.e. empathy) and found that psychopaths are deficient in the latter but not the former. (James et al. 1997 and Shamay-Tsoory et al 2008) The idea is that psychopaths have no problem interpreting the behavior of others and assigning intentional states to them, including negative affective states of suffering and distress - they are actually quite good at that, and use that understanding to manipulate people for their own purposes - they are just not emotionally moved by the bad feelings of others, i.e. they don't feel bad that other people or animals are suffering.

This explanation of psychopathy, that it is due to emotional, not cognitive or intellectual deficits is complicatedly related to the classic 'moral insanity' diagnosis (Prichard 1835), whereby psychopaths lack the capacity to recognize the difference between right and wrong. This complication cuts to the heart of moral philosophy, the question of whether or not one can truly appreciate the difference between right and wrong, yet not be moved to do what is right, without suffering from any weakness of will.

Psychopaths do seem to appreciate that there are moral rules, what counts as following them, and that other people take themselves to be bound by them. According to the findings of Cima et al. (2010) and Koenigs et al. (2012) psychopaths do not differ from non-psychopaths in their judgments about various moral dilemmas. Cima et al. (2010) conclude from this that psychopaths understand what counts as morally right or wrong -- ordinary emotional reactions to morally salient scenarios are not necessary for such judgments -- but that they don't care about whether or not their own actions are right or wrong. One aspect of morality that they do fail to appreciate is the difference between moral and conventional rules. In one respect they see all rules governing behavior as moral rules, because they are serious and authority independent, however, they don't take themselves to be bound by any of those rules, seeing them as mere expressions of particular values which they need not share, and in that respect treat them as conventional.

Consider the following quote from the paradigmatic psychopath, Ted Bundy, who recorded these comments about committing the rape and murder of a woman:

> Then I learned that all moral judgments are 'value judgments,' that all value judgments are subjective, and that none can be proved to be either 'right' or 'wrong'. Believe it or not but I figured out that for myself that if the rationality of one moral judgment was zero, multiplying it by millions would not make it one whit more rational. Nor is there any 'reason' to obey the law for anyone, like myself, who has the boldness and daring – the strength of character – to throw off its shackles. I discovered that to become truly free, truly unfettered, I had to become truly uninhibited. And I quickly discovered that the greatest obstacle to my freedom, the greatest block and limitation to it, consists in the insupportable 'value judgment' that I was bound to respect the rights of others. I asked myself, who were these 'others'? Why is it more wrong to kill a human animal than any other animal, a pig or a sheep or a steer? Is your life more to you than a hog's life

> to a hog? Why should I be willing to sacrifice my pleasure more for the one than for the other? Surely, you would not, in this age of scientific enlightenment, declare that God or nature has marked some pleasures as 'moral' or 'good' and others as 'immoral' or 'bad'? In any case, let me assure you, my dear lady, that there is absolutely no comparison between the pleasure I might take in eating ham and the pleasure I anticipate in raping and murdering you. That is the honest conclusion to which my education has led me – after the most conscientious examination of my spontaneous and uninhibited self. (quoted in Michaud and Aynesworth 1989)

There doesn't seem to be any cognitive impairment in Bundy's reasoning. His view of morality is not so different from that of an emotivist like Charles Stevenson, who takes moral judgments to be mere expressions of feelings, rather than claims about the world that may or may not be confirmed or disconfirmed by facts. (Stevenson 1937) The difference between them is that, unlike Stevenson, I suppose, Bundy just doesn't seem to feel bad about those he is harming (nor would he for harming a hog or a steer, though his reasoning would seem equally to imply that he should be a vegetarian as it would that he may rape and murder people). One can understand moral rules as non-cognitive, emotive value judgments or as social conventions, but think that they are very special value judgments or conventions because of the severe consequences of their violation, consequences that would cause emotional distress in most people. Bundy's attitude towards his victims is then best understood as a deficit in empathic distress, not a deficit in cognitive abilities. He appreciates the facts of the situation that people take to be morally relevant, but he does not respond to them emotionally the way most people do. This is due both to a lack of empathy and a lack of the fear response most people would have when considering how one would feel if roles were reversed and someone

behaved in this way toward oneself. Bundy sees that attitude as a type of courage and therefore lauds himself for it, contemptuous for those whose weakness binds them to morality.

However, despite this lack of concern for others and deficit in some emotions such as guilt, shame and fear, psychopaths like Bundy do have concern, in the sense that they care that their own desires be satisfied. They do experience many emotions, such as anger, frustration, joy, and amusement. Furthermore, as the Bundy quote shows, they are fully self-conscious, appear to appreciate their own desires and affirm them as desires they are pleased to have. They proudly claim ownership of their desires as their own. Therefore, I see no reason to think they are not responsible for their actions or lack personhood. If anyone is responsible for doing something that is wrong, it is the unconflicted psychopath and not individuals who suffer from weakness of will or ignorance. Psychopaths are persons, just extremely bad ones. Intuitively, when a self-consciously concerned individual does things that we disapprove of, and does not display any inner conflict, but seems perfectly content to do those things, we do not think their personhood or responsibility diminished, but only think of that individual as a bad person responsible for bad things. That concern, but not concern for others, is necessary for personhood, explains and justifies the intuition that psychopaths are responsible for their actions and are, therefore, persons. Concern for others is unequally distributed among persons, as it is within many animal species, so that being a person

or non-person says nothing about the extent to which one may exhibit positive moral behavior.

II. Irrationality, insanity and immorality

Some philosophers offer arguments for the claim that behaving immorally entails a lack of rationality or sanity, and therefore a lack of responsibility or personhood, because, they claim, there is normativity bound up in those very notions. In chapter one, I engaged with Dennett's "Conditions of Personhood" (1978), examining his six 'themes', basically agreeing with him (with a few caveats) that they, particularly the sixth theme of 'self-consciousness', are necessary for personhood, though arguing that he left out one other essential condition, that of 'concern'. Dennett himself is not satisfied that his list, which he takes to be a more or less complete analysis of the metaphysical features associated with the notion of a person, gives us exclusive and exhaustive criteria for 'moral personhood', but not because any further criteria were left off his list. His remarks are rather enigmatic, so I here quote them at length before offering analysis:

> Now, finally, why are we not in a position to claim that these necessary conditions of moral personhood are also sufficient? Simply because the concept of a person is, I have tried to show, inescapably normative. Human beings or other entities can only aspire to being approximations of the ideal, and there can be no way to set a "passing grade" that is not arbitrary. Were the six conditions (strictly interpreted) considered sufficient they would not ensure that any actual entity was a person, for nothing would ever fulfill them. The moral notion of a person and the metaphysical notion of a person are not separate and distinct concepts but just two different and unstable resting points on the same continuum. This relativity infects the satisfaction of conditions of personhood at

> every level. There is no objectively satisfiable sufficient condition for an entity's *really* having beliefs, and as we uncover apparent irrationality under an intentional interpretation of an entity, our grounds for ascribing any beliefs at all wanes, especially when we have (what we always *can* have in principle) a non-intentional, mechanistic account of the entity. In just the same way our assumption that an entity is a person is shaken precisely in those cases where it matters: when wrong has been done and the question of responsibility arises. For in these cases the grounds for saying that the person is culpable (the evidence that he did wrong, was aware he was doing wrong, and did wrong of his own free will) are in themselves grounds for doubting that it is a person we are dealing with at all. And if it is asked what could *settle* our doubts, the answer is: nothing. When such problems arise we cannot even tell in our own cases if we are persons. (Dennett 1978, 193-4)

Dennett says that the metaphysical conditions he has claimed are necessary for being a person cannot be sufficient for being a person in the full moral sense, but I think this way of putting the point is confused. Here is what I think he is really getting at: He is suggesting that whenever a person acts immorally, the rationality of their behavior is called into doubt. Because rationality is a necessary condition of personhood, we have reason to doubt that anyone who acts immorally really does meet the criteria for personhood. Since this doubt arises in just the cases where the concept of moral personhood is important, i.e. when someone has committed an immoral action and we want to know whether or not he or she is responsible for doing so, we are left with a paradox. An individual who self-consciously performs an immoral action acts irrationally and therefore does not act intentionally in the way necessary for being self-conscious. Therefore, what follows from Dennett's reasoning isn't really about the sufficiency of his conditions, but instead, that there are no metaphysical conditions that can be seen as *necessary* for moral personhood. If rationality is necessary for intentionality,

intentionality necessary for self-consciousness, self-consciousness necessary for responsibility and responsibility necessary for personhood, and to act immorally is irrational, then anyone who acts immorally is, in respect of that action, not a person. If 'person' is essentially normative, implying a moral ideal of good conduct, then almost no one really is or has ever been a person.

Citing Locke's passage about the forensic nature of 'person', Dennett explains the moral notion of a person as that of a being "who is accountable, who has both rights and responsibilities," (Dennett 1978, 176) and then goes on to wonder whether being a self-conscious being (or regarded as one) is equivalent to being an end-in-oneself or merely a precondition of being one. However, this characterization goes beyond Locke's forensic use of person in a way that muddles the issue. Being responsible in a sense that is more or less synonymous with 'accountable' is central to personhood, however, as I have argued above, being an individual with rights or particular 'responsibilities' (i.e. duties or obligations)[34] or being and end-in-oneself does not go hand in hand with responsibility/accountability. So Dennett's target in attempting to fit the metaphysical conditions of personhood to the moral notion is broader and more heterogeneous than he thinks. 'Moral personhood', for Dennett consists not just in the capacity for acting in such a way that one is responsible for those actions, but also being the bearer of 'most rights' (which ones he does not enumerate) and of having and recognizing duties to others, i.e. treating them as ends-in-themselves. Given such a maximal conception of

[34] One can be responsible for one's actions even if there are no particular things that one is obligated to do.

what it takes to be a moral person, Dennett is right to think that most, perhaps all of us, fall short of the mark. However, where I disagree with Dennett is about his claim that the normativity associated with the moral notion of personhood is built into the metaphysical conditions themselves, such that one can't even claim with any assurance that any individual is a genuine person in even the metaphysical sense of the term.

Dennett's claim, that an individual's rationality may be called into doubt whenever that individual does something morally wrong, is dubious. That would only be the case if it is always or at least usually irrational to do things that are morally wrong. To assume that is so is to endorse a conception of morality where doing the right thing is entirely a matter of grasping truths about what constitutes right or wrong action, the principal truth being either, following Kant, that one should treat others as ends in themselves and never as means to one's own ends; or, following the utilitarians, that one should act to secure the greatest possible benefit and least harm, taking the interests of all morally relevant beings into account. One could argue in favor of either principle by appealing to the fact that there are no ethically relevant differences between one person and another, so that one has no good reason to privilege one's own interests over those of another. However, it's not clear that there are decisive reasons for *not* privileging one's own desires over those of others.

Rawls' ethical theory attempts to provide some such reasons, supplementing Kant's or the utilitarian's principles (though not necessarily in such a way that Kant or the utilitarians would have approved) by providing a justification for taking the interests

of everyone else into account. If one had no idea what station in life one would end up in, one would choose the most equitable distribution of benefit possible. Granting Rawls' view that in the "original position" where individuals are asked to make judgements about the normative principles that should govern society, an ideally rational agent would opt for justice (as Rawls conceives of it), Dennett claims that "just as part of our warrant for considering an entity to have any beliefs or other Intentions is our ability to construe the entity as *rational*, so our grounds for considering an entity a person include our ability to view him as abiding by the principles of justice." (Dennett 1978, 190) However that would only be true if we were in the original position so that the most rationally self-interested choice (assuming that the kind of rationality necessary for being ascribed intentions really does require one act to maximize self-interest in the first place) were the choice that is also the most just. As it is we do not in fact make our decisions behind a veil of ignorance, so it is not clear that it is most rational to care what happens to anyone besides oneself. If someone steals or kills for profit, it is clear that the person is rational in both the minimal sense implied by goal-directedness that Dennett takes to be the ground for thinking of something as an intentional system and in the sense that the person is acting in her own self-interest. There need be no fault in that individual's ability to reason in either sense. Only if one's goal is to be egalitarian or altruistic in one's actions, would selfish behavior be a sign of irrationality, though even then it might also be explained by weakness of will.

A related challenge to the claim that persons needn't act morally comes from Susan Wolf's criterion of *sanity* for responsibility. Wolf develops her criterion as a supplement to what she calls "deep-self" views of responsibility, such as Frankfurt's, but also those of Watson (1975) and Taylor (1976). What she takes to be common to these views is that they

> share the idea that responsible agency involves something more than intentional agency. All agree that if we are responsible agents, it is not just because our actions are in control of our wills, but because, in addition, our wills are not just psychological states *in* us, but expressions of characters that come *from* us, or at any rate that are acknowledged and affirmed *by* us… In one way or another, all these philosophers seem to be saying that the key to responsibility lies in the fact that responsible agents are those for whom it is not just the case that their actions are in the control of their wills, but also the case that their wills are in the control of their *selves*, in some deeper sense. (Wolf 1987, 49)

I have appealed to the kind of view Wolf is talking about in my account of personhood. Wolf agrees that such a view distinguishes responsible agents from, on the one hand, kleptomaniacs and other compulsives, who reject their desires or see them as alien to themselves, so that their first order desires and the actions that result from them are at odds with their deep selves (given my account of the self in the previous chapter, it seems to me that the adjective 'deep' doesn't add much, but I'll use it for now) and on the other hand, from animals or other intentional agents who cannot be responsible for their actions in the way that persons can because they lack deep selves altogether. I'm more interested in the second distinction than the first because my view is about the conditions necessary and sufficient to be a responsible agent in general,

not those for being responsible for some particular action. A kleptomaniac may be responsible for some other actions even if he is not responsible for stealing.

However, Wolf thinks that having the capacity for second order volitions or any other kind of deep self is not sufficient for being a responsible agent, because one might still lack a further necessary condition of responsibility, which she calls ‘sanity’. She introduces this condition by way of illustration through the story of ‘Jojo’, raised by a sadistic dictator of a father who teaches Jojo to delight in hurting people. Jojo learns to fully identify with the sadistic desires he develops as a result of this upbringing in the way Frankfurt and the others think is sufficient for responsibility. However, for Wolf, it is clear that Jojo is not a responsible agent, not because he lacks some kind of self-control over his actions, but because Jojo is not sane, his values are not “controlled by processes that afford an accurate conception of the world.” (Wolf 1987, 55) Now, keeping in mind my distinction between being responsible for some action vs. being a responsible agent in general, I’m not sure if Wolf would deny that Jojo meets the latter condition and is therefore, on my view not a person at all (Wolf herself does not seem to agree that being a responsible agent goes hand in hand with being a person). However, Wolf explains the scenario as if *all* of Jojo’s actions follow from the wicked nature cultivated in him by his father, so it would follow that, on Wolf’s account, Jojo cannot be responsible for any actions and therefore, given my account of the criteria for being a person, lacks personhood.

Wolf's definition of sanity is taken from the legal domain, specifically the M'Naghten rule or 'right-wrong' test, according to which "a person is sane if: (1) he knows what he is doing and (2) he knows that what he is doing, as the case may be, is right or wrong," the second of which requires, for Wolf, "the minimally sufficient ability, cognitively and normatively to recognize and appreciate the world for what it is." (Wolf 1987, 55) On her view an immoral individual such as Jojo counts as insane because such an individual violates the second of the two conditions, by failing to recognize or appreciate certain facts about the world, i.e. moral facts. For example, she says that "a person who, even on reflection, cannot see that having someone tortured because he failed to salute you is wrong plainly lacks the requisite ability [to distinguish right from wrong]" and therefore, "although like us, Jojo's actions flow from desires that flow from his deep self, unlike us, Jojo's deep self is itself insane." (Wolf 1987, 56)

The criterion of sanity, for Wolf, supplements the deep-self views, by offering what she thinks should be a satisfying response to the hard determinists' objection to the compatibilism of the deep-self views that we are no more free to choose our deep selves than we can our first order desires. Her point is that what is needed, beyond the ability to scrutinize and revise ourselves, is not the ability to *create* ourselves, as the hard determinists would have it, but rather only to *correct* ourselves, to "self-evaluate sensibly and accurately." (Wolf 1987, 60). This, for Wolf, marks a distinction between sane individuals, who are capable of evaluating themselves sensibly and accurately and only they can transform themselves as that evaluation tells them to, and therefore, can

“take responsibility for the selves [they] are but did not ultimately create,” (Wolf 1987, 60) and insane individuals, who do not possess that capacity.

Aside from the fact that Wolf’s criterion does little that should satisfy the hard determinists, there are several objections to her argument that such a morally loaded attribute of sanity is necessary for responsibility. She has considered some of these objections herself, and has attempted to respond to them, though, as I will try to show, unsuccessfully. Furthermore, the more powerful objections are those which do not appear in her essay, some of which are actually made salient by her answers to the others.

The first objection Wolf considers targets her apparent confidence in her own and other sane people’s judgments that they are in fact sane, that they appreciate the morally significant aspects of the world and that their deep selves are appropriately calibrated in response to those aspects. How can one ever be sure that one is any saner in one’s moral judgments than Jojo, a Nazi or a slave owner. Wolf’s answer is that “nothing justifies this except widespread intersubjective agreement and the considerable success we have getting around in the world and satisfying our needs.” (Wolf 1987, 60) She admits that in time we may discover that some things in our cognitive and normative outlook may be revealed to be mistaken, “but our judgments of responsibility can only be made from here on the basis of the values and understandings that we can develop by exercising the abilities we do possess as well and as fully as possible.” (Wolf 1987, 61)

There are two main ideas of what justifies a judgment of sanity in this response that I will list in reverse order: 1. what lets us get around in the world and satisfy our needs and 2. widespread intersubjective agreement. The first can be no help in deciding who is sane and who isn't, assuming that Nazis, slave owners and Jojo aren't sane, because such people are quite good at satisfying their own needs at the expense of others. Wolf might answer, that it is not what satisfies one's individual needs that matters, but what satisfies the needs of everyone. However, as ethical puzzle cases are meant to show, it is often the case that a given action must satisfy some people's needs at the expense of others. Furthermore, to assume such a general moral principle that one should act to satisfy the needs of all or of the greatest number, etc. cannot be the basis on which judgments of moral sanity are made, because they would then be question-begging. We cannot decide who is sane between the egoist and utilitarian just by assuming that utilitarianism is true.

As for the second idea, while intersubjective agreement does play a role in how we decide what is reasonable to expect from others, it cannot by itself determine who is morally sane and who is not, because that would make righteous dissent impossible in the same way that cultural-bound ethical relativism does. The dominant voice in society could then never be wrong in its ethical judgments, so that, contrary to Wolf's claims (which can be challenged for other reasons) slave owners of the past would actually turn out to be sane and abolitionists insane.

Nevertheless, these objections are not by themselves fatal to Wolf's view because they only target an epistemic claim about how we decide who (including ourselves) is sane or insane. That we could be largely wrong about that implies that there is an objective truth about who is sane or insane, in Wolf's sense, although it is not clear that there is such a truth anyway. The issue of the truth or falsity of moral realism, the view that there are objective moral facts about the world, on which partially rests the issue of the objectivity of judgments of moral sanity, is an enormous one in moral philosophy and resolving it is certainly beyond the scope of this study. That Wolf seems to assume its truth might by itself be taken as a weakness of her position, particularly because of the epistemic limits addressed in the preceding paragraph. More importantly, though, even if there are objective moral facts, it is even more doubtful that failure to recognize them or be moved by them constitutes insanity in a way that mitigates responsibility.

Wolf herself admits that "it would unduly distort ordinary linguistic practice to call the slaveowner, the Nazi, the male chauvinist even partially or locally insane," (Wolf 1987, 57) but maintains that "the reason for withholding blame from them is at bottom the same as the reason for withholding it from Jojo," (Wolf 1987, 57) i.e. they are unable to judge the world for what it is, and therefore their deep selves are not sane. I'm not sure how Wolf means to iron out the apparent contradiction in these two pronouncements. Perhaps she means that 'not sane' is the appropriate descriptor, though 'insane' is not, but I don't see why there is a difference between them. In any

case, sanity does not seem to be the issue. A slave owner may believe that her slaves are naturally inferior individuals, and would certainly be wrong in that opinion, but being wrong is not the same as being insane. Ignorance is not the same thing as insanity (or non-sanity). A Nazi may have been brought up to believe that there is an Aryan race whose members are the rightful rulers of humanity. Simply believing this, given a lack of evidence to the contrary, is not insanity. Believing it in the face of such evidence might be. So being a Nazi in a humane society is really a better ground (though not a decisive one) for attributing a lack of sanity to an individual than is an individual's being a Nazi having grown up in Nazi Germany.

Jojo's insanity, for Wolf, is due to the fact that Jojo was deprived of the sort of experiences necessary for developing an appreciation for the difference between right and wrong. I wonder, however, what sort of experiences would have been sufficient for such appreciation. Would any kind of countervailing influence suffice? A humanitarian leaflet found in the gutter? An overheard diatribe spoken by a prisoner being hauled off in shackles? A cousin with anti-authoritarian leanings who visited the house twice a year? Or looking directly into the eyes of a suffering person who pleads for his help? Suppose there are some cases in which these meager sources of dissent do sufficiently contribute to the 'sanity' of some individuals, but not in Jojo's case, where they are similarly available. Does Jojo, then, still get off the hook? If so, then it can't be because the right experiences were unavailable.

An even more powerful objection to Wolf's position, one that she anticipates, is that if her view is correct then only morally good people are sane. This is similar to the problem Dennett saw with his own conditions of responsibility and personhood. If responsibility requires rationality and it is irrational to be immoral, then no one is ever responsible for doing something morally wrong.[35] Wolf responds to this objection, in a way that she admits is not entirely satisfying, by distinguishing between the ability to judge the world for how it is and the exercise of that ability. However, as she herself recognizes, it is difficult to see what that distinction amounts to. If not exercising such an ability is a matter of weakness of will, then it seems that such a failure is the kind of case that is a paradigmatic example of something that, on the deep self views, one is not responsible for, because one does not fully identify with or approve of one's action. On the other hand, if the failure to do what one judges to be right or to refrain from doing what one judges to be wrong is not weakness of will, what it means to judge something to be right or wrong must be clarified. One might believe that there is a certain rule, and yet not wish to obey it. This seems to be the case with psychopaths. So a sane but immoral person would be one who knows what the moral facts are, because she knows what the rules others abide by are, but does not follow those rules. Perhaps what distinguishes Jojo from that sort of person is that Jojo hasn't ever learned of the moral rules of others, and so cannot judge what would count as an instance of following them versus not doing so. But again that seems to be a matter of ignorance, not insanity. We

[35] R. Abelson (2014) raises this objection and thinks that it also applies to Kantian morality in general.

might not punish an individual for committing a crime or an immoral action because that individual was ignorant of the rule forbidding it, but that is not the same as judging the individual to be insane. If Jojo fails to do what is right, to want to do what is right, and to want to want to do it, after being educated in moral principles, then the problem would have to be that he is not moved to do what he has been taught is morally right for he has not internalized those values. Such a failure cannot then be one of knowledge or correct perception, and so I think must be of *feeling*.

Cases where an individual self-consciously affirms something immoral are best understood not as irrational or insane, but as lacking concern for the interests of others. This can seemingly be either programmed by genetics or taught from experience, but in either case individuals who delight in torture, lie, cheat, steal, rape and murder without compunction, do not fail to see the world for what it is, but rather fail to care about it in the way that most of us do. Again, concern is necessary for responsibility, and therefore, for personhood, but that does not require that one is concerned for anyone other than oneself. Assuming someone like Jojo is aware of but does not abide or wish to abide by the principle that causing unnecessary suffering in others is wrong, then what Jojo lacks is empathy. However, a lack of empathy does not mitigate responsibility. It is precisely those who are aware of the harm they are causing to others but who do not care about that harm who are most justifiedly judged culpable for causing it.

III. Responsibility, rights and personhood

Another way in which personhood has often been thought to be bound up with ethics and morality is in the domain of rights. I have already mentioned Chappell who takes the class of persons to be our primary moral constituency (PMC). Similarly, Anderson (2000) defines a person as "any entity that has the moral right of self-determination," a right that distinguishes "*persons* from *pets* and from *property*. A person is the kind of entity that has the moral right to make its own life choices, to live its life without (unprovoked) interference from others." Other authors (e.g. Campbell 2011) who take persons to be the unique bearers of various rights have used the concept of a person to decide bioethical debates such as abortion and euthanasia. If only persons have a right to life, then deciding whether or not a fetus is a person should decide whether or not abortion is wrong. Similarly for euthanasia, if it is only wrong to kill persons or let persons die and not other beings, then it is crucial to know whether or not a fetus or someone in an irreversible persistent vegetative state, being kept alive only by artificial life support, is or is not a person, in order to know whether it is morally permissible to pull the plug.

However, there have been several challenges to the idea that persons are unique bearers of rights or moral significance. Most visibly, there is the animal rights movement, which claims rights to life, fair treatment, and freedom on behalf of non-human animals. One way to reconcile this tension might be to extend the sphere of personhood to include all animals, but this would be to abandon the primary usefulness

of the term ‘person’ since most of what we consider animals are clearly not responsible for their actions in the way that normal human beings are, because they lack self-consciousness. This fact is the reason for the absurdity of granting the status of ‘person’ to dogs and cats, and even non-living beings such as rivers. Avoiding this absurdity would require saying that beings other than persons can have rights. The example of Bolivia is instructive, where ‘nature’ has been granted rights, though there is no mention of personhood. (Vidal 2011) A further reason for which the usefulness of the concept of the person for resolving ethical debates has been questioned comes from the motley way it has been employed in the bioethical sphere, leading some writers to hold that it is too messy a concept, metaphysically speaking, to help us resolve the all-important life or death issues surrounding abortion, euthanasia and animal cruelty.

Gordijn (2011) is one writer who has has argued that the concept of a person should be removed from the bioethical arena. The main thrust of the attack is that because neither philosophers nor ordinary people can come to a consensus on what a person is, the concept can only be employed irresponsibly when it is used to demarcate a class of beings with a special moral standing as regards our freedom to harm or kill them. I agree that the danger Gordijn fears is real and that personal status does not decide who or what falls within our sphere of ethical concern, though I don’t endorse his argument for that claim.

Gordijn thinks that the concept of a person is hopelessly vague. There is no common usage of the term. It is largely an invention of philosophers and none of them

agree on what it means. For this reason, the term cannot be usefully employed in bioethical debates. It can only lead to deception and confusion, and it doesn't look like there is any emerging philosophical consensus that can fix it:

> a purely pragmatic use of the concept of the person as gathering the different qualities that transform an entity into a moral agent cannot be defended, since using the concept of the person only leads to confusion within the debate. This is… because the variety of lists of necessary conditions for personhood that the participants have in mind is so great, that the concept of the person is far from unambiguous. Therefore, using the concept does not contribute to mutual understanding and thus has no pragmatic use at all. (Gordijn 1999, 354)

The problem, as Gordijn sees it, is that since Locke divorced the concept of a person from any reference to a particular substance (physical or mental) personhood has been treated as a matter of a being (made of any kind of substance) having certain properties. While most contemporary thinkers assume that persons are always physical, since everything is, the focus is on the properties (biological or psychological) not the substance that possesses them (if there are substances at all.) The possession of these special properties not only makes something a person, but is supposed to give it a special moral status. A being with these properties, however they are delineated, is supposed to have special rights and privileges, e.g. a right to life and fair treatment. The problem is that because there is such a plurality of conflicting lists of person-making properties, one cannot be certain what someone is talking about when they use the word "person". If this is the problem, however, then it isn't a consequence of Locke divorcing the concept from the concept of any particular substance. If persons were brains, bodies, psyches or pneuma, the problem of disagreement over properties would

remain. It may have seemed easier before Locke, because it was assumed that mental substance had certain essential properties which are characteristic of persons. However, even if we accepted that, there could still be disagreement over what those properties are.

Gordijn claims that due to the vagueness of the concept ‘person’, it can easily be used as what he calls a “cover-up concept”. It will be useful to quote what he says here at length:

> Since there is no independent external criterion of demarcation of qualities that are and those that are not necessary conditions for personhood, a participant in an (sic) bioethical debate can simply choose a specific set of properties as being necessary for personhood in order to corroborate his own moral views. As it happens, his particular choice of certain qualities as being necessary conditions for personhood cannot be decisively criticized by his opponents, since there is no consensus on any ontology or metaphysics of the person that could deliver the necessary tools for such criticism. Through this circumstance, participants in bioethical debates can use the concept of the person as a tactical instrument, for by fixing a broader or a narrower concept of the person they can enlarge or diminish the group of human beings that can be looked upon as possessing moral status. In this way, they can morally justify their own acts with respect to certain groups of human beings as well as condemn certain other practices of which they, for some reason or another, do not approve. In this way, arguments using the concept of the person are a form of begging the question. (Gordijn 1999, 355)

The kind of question-begging Gordijn is worried about surely does take place. An example would be defining persons as fully self-conscious beings, claiming that therefore only fully self-conscious beings have a right to life, and then concluding that since fetuses are not fully self-conscious they are not persons, and therefore do not have a right to life. Another example would be to claim that anything that feels pleasure

and pain is a person, all persons have a right to life, a fetus can feel pleasure and pain, therefore all abortion is wrong.

Similarly, Ridley (1998) objects to philosophers employing personhood as "a specifically ethical concept, intended to indicate the possession of whatever properties are held to account for the ethical significance of adult human beings," (Ridley 1998, 115), because the question of how to decide which properties are ethically significant is usually left unanswered and the properties are pre-selected to fit the ethical views of whichever philosopher is employing the term. By way of *reductio*, Ridley offers the following example:

> I might define person as 'whatever has two legs and no feathers' (these are certainly properties enjoyed by most human adults); then go on to claim that persons and persons alone have ethical value. But this would be purely arbitrary. I have given you no reason to suppose that my definition of person captures anything of ethical significance at all…. personhood theorists almost always end up by choosing their technical definition of 'person' simply in order to get the conclusion they want. If the conclusion they want is that it is justifiable to kill fetuses, it is hardly surprising if they end up defining 'person' in terms that no fetus could match. (Ridley 1998, 115)

However, the problem, as I see it, is not that there is no good way to decide among competing accounts of personhood. The problem arises, rather, from trying to make one's metaphysical notion of personhood encompass the unique moral status one thinks persons must have.[36] Proceeding in that way will surely lead to circularity as far as bioethical debates are concerned. The mistake is not in providing a definition of

[36] Thomson (1971) seems to express a similar view when she grants for the sake of argument that a fetus is a person but then raises doubts about whether that entails that the fetus has a right to life and if it does, whether that right must always trump the right of the mother to sovereignty over her body.

"person" that other people might disagree with, but, first of all, in assuming that the sets of morally relevant properties and person-constituting properties are identical. That assumption requires argument and there is good reason to believe it is mistaken. For one thing, it is not obvious that only persons deserve our moral attention and concern. Animal rights advocates will certainly deny it. There may be morally relevant properties that are not unique to persons, e.g. the desire to continue living, the capacity for pleasure and pain, etc.[37] Similarly, it is not obvious that fetuses and irreversible coma victims, even if they are not persons, do not have a right to life, or are not deserving of dignity and respect. To say they don't requires further argument beyond saying that they aren't persons. (Not that such argument isn't available, at least in some cases). We might even think it right to treat some humans and non-human animals in the way we normally think to treat persons even while recognizing that they are not strictly so. Also, one might think that some persons are not deserving of some rights. For example, supporters of capital punishment might think that murderers have given up their right to life. Less radically, criminals in general often lose their right to move about unconstrained in the world. Whatever one's position is, this issue is not settled by appeal to personhood.

Now, it should already be clear from what I have said in chapter one that I don't think the task of clarifying the concept of a person is hopeless. I have offered an

[37] Gordijn includes the capacity for pleasure and pain in his list of possible person-constituting properties, but no serious account of the nature of personhood takes this capacity to be by itself sufficient.

account of the conditions necessary and sufficient for being a person, i.e. possession of the capacities for self-consciousness and concern, but I have done so with no eye toward solving ethical debates beyond the issue of responsibility. However, I have less confidence that we can conclusively settle the question of which properties to count as morally relevant and which properties a being must have in order to be the appropriate bearer of rights. Gordijn suggests that to avoid the circularity engendered by the use of "person", we should abandon the term and instead direct our attention to the morally relevant properties it is supposed to cover. Ridley, on the other hand thinks that "the search for special properties that give people the value they have seems mistaken from the start. I don't value you *for* your autonomy or *for* your higher brain function. I value you for being you, for being a person (whatever that involves)." However, neither approach is very promising. It is not clear which properties get to count as morally relevant and it's also not clear that we only value individuals for being persons. I value my cats, maybe for being cats, but mostly for being beautiful, sensitive, and fascinating among other reasons, just as I value persons for many, though not all, of the same reasons and other reasons besides. Arguments about which properties or beings have greater value are unlikely to succeed because fundamentally our values are based on our concerns which are not metaphysically grounded. Some theorists, however, particularly the ones who think persons have a unique ethical status, take higher cognitive capacities such as high intelligence, language or self-consciousness and responsible agency to be of special ethical significance, often claiming that only

responsible beings can the the bearers of rights, thinking that the two designations go hand in hand.

The crucial difference between responsibility and rights, is that whether or not someone is responsible for an action is an objective, metaphysically grounded fact, whereas what rights an individual or type of individual has is not determined by the facts alone, but requires an individual or collective attribution. Rights are conferred by persons on persons and other beings, they do not, as R. Abelson (2014) puts it, “grow on trees,” but, like our moral principles, duties and judgments (he thinks), are products of our explicit and tacit commitments to one another. I agree with the general claim about moral principles, duties and judgments, but will not argue for that more general claim here. However, it is clear to me that, both legally and morally speaking, rights are things that must be granted by a society. That does not mean that there aren’t better and worse reasons why some beings have rights and others don’t, but those reasons ultimately reflect our values and not objective facts about the world. That is why they often must be fought and campaigned for, they cannot merely be pointed out or argued for based on what is already known to be true. Arguments for rights can only proceed based on already agreed upon values or inconsistencies in the way that already recognized rights are conferred. For instance, if we already agree that causing unnecessary suffering is wrong, then anything with the capacity for suffering should have the right to live without being made to suffer unnecessarily (though of course what counts as ‘unnecessary’ is a problem in itself), or if a right is given to some individuals

on the basis of their being members of the human species and some group of humans is denied that right, then one can argue that the species membership of the latter group makes the denial of that right in their case unjust. However, extending that right beyond the human species cannot come merely from appeal to principle, but requires a change in values.

Abelson, however, thinks that because rights are based on agreed upon conventions, beings that cannot make agreements cannot have rights. All rights, according to him, must be claimable, as well as respected in other individuals who are granted the same rights, by the individuals who would have them. In other words, if one is unable to claim one's own rights or one is incapable of respecting the rights of others, then one cannot have any rights of one's own. Such criteria would entail that human infants, the severely cognitively disabled, and non-human animals that lack speech and theory of mind could not have rights.

Similarly, Hart (1955) holds that "animals and babies" are not appropriate bearers of rights.[38] For Hart, it is not sufficient for having a right that one is "capable of benefiting from the performance of a duty."(Hart 1955, 180) It is because that condition is generally taken to be sufficient "that animals and babies who stand to benefit from our performance of our 'duty' not to ill-treat them (which is to say only that ill-treating them would be wrong) are said *therefore* to have rights to proper treatment." (Hart 1955, 180) It is not, however, sufficient, according to Hart, because one does not determine who

[38] Though, unlike Abelson, he thinks there is one 'natural right', one "not created or conferred by men's voluntary action," (Hart 1955, 175), namely, the right of all men to be free.

has a right simply by determining who stands to benefit from the performance of a duty. One must instead examine “the transaction or antecedent situation or relations of the parties out of which the ‘duty’ arises.” (Hart 1955, 181) For instance, if a person X has promised to look after another person Y’s infirm mother while Y is away, then while the mother stands to benefit from X’s performance of his duty to look after her, it is not she, but Y who has the right to compel X to perform that duty, because it is Y to whom X made the promise and

> so it is Y, not his mother, whose right X will disregard and to whom X will have done *wrong* if he fails to keep his promise, though the mother may be physically injured. And it is Y who has a moral *claim* upon X, who is *entitled* to have his mother looked after, and who can *waive* the claim and *release* Y from the obligation. Y is, in other words, morally in a position to determine by his choice how X shall act and in this way to limit X’s freedom of choice; and it is this fact, not the fact that he stands to benefit, that makes it appropriate to say that he has *a right*. (Hart 1955, 180)

What having a right entails, for Hart, is having a moral justification for limiting someone else’s freedom to act. Therefore, having a right means that the individual who has it can appeal to that justification when compelling another individual to do other than the second individual wants. Appealing to such a justification or stating that someone is bound or held by a claim, along with the opposite actions of waiving a claim or releasing someone from an obligation seem to be things that only persons can do, because they require that one is self-consciously aware of one’s justification and can choose, in a sense implying responsibility for that choice, whether or not to exercise one’s right by holding another individual to an obligation.

According to Hart:

> [the above] considerations should incline us not to extend to animals and babies whom it is wrong to ill-treat the notion of a right to proper treatment, for the moral situation can simply and adequately described here by saying that it is wrong or that we ought not to ill-treat them. If common usage sanctions talk of the rights of animals or babies it makes an idle use of the expression 'a right,' which will confuse the situation with other different moral situations where the expression 'a right' has a specific force and cannot be replaced with the other moral expressions which I have mentioned. (Hart 1955, 181)

However, it is not clear to me that the case of the ill-treatment of animals, babies, as well as the severely mentally handicapped, differs substantially from the cases Hart thinks require use of the expression 'a right,' in the way Hart thinks they do. This is because we often do appeal to a moral justification for limiting individuals' freedom to harm babies and animals, and not always because doing so would be an infringement of *our* rights. Rather, we may claim the right not to be ill-treated *for* the animals, etc. In such a case, that it is their rights and not our own that are being appealed to follows from the fact that we do not have the option of waiving them or releasing others from their obligations to respect those rights. If the non-persons at issue could waive their rights then ill-treating them would not be wrong, but the fact that they are incapable of communicating that waiver does not mean that they don't have the right. This is why children who are neglected or abused can justifiably be taken away from their parents. The parents have a duty to treat the child well, because the child has the right to be well-treated. Having this right justifies us in limiting the parents' freedom to treat the child as they wish.

While I agree that only persons can make moral commitments, we can make those commitments to beings who are themselves unable to make them. We can commit to protecting, caring for, and otherwise being concerned for the interests of our pets or our cognitively disabled brothers, sisters and children, and on that basis confer rights onto them. Having a right means that a society commits itself to protecting the individual who has the right from being wronged in some particular way. Many creatures that are not able to explicitly claim their rights can still behave in ways that demonstrate that they have their own concerns, particularly that they not be caused suffering, and that they continue to be alive and healthy, such that we can get a sense of their interests and what constitutes a wrong done to them. Now, that doesn't mean that all beings are to be so protected from any sort of wrong being done to them, nor does it mean that some beings' rights and concerns might not trump those of others. This is an extremely difficult and complicated issue that I can't do justice to here, other than to show just how difficult and complex it is. We claim rights for ourselves and others based on which wrongs we take to be intolerable in our society, such as the infliction of physical or psychological pain or the denial of the freedom to pursue happiness. However, if someone's pursuit of happiness requires them to hurt someone else, the right not to be caused pain may trump the right to pursue happiness. On the other hand, if pain must be inflicted on some beings in order to design medical technologies that will reduce the chances of pain or death in many others, the rights of the many may outweigh the rights of the few. In most cases, we won't agree to force an unconsenting

human to undergo a painful experiment no matter the possible benefits, whether that human has the capacities of a person or not. Most of us are less inhibited when it comes to non-human animals, though some of us draw a line where we think the cost in suffering is not outweighed by the potential benefits. Testing a vaccine on an animal which may prevent millions of humans from contracting a deadly disease may be worth the animal's suffering, while trying out a new cosmetic on an animal to figure out whether or not it will irritate skin is not. Why we are more reluctant to sacrifice a human non-person than we are an animal does not merely track the capacities of the beings involved, but also various other factors including our emotional bonds and the roles that the beings play in society. Decisions about which rights trump which others, then, should not rest on a human/animal distinction nor a person/non-person distinction. In a situation where there is a quickly spreading fatal disease that could wipe out all of humanity, we might agree to ignore the rights of a few humans to save the rest. Whatever decisions we make will involve a weighing of our values not an easy appeal to metaphysical categories.

Anderson (2000) offers an example meant to show the difference between how we treat persons, versus how we treat pets or property, with respect to rights. He says that

> Property is the kind of thing that can be bought and sold, something I can "use" for my own interests. Of course, when it comes to animals there are serious moral constraints on how we may treat them. But we do not, in fact, give animals the same kind of autonomy that we accord persons. We buy and sell dogs and cats. And if we live in the city, we keep our pets "locked up" in the house, something that we would have no right to do to a *person*.

And generally of persons, he holds that it is:

> *morally wrong* to buy or sell them as property the way we do with dogs and cats or to otherwise use them for our own interests without taking into account the fact that they are moral agents with interests that deserve the same respect and protection that ours do…. Many of us would be prepared to say, I think, that any entity judged to be a person would be the kind of thing that would deserve protection under the constitution of a just society. It might reasonably be argued that any such being would have the right to "life, liberty and the pursuit of happiness." (Anderson 2000)

Anderson is correct that as a matter of fact we do sell pets and don't sell people, but the latter hasn't been or isn't always the case at all times or in all places and it is not out of the question that we could come to consider the former morally wrong and unacceptable. It is not inconceivable that dogs and cats could be granted the right to not be bought and sold as property, but only adopted by their caretakers. We don't give our pets complete freedom of movement, but that need not be because we are possessive of our property, but might be only because we think it is in the pet's long term interests to stay within certain bounds. In the same way, persons are restricted in their freedom of movement, and not only when we commit crimes that provoke a suspension of our rights, but we are not free to go into places that pose great hazards to our health or would constitute an invasion of someone else's privacy. While there might be some rights, e.g. free speech, that non-persons are incapable of enjoying, that does not mean they cannot have any rights. I see no reason why dogs and cats could not, and should not, in principle, be granted the rights to life, liberty and the pursuit of happiness - all things they are perfectly capable of enjoying.

Furthermore, I see no reason to think that just because a being is incapable of respecting the rights of others means that it cannot itself have rights. We respect some of the rights of persons who do not in fact respect the rights of others. For instance, freedom of speech is granted to fascists, racists, bigots, and psychopaths. Therefore, there is no reason why a being completely incapable of respecting others' rights, could not itself have rights. We may reasonably regard it as our duty to protect the interests of those who cannot protect themselves or claim their own rights, even though those others do not and cannot reciprocate. After all, our very young children have a right to our protection and care, but we have no right to their protection, nor could they offer it.

Rights aside, I see the general issue of which properties are morally relevant, and therefore which beings deserve our moral consideration, as an issue of how far and in what directions our concern extends. Different people, and therefore different societies, have different degrees of concern for the interests of severely cognitively disabled humans (such that they are incapable of self-consciousness), non-human animal pets, and non-human wildlife. This seems to be a function of the degree to which those people can empathize with those interests, which has to do with myriad properties that such beings possess. Some people take intelligence, self-consciousness, responsible agency, and species membership to be especially morally relevant, but I see no reason why we necessarily should be concerned with them. First of all, even among persons we do not usually take individuals with greater intelligence as deserving of greater moral concern, so I don't see why it should be relevant when comparing

persons to non-persons. Secondly, animals, like children, are often loved particularly for their supposed "innocence," the fact that they can't be responsible for their actions (particularly the bad ones), due to their lack of self-consciousness. Individuals who are weak, disabled, or unable to care for themselves require greater moral consideration to protect their interests, not less. Therefore, having the capacities necessary and sufficient for personhood does not put one in a special class, the members of which are uniquely deserving of the highest moral consideration.

Singer (1975/2002) famously argues for equal consideration of the interests of animals.[39] He claims that this follows from the general moral idea of equal consideration of the interests of all human beings. According to Singer, the equality asserted in such a principle is not a statement of fact and so:

> does not depend on intelligence, moral capacity, physical strength, or similar matters of fact… There is no logically compelling reason for assuming that a factual difference in ability between two people justifies any difference in the amount of consideration we give to their needs and interests. *The principle of equality between human beings is not a description of an alleged actual equality among humans: it is a prescription of how we should treat human beings*…. our concern for others and our readiness to consider their interests ought not to depend on what they are like or on what abilities they might possess. (Singer 1975/2002, 4-5)

However, our concern for the interests of others does require that they be capable of suffering or enjoyment of happiness, because "the capacity for suffering and enjoyment is *a prerequisite for having interests at all*." (Singer 1975/2002, 7) In other

[39] I am somewhat sympathetic to Singer's view that "the language of rights is a convenient political shorthand" (Singer 1975/2002, 8) for moral consideration, but I will not argue for that view here and anyway think I have already refuted what I take to be the most compelling argument for denying that non-persons can have rights.

words, having interests requires that one is concerned. Suffering and enjoyment are affective states tied to the satisfaction of one's goals. As Singer, following Bentham (1789/1907), argues, empathy and therefore consideration of interests is morally relevant in a way that other characteristics are not, because our feeling empathy for something depends upon the capacity of the being we feel empathy for (or at least our belief that the being has such capacity[40]) to have such states.

However, while I agree that the capacity for suffering and enjoyment of happiness is necessary (perhaps constitutive) of having *concern*, I understand the possession of *interests* to be somewhat broader. Trees, rivers and robots incapable of concern, might still have interests in the sense that they can be in better or worse condition, whether or not such conditions can have any significance for them. If it is appropriate to talk about the interests of these unconcerned beings, then there is no reason, besides our lack of concern for their interests, why they too cannot be part of our moral community, though they are not persons or even sentient beings. Nevertheless, it seems that we are primarily concerned for other beings insofar as they are themselves concerned.

If the primary morally significant property is the capacity to experience joy and suffering, then many non-human animals and severely cognitively disabled humans deserve as much moral consideration as persons. Being persons, with our capacities for both concern and self-consciousness, we are in a better position than non-persons to

[40] Or the relevant type of state necessary to feel for a fictional character, whether that is a real belief or some kind of pretend belief.

make choices that may benefit others, but that does not entail that we are deserving of greater benefit than others, except perhaps in the utilitarian sense that preserving the well-being of an individual who is in such a better position to help others, may allow for more good to be done in the long run, than preserving the well-being of an individual that is in a worse position to help others, that is, so long as one expects the former individual to actually act altruistically. However, even in that case, it might be that some animals, for instance a mother who must care for many of her young, may deserve greater consideration than some persons. In general, as Singer himself grants, the interests of persons may trump those of animals for various reasons. Perhaps, given our capacity for self-consciousness, our capacity for joy and suffering is greater than that of animals, or again we can potentially do more good for the world (though also more evil) than an animal can. Still, that does not mean that animals' and other nonpersons' interests should not be considered or that they cannot be part of our primary moral constituency. The differences in interests are of degree, not kind. Finally, some people might argue that we have special duties to members of our own species because our genetic destiny is linked with theirs. That might be true, but then the issue is not about persons versus non-persons, but about members of the human species versus members of other species.

IV. Metaphysical and Moral Personhood reconsidered

Dennett (1978) claims that the metaphysical and moral conceptions of personhood are inextricably connected, resting as unstable points on a continuum, and that neither can be analyzed in terms of necessary and jointly sufficient conditions. Scott (1990), responding to Dennett, attempts to extricate the two sorts of personhood into distinct concepts. Though I don't think that aim is fully achievable I am in other ways sympathetic to Scott's view. He defines metaphysical persons as "malleable higher order intentional systems." Their malleability or flexibility in the kinds of intentional states they are capable of having gives them the capacity for evil as well as good. (Scott 1990, 78) Therefore, he holds that:

> whether we like it or not... it is entirely appropriate to say that being moral is not a necessary condition of personhood. The figures we look upon as the most evil and morally depraved in human history were, no less than those of us who today consider ourselves to be persons, certainly persons. They possessed all the features we have so far attributed to persons; it just happened that malleability, in their cases, led to results of a horrifying nature. Whether we would really want to say, as Dennett has suggested, that we all as persons and moral persons might properly be said to be located on the same continuum, is, because of this, a more troublesome claim… (Scott 1990, 78)

I have argued in support of the claim that persons aren't necessarily moral and have explained how that is so, by distinguishing between the general capacity for concern, which is necessary for personhood, and the having of particular concerns for the interests of others, which is not. Scott suggests that being a moral person might essentially consist in having the capacity for "*caring*", and so doesn't distinguish that general capacity from specifically caring for others. He holds generally that a moral

person is one who chooses not to engage in actions that would interfere with or destroy the capacities that make another person a person, but instead acts in such a way as to contribute to the development and maintenance of those capacities in others. That's not far from the mark, I think, though I would add that a moral person should not interfere with and should instead contribute to the general flourishing of all beings with the capacity for joy, suffering and empathy, to the best of her ability insofar as that does not interfere with her own flourishing. In any case, Scott's definition of a moral person belies his attempt to fully disentangle the metaphysical concept of a person from it. If a moral person is a person who chooses to benefit other persons (and maybe non-persons as well), then a person, in the first place, is a being with, in some sense, the capacity to choose, i.e. with the capacity for responsible action. Therefore, as I have claimed above, the metaphysical concept of a person is inextricably linked with one aspect of the moral concept, responsibility, though not with all of them, i.e. moral goodness or possession of rights. So instead of talking about the metaphysical concept of a person versus the moral one, I propose that we just recognize one, metaphysical, concept, which has some relation to morality, in that it is the concept of a being that can be responsible for its actions, and from there just talk about morally good and morally bad persons, or even better (since most persons do some good and some bad things), morally good and bad actions performed by persons.

Bibliography

Abelson, Raziel. 1988. *Lawless Mind*. Philadelphia: Temple University Press

Abelson, Raziel. 2014. *Common Sense Morality*. Global Scholarly Publications: New York.

Aboitiz, F; Morales D, Montiel J. 2003. "The Evolutionary Origin of the Mammalian Isocortex: Towards an Integrated Developmental and Functional Approach". *Behav. Brain Sciences* 26 (5): 535–52.

Acampora, Christa. 2013. “Nietzsche, Agency, and Responsibility: Das Thun ist Alles”. *The Journal of Nietzsche Studies* 44(2):141-57

Anderson, David L. 2000. “What is a Person?” *Consortium on Cognitive Science Instruction*
http://www.mind.ilstu.edu/curriculum/what_is_a_person/what_is_a_person.php

Armstrong, Sharon Lee, Lila R. Gleitman, and Henry Gleitman. 1999. “What Some Concepts Might Not Be.” In *Concepts: Core Readings*, edited by Eric Margolis and Stephen Lawrence, 225-260. Cambridge: MIT Press.

Arnadottir, Steinvor Tholl. 2010 “Functionalism and Thinking Animals.” *Philos Stud* 147: 347-354

Babiak, P., C. S. Neumann, & R.D. Hare. 2010. "Corporate Psychopathy: Talking the Walk." *Behavioral Sciences & the Law*, 28 (2), 174-93

Baker, Lynne Rudder. 2008. “Big Tent Metaphysics.” *Abstracta* 2:8-15

Baker, Lynne Rudder. 2008. Review of *What Are We? A Study in Personal Ontology*, by Eric T. Olson. *Mind* 117 (468): 1120-1121

Baron-Cohen, Simon. 1997. *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge: MIT Press

Bartels, Daniel M. and Oleg Urminsky. 2011. “On Intertemporal Selfishness: How the Perceived Instability of Identity Underlies Impatient Consumption,” *Journal of Consumer Research* 38:182-198

Bartels, Daniel M. et al. 2013. “Selfless Giving,” *Cognition* 129:392-403

Batson, Daniel C. 1990. “How Social an Animal?: The Human Capacity for Caring.”

American Psychology 45 (3): 336-346

Bayne, Tim. 2008. “The Unity of Consciousness and the Split-Brain Syndrome,” *The Journal of Philosophy*, 105(6) 277-300

Bennett, Karen. 2009. “Composition, Colocation, and Metaontology,” In *Metametaphysics: New Essays on the Foundations of Ontology*, Edited by David Chalmers, David Manley, and Ryan Wasserman. Oxford: Clarendon Press

Bentham, Jeremy. 1789/1907. *An Introduction to the Principles of Morals and Legislation*. Clarendon Press: Oxford

Bermudez, Jose Luis. 2005. *Philosophy of Psychology: A Contemporary Introduction*. New York: Routledge.

Berns, Gregory. 2013. “Dogs are People, Too,” *The New York Times Sunday Review*, October 5.

Blatti, Stephan. 2006. “Animalism.” In *The Continuum Encyclopedia of British Philosophy*, edited by A.C. Grayling, Andrew Pyle, and Naomi Goulder, 1: 108-09.

Blatti, Stephan. 2007. “Animalism, Dicephalus, and Borderline Cases.” *Philosophical Psychology* 20 (5): 595-608

Block, Ned. 2002. “Concepts of Consciousness,” In *Philosophy of Mind*, edited by David Chalmers. Oxford: Oxford University Press.

Borowski, E.J. 1976. “Identity and Personal Identity.” *Mind* 85 (340): 481-502

Brown, Mark T. 2001. “Multiple Personality and Personal Identity.” *Philosophical Psychology* 14

Bruno, Michael and Shaun Nichols. 2010. “Intuitions About Personal Identity: An Empirical Study,” *Philosophical Psychology* 23(3):293–312

Butler, Joseph. 1736. “Of Personal Identity” In *Personal Identity* (2008), ed. John Perry. Berkeley: University of California Press.

Campbell, Colleen Carroll. 2011. "Personhood Begins When Life Begins". *The Washington Post,* November 2.

Carpenter, Amber D. 2014. *Indian Buddhist Philosophy*. New York: Routledge.

Carnap, Rudolf. 1950. “Empiricism, Semantics, and Ontology.” *Revue Internationale de Philosophie* 4:20-40

Carter, W.R. 1999. “Will I Be a Dead Person?” *Philosophy and Phenomenological Research* 59: 167-171

Caruso, Gregg. 2013. *Free Will and Consciousness: A Deterministic Account of the Illusion of Free Will.* Lexington Books: Lanham

Catterson, Troy. 2008. “Changing the Subject on Subjectivity” *Synthese* 162:385-404

Chalmers, David. 1995. “Facing up to the Problem of Consciousness,” *Journal of Consciousness Studies* 2(3): 200-219.

Chalmers, David. 1996. *The Conscious Mind*. Oxford: Oxford University Press

Chappell, Timothy. 2011. “On the Very Idea of Criteria for Personhood.” *The Southern Journal of Philosophy* 49

Chisholm, Roderick. 1976. *Person and Object: A Metaphysical Study*, La Salle, IL: Open Court.

Coburn, Robert C. 1985. “Personal Identity Revisited.” *Canadian Journal of Philosophy* 15 (3): 379-403

Collins, Steven. 1982. *Selfless Persons: Imagery and Thought in Theravada Buddhism*. Cambridge, New York: Cambridge University Press.

Craver, Carl. 2012. "Amnesia, Time and Memory: Perspectives from Clinical Moral Psychology," Paper presented at the Pittsburgh Area Philosophy Colloquium, Washington, PA

Dainton, Barry and Tim Bayne. 2005. “Consciousness as a Guide to Personal Persistence,” *Australasian Journal of Philosophy* 83(4): 549-571

Damasio, Antonio. 1994. *Descartes’ Error: Emotion, Reason and the Human Brain.* New York: HarperCollins

Damasio, Antonio. 2010. *Self Comes to Mind*: *Constructing the Conscious Brain*. New York: Pantheon.

Davis, Lawrence H. 1998. “Functionalism and Personal Identity.” *Philosophy and Phenomenological Research* 58 (4): 781-804

Degrazia, David. 2002. “Are We Essentially Persons?: Olson, Baker, and a Reply.” *The Philosophical Forum* XXXIII (1)

Dennett, Daniel. 1978. *Brainstorms*. Montgomery: Bradford Books.

Dennett Daniel. 1984. “Cognitive Wheels: The Frame Problem in AI.” In *Minds, Machines and Evolution*, edited by C. Hookway, 129-151, Cambridge: Cambridge University Press

Dennett, Daniel. 1987. *The Intentional Stance*. Cambridge: MIT Press.

Descartes, Rene. 1641/1993. *Meditations on First Philosophy.* Translated by Donald Cress. Indianapolis: Hackett.

Doris, J. 1998. “Persons, Situations, and Virtue Ethics.” Nous, 32:4: 504-530.

Doris, J. M. 2002. *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press.

Edwards, Paul. 1949. “Russell’s Doubts About Induction.” *Mind* 58(230):141-63

Elliot, Robert. 1991. “Personal Identity and the Causal Continuity Requirement.” *The Philosophical Quarterly* 41: 55-75

Flanagan, Owen. 2009. “Buddhist Persons and Eudaimonia[Buddha],” In *The Routledge Companion to Philosophy of Psychology*, edited by John Symons and Paco Calvo, 659-669. Abingdon: Routledge

Frankfurt, Harry G. 1971. “Freedom of the Will and the Concept of a Person” *The Journal of Philosophy*, 68(1): 5-20

Friquegnon, Marie. 2011. “Free Will, Liberation, and Buddhist Philosophy.” Presented at the Columbia University Seminar on Buddhist Philosophy.
Geach, Peter. 1962. *Reference and Generality*. Ithaca: Cornell University Press.

Goldberg, I., M. Harel, and R. Malach. 2006. “When the Brain Loses Itself: Prefrontal Inactivation During Sensorimotor Processing,” *Neuron.* 50:329-339.

Goldwater, Jonah. 2014 (Under revision/review). “The Metaphysics of Category Mistakes. A Two-dimensional Theory.”

http://media.wix.com/ugd/4ce7a5_a45becb6d7447eaaecb6469f7bf194ca.pdf

Goldwater, Jonah. 2015. “No Composition, No Problem. Ordinary Objects as Arrangements.” Forthcoming in *Philosophia*: 1-13

Gordjin, Bert. 1999. “The Troublesome Concept of The Person.” *Theoretical Medicine and Bioethics* 20: 347-349

Gordon-Roth, Jessica. 2015. “Locke on the Ontology of Persons.” *The Southern Journal of Philosophy* 53(1):97-123

Greenwood, John D. 1994. *Realism, Identity, and Emotion: Reclaiming* Social *Psychology*. London: Sage Publications Ltd.

Greenwood, Terence. 1967. “Personal Identity and Memory.” *The Philosophical Quarterly* 17 (69): 334-344

Hall, Stephen S. 2013. “Repairing Bad Memories,” *MIT Technology Review*, June 17. http://www.technologyreview.com/featuredstory/515981/repairing-bad-memories/

Hanley, Richard. 2014 (Unpublished manuscript). “How to Beam and Keep Your Body.”

Hare, R. D. 2003. *Hare PCL-R 2nd Edition*. Multi-Health Systems: Toronto..

Harman, G. 1999. “Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error.” Proceedings of the Aristotelian Society 99: 315-331.

Harman, G. 2000. “The Nonexistence of Character Traits,” Proceedings of the Aristotelian Society 100: 223-2 26

Hart, H.L.A. 1955. “Are There any Natural Rights?” *The Philosophical Review* 64(2):175-191

Heilman, Kenneth. 2002. *Matter of Mind: A Neurologist’s View of Brain-Behavior Relationships*. Oxford: Oxford University Press.

Hershenov, David. 2002. “Olson's Embryo Problem.” *Australasian Journal of Philosophy* 80 (4): 502-511

Hershenov, David. 2005. “Do Dead Bodies Pose a Problem for Biological Approaches to Personal Identity?” *Mind* 114 (453)

Horwich, P. 1987. *Asymmetries in Time*, Cambridge MA: MIT Press

Hume, David. 1738/2001. *A Treatise of Human Nature.* Oxford: Oxford University Press.

Inwood, Michael. 1999. “Care. Concern, Solicitude.”In *A Heidegger Dictionary*. Blackwell Reference Online http://www.blackwellreference.com/public/tocnode?id=g9780631190950_chunk_g97806311909506_ss1-1

James, William. 1952. *The Principles of Psychology*. Chicago: William Benton.

Johansson, Jens. 2010. “Parfit on Fission.” *Philos Stud* 150: 21-35

Katsafanas, Paul. 2013. “Nietzsche’s Philosophical Psychology” In *The Oxford Handbook of Nietzsche*, 727-755, Oxford: Oxford University Press.

Ketler, Alanna. 2013. “India Declares Dolphins and Whales as ‘Non-Human Persons,’ Dolphin Shows Banned”, *Collective Evolution*, September 17th. http://www.collective-evolution.com/2013/09/17/india-declares-dolphins-whales-as-non-human-persons/

Kim, Jaegwon. 1998. *Mind in a Physical World*. Cambridge: MIT Press.

Koenigs, Michael, Michael Krupke, Joshua Zeier, and Joseph P. Newman. 2012. “Utilitarian Moral Judgment in Psychopathy,” *Scan* 7:708-714

Korsgaard, Christine. 1989. “Personal Identity and the Unity of Agency: A Kantian Response to Parfit.” *Philosophy and Public Affairs* 18(2):103-31

Ladyman, James and Don Ross. 2009. *Everything Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Lau, Hawkwan and David Rosenthal. 2011. “Empirical Support for Higher-Order Theories of Conscious Awareness,” *Trends in Cognitive Science*. 15(8):365-373

Ledoux, Joseph. 2002. *The Synaptic Self: How Our Brains Become Who We Are*. New York: Penguin.

Levin, Paula F. and Alice M. Isen. 1972. “The Effect of Feeling Good on Helping: Cookies and Kindness,” *Journal of Personality and Social Psychology.* 21:384-388

Levin, Paula F. and Alice M. Isen. 1975. “Further Studies on the Effect of Feeling Good on Helping,” *Sociometry*. 38(1):141-147

Lewis, David. 1976. “The Paradoxes of Time Travel” *American Philosophical Quarterly*,

13(2):145–152.

Lewis, David. 1983 “Survival and Identity”, In *Philosophical Papers Volume I*. Oxford: Oxford University Press.

Levy. J. 1990. “Regulation and generation of perception in the asymmetric brain.” In *Brain Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry*. Edited by Colwyn Trevarthen. New York: Cambridge University Press, p.231-246.

Lindemann (Nelson), Hilde. 2001. *Damaged Identities, Narrative Repair*, Ithica: Cornell University Press.

Lindemann, Hilde. 2009. “Holding on to Edmund: The Relational Work of Identity,” In *Naturalized*
Bioethics: Toward Responsible Knowing and Practice. Edited by Hilde Lindemann, Marian Verkirk, and Margaret Urban Walker. New York: Cambridge University Press.

Locke, John. 1690/1979. *An Essay Concerning Human Understanding*. Oxford: Oxford University Press.

Low, Philip et al. 2012. “The Cambridge Declaration on Consciousness” Publicly proclaimed in Cambridge, UK on July, 7 2012.

Lurz, Robert. 2011a. “How Could We Know Whether Non-human Primates Understand Others’ Internal Goals and Intentions: Solving Povinelli’s Problem,” *Rev.Phil.Psych* 2:449-481

Lurz, Robert. 2011b. *Mindreading Animals: The Debate over What Animals Know About Other Minds*. Cambridge :Bradford Books

Maaike, Cima, Franca Tonnaer, and Marc D. Hauser. 2010. “Psychopaths Know Right from Wrong but Don’t Care,” *Scan* 5:59-67

Mackie, David. 1999. “Animalism *Versus* Lockeanism: No Contest.” *The Philosophical Quarterly* 49 (196): 368-376

Panksepp, J. 2007. “Neuroevolutionary Sources of Laughter and Social Joy: Modeling Primal Human Laughter in Laboratory Rats,” *Behav. Brain Res.* 182:231-244

Panksepp, J. and J. Burgdorf. 1999. “Laughing Rats? Playful Tickling Arouses High Frequency Ultrasonic Chirping in Young Rodents,” In *Toward a Science of Consciousness III*, Edited by S.R. Hameroff et al. Cambridge: MIT Press, 231-244.

Mackinnon, D.M.,F. Waismann and W.C. Kneale. 1945. "Verifiability," *The Aristotelian Society for the Study of Philosophy*, Supp. 19:119-50.

Makowski, I.J. and D.M. Weary. 2013. “Assessing the Emotions of Laboratory Rats,” *Applied Animal Behaviour Science* 148:1-12

Malcolm, Norman. 1958. “Knowledge of Other Minds,” *Journal of Philosophy* 55:35-52.

Manicavasgar, V., G. Parker, and T. Perich. 2011. “Mindfulness-Based Cognitive Therapy Vs. Cognitive Behaviour Therapy as a Treatment for Non-Melancholic Depression,” *Journal of Affective Disorders*. 130(1–2):138–144.

Martin, Raymond and John Barresi. 2006. *The Rise and Fall of Soul and Self*. New York: Columbia University Press

Mcmahan, Jeff. 2002. *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford

McTaggert, John Ellis. 1908. “The Unreality of Time,” *Mind: A Quarterly Review of Psychology and Philosophy* 17: 456-473.

Melden, A.I. 1977. *Rights and Persons.* Berkeley and Los Angeles: University of California Press

Mele, Alfred R. 2009. “Action and Mind.” In *The Routledge Companion to Philosophy of Psychology*, edited by John Symons and Paco Calvo, 609-621. Abingdon: Routledge

Mellor, H. 1998. *Real Time II*, London: Routledge

Merricks, Trenton. 2001. *Objects and Persons*. Oxford: Clarendon Press.

Messenger, Stephen. 2012. “New Zealand Grants a River the Rights of Personhood.” *Treehugger*, September 6. http://www.treehugger.com/environmental-policy/river-new-zealand-granted-legal-rights-person.html

Michaud, Stephen G. and Hugh Aynesworth. 1989. *Ted Bundy: Conversations with a Killer.* Authorlink Press: Irving, TX

Moor, James. 1982. “Split Brains and Atomic Persons,” *Philosophy of Science* 49(1):91-106

Moyer, Mark. 2008. “A Survival Guide to Fission.” *Philos Stud* 141:299-322

Nagel, Thomas. 1971. “Divided Minds and the Nature of Persons,” *Synthese* 22(¾):396-413

Nagel, Thomas. 1974. “What is it Like to Be a Bat?” *The Philosophical Review* 83(4):435-450

Nagel, Thomas. 1986. *The View from Nowhere*. New York: Oxford University Press.

Nichols, Shaun and Nina Strohminger. 2014. “The Essential Moral Self,” *Cognition* 131:159-171.

Nietzsche, Friedrich. 1885/2006. *Thus Spake Zarathustra.* Translated by Robert Pippin. Cambridge: Cambridge University Press.

Nietzsche, Friedrich. 1886/1989. *Beyond Good and Evil*. Translated by Walter Kaufmann. New York: Vintage Books.

Nietzsche, Friedrich. 1887/1967. *On the Genealogy of Morals.* Translated by Walter Kaufmann. New York: Vintage

Nisbett, Richard and Timothy Wilson. 1977. “Telling More Than We Can Know: Verbal Reports on Mental Processes,” *Psychological Review.* 84:231-259.

Noë, Alva. 2013. “If You Have to Ask You’ll Never Know,” in *NPR Blogs: Cosmos and Culture* October 11th

Noonan, Harold W. 2001. “Animalism *Versus* Lockeanism: Reply to Mackie.” *The Philosophical Quarterly* 51 (202)

Oderberg, David S. 1989. “Reply to Sprigge on Personal and Impersonal Identity.” *Mind* 98 (389): 129-133

Olson, Eric T. 1997. *The Human Animal*. New York: Oxford University Press.

Olson, Eric T. 2004. “Animalism and the Corpse Problem.” *Australasian Journal of Philosophy* 82 (2): 265-274

Parfit, Derek. 1971. “Personal Identity.” *The Philosophical Review*, Vol. 80, No. 1, pp. 3-27.

Parfit, Derek. 1982. “Personal Identity and Rationality.” *Synthese* 53 (2): 227-241

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Parfit, Derek. 1987. “Divided Minds and the Nature of Persons.” In *Mindwaves*, Edited by Colin Blakemore and Susan Greenfield. p.19-25. Oxford: Basil Blackwell.

Parfit, Derek. 1995. “The Unimportance of Identity” in *Identity*, Edited by Henry Harris. Oxford: Clarendon Press.

Parfit, Derek. 1999. “Experiences, Subjects and Conceptual Schemes” *Philosophical Topics* 26(½):217-70

Perrett, Roy W. 2002. “Personal Identity, Minimalism, and Madhyamaka.” *Philosophy East and West*, 52 (3): 373-385

Perry, John. 1970. “The Same *F*,*” The Philosophical Review* 79(2):181-200

Peterson, Dale. 2011. *The Moral Lives of Animals*. New York: Bloomsbury Press.

Prichard, James Cowles. 1835. *Treatise on Insanity and Other Disorders Affecting the Mind*.

Priest, Graham. 1998. “Numbers,” In *Routledge Encyclopedia of Philosophy*. Edited by Edward Craig. London: Routledge.

Priest, Graham. 2014. *One: Being an Investigation into the Unity of Reality and of its Parts, including the Singular Object which is Nothingness*. Oxford: Oxford University Press

Prinz, Jesse. 2012. *The Conscious Brain: How Attention Engenders Experience*. Oxford: Oxford University Press.

Puccetti, Roland. 1973. “Brain Bisection and Personal identity,” *The British Journal for the Philosophy of Science* 24(4):339-355

Putnam, Hilary. 1964. “Robots: Machines or Artificially Created Life?” *The Journal of Philosophy* 61: 668-91.

Putnam, Hilary. 1975. “The Analytic and the Synthetic,” In *Mind, Language and Reality: Philosophical Papers*. Cambridge: Cambridge University Press.

Quine, Willard Van Orman. 1951. “Two Dogmas of Empiricism”, *The Philosophical Review* 60 (1951): 20-43

Rea, Michael C. and David Silver. 2000. “Personal Identity and Psychological Continuity.” *Philosophy and Phenomenological Research* 61 (1): 185-193

Reed, Brian. 2008. *Secret Invasion: Captain Marvel*. New York: Marvel Comics.

Reid, Thomas. 1785. “Of Mr. Locke’s Account of Our Personal Identity.” In *Personal Identity* (2008), ed. John Perry. Berkeley: University of California Press.

Ridley, Aaron. 1998. *Beginning Bioethics: A Test with Integrated Readings*. St. Martin’s Press: New York.

Roth, Abraham Sesshu. 2000. “What was Hume's Problem with Personal Identity?” *Philosophy and Phenomenological Research* 61 (1): 91-114

Rosenthal, David. 2003. “Unity of Consciousness and the Self,” *Proceedings of the Aristotelian Society*, New Series 103: 325-352.

Rosenthal, David. 2005. *Consciousness and Mind*. New York: Oxford University Press.
Ryle, Gilbert. 1949. *The Concept of Mind*. London, New York: Hutchinson's University Library

Rosenthal, David. 2008. “Consciousness and its Function,” *Neuropsychologia* 46:829-840.

Sartre, Jean-Paul. 1956. *Being and Nothingness*. Translated by Hazel Barnes. New York: Philosophical Library

Sattig, Thomas. 2008. “Identity in 4D.” *Philos Stud* 140:179-195

Schechtman, Marya. 1990. “Personhood and Personal Identity.” *The Journal of Philosophy* 87 (2): 71-9

Schechtman, Marya. 1996. *The Constitution of Selves*. Ithaca: Cornell University Press.

Schechtman. Marya. 2014. *Staying Alive: Personal Identity, Practical Concerns and the Unity of a Life*. New York: Oxford University Press.

Sclafani, Richard T. 1971. “'Art', Wittgenstein, and Open-Textured Concepts.” *The Journal of Aesthetics and Art Criticism*, 29 (3): 333-341

Schiller, Daniela et al. 2010. “Preventing the Return of Fear in Humans Using Reconsolidation Update Mechanisms,” *Nature* 463:49-53

Schneider, Susan. 2009. “The Language of Thought.” In In *The Routledge Companion to Philosophy of Psychology*, edited by John Symons and Paco Calvo, 280-295.

Abingdon: Routledge

Scott. G.E. 1990. *Moral Personhood: An Essay in the Philosophy of Moral Psychology*. Albany: SUNY Press.

Sendak, Maurice. 1962. *Pierre: A Cautionary Tale in Five Chapters and a Prologue*. New York: HarperCollins

Shamay-Tsoory, Simone G., Hagai Harari, Judith Aharon Peretz, and Yechiel Levkovitz. 2008. "The Role of the Orbitofrontal Cortex in Affective Theory of Mind Deficits in Criminal Offenders with Psychopathic Tendencies" *Cortex* 46:668-677.

Shantideva. c.600/2006. *The Way of the Bodhisattva.* Translated by the Padmakara Translation Group. Boston: Shambala Publications.

Shoemaker, David. 2007. "Personal Identity and Practical Concerns" *Mind*, 116 (462): 317-57

Shoemaker, S. 2008. "Persons, Animals, and Identity." *Synthese* 162: 313-324

Sider, Theodore. 2001. "Criteria of Personal Identity and the Limits of Conceptual Analysis," *Philosophical Perspectives* 15(s15):189-209

Siderits, Mark. 2003. *Personal Identity and Buddhist Philosophy*. Aldershot, England ; Burlington, VT : Ashgate

Singer, Peter. 1975/2002. *Animal Liberation*. Harper Collins: New York.

Smith, J. David. 2009. "The Study of Animal Metacognition," *Trends in Cognitive Science.* 13(9): 389–396

Smith, JD et al. 1995. "The Uncertainty Response in the Bottlenose Dolphin" (*Tursiops truncatus*). *J. Exp. Psychol. Gen.* 124, 391-408

Smythe, Thomas W. "Chisholm on Personal Identity" *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 27(5):351-360

Sorabji, Richard. 2006. *Self: Ancient and Modern Insights about Individuality, Life, and Death*. Chicago: University of Chicago Press

Sperry, R.W. 1968. "Hemisphere Deconnection and Unity in Conscious Awareness." Invited address presented to the American Psychological Association in Washington, D.C., September 1967, and to the Pan American Congress of Neurology in San Juan,

Puerto Rico, October 1967.

Sprigge, T.L.S. 1988. “Personal and Impersonal Identity.” *Mind* 97 (385): 29-49

Sprigge, T.L.S. 1989. “Personal and Impersonal Identity: A Reply to Oderberg.” *Mind* 98 (392): 605-610

Sprigge, T.L.S. 1994. “Consciousness” *Synthese* 98(1):73-93

Steinbock, Bonnie. 2009. "Speciesism and the Idea of Equality." In *Morality and Moral Controversies: Readings in Moral, Social, and Political Philosophy*. Eighth edition. Pearson/Prentice Hall: New York.

Stevenson, Charles. 1937. “The Emotive Meaning of Ethical Terms,” *Mind* 46(181):14-31

Thomasson, Amie. 2007. *Ordinary Objects.* Oxford, New York: Oxford University Press.

Thomson. 1971. “A Defense of Abortion”, *Philosophy & Public Affairs* 1(1)

Thomson, Judith Jarvis. 2008. “People and Their Bodies,” In *Contemporary Debates in Metaphysics*, edited by Theodore Sider, John Hawthorne, and Dean W. Zimmerman. New York: Blackwell Publishers.

Unger, Peter. 1990. *Identity, Consciousness and Value*. New York: Oxford University Press. Reprinted in?

Van Inwagen, Peter. 1995. *Material Beings*. Ithaca: Cornell University Press.

Vidal, John. 2011. “Bolivia Enshrines Natural World’s Rights with Equal Status for Mother Earth.” *The Guardian,* April 10. http://www.theguardian.com/environment/2011/apr/10/bolivia-enshrines-natural-worlds-rights

Wegner, Daniel M. 2002. *The Illusion of Conscious Will*. Cambridge: MIT Press.

Weitz, Morris. 1956. "The Role of Theory in Aesthetics," *Journal of Art and Art Criticism* 15: 27-35.

Whiting, Jennifer E. 2002. “Personal Identity: The Non-Branching Form of 'What Matters.'” In *The Blackwell Guide to Metaphysics,* edited by Richard M. Gale. Oxford: Blackwell Publishers

Williams, Bernard. 1970. “The Self and the Future.” *The Philosophical Review* 79 (2): 161-180

Wilkes, Kathleen. 1988. *Real People: Personal Identity Without Thought Experiments*. Oxford: Oxford University Press

Wittgenstein, Ludwig. 1958. *Philosophical Investigations*. Translated by G.E.M. Anscombe. New York, Macmillan.

Wolf, Susan. 1987. “Sanity and the Metaphysics of Responsibility.” In *Responsibility, Character, and the Emotions*, edited by Ferdinand David Schoeman, 46-62. New York: Cambridge University Press.