LAWRENCE A. LOCKE[1]

PERSONHOOD AND MORAL RESPONSIBILITY

INTRODUCTION

There are several common justifications for society imposing sanctions on individuals. Society might wish to impose criminal sanctions upon me, for instance, to deter others from criminal behavior, to make a social statement about some activity, or to punish me because I am morally responsible for some reprehensible act.

The first two justifications have often been the subject of controversy. Possible complaints are obvious: What gives society the right to punish me for the purpose of deterring others? By what right may society use me as the means for a statement of social policy? Additionally, a general problem for these first two justifications is that for all their intents and purposes, my moral responsibility is irrelevant. Theoretically at least, punishing me could deter others, or could make a social statement, regardless of my moral guilt. And these goals could be satisfied by merely inventing a guilty party, or by claiming to punish and letting me go free. All that would be required is a clever story.

There has been less controversy over the third justification for imposing criminal sanctions. If I have intentionally done an act that I knew to be morally reprehensible, then there is great intuitive appeal to the argument that society ought to punish me because I am morally blameworthy and deserve punishment. Additionally, where moral responsibility is the issue, many of the objections described above do not apply. My moral guilt, of course, is relevant. And, where moral desert

[1] Associate, Stoel Rives Boley Jones & Grey, Portland, Oregon; J.D., Yale Law School, April, 1989; B.A., Reed College, 1979. I am grateful to Professor Jules Coleman, Yale Law School and Professor Robert Merges, Boston University Law School, for comments on earlier drafts of this article.

Law and Philosophy 9: 39–66, 1990.

is the concern, it would generally make no sense for society to punish the innocent, to invent a "guilty" party or to merely claim to punish and let a guilty party go free.

The imposition of sanctions using moral responsibility as the justification is not without problems, however. Society justifies a finding of moral culpability solely on the fact that "he is the one who did it" far too easily. This should not be the case. In the relationship that I bear to you, there is generally something lacking that precludes you being morally responsible for something I have done. So it is generally unfair for society to punish you for something I have done. There can also be something lacking in the relationship that I bear to myself that precludes my being morally responsible for something I have done. Under that condition it would be unfair for society to punish me for something I have done. (At least when then the punishment is based on moral responsibility.) That this is true is evidenced by our reluctance to punish, or to find moral responsibility, in cases involving insanity, extreme youth, hypnotism, etc. The problem is that our intuitions about moral responsibility in such cases can be unclear. This unclarity can lead to injustice. The debate, for instance, over the insanity defense rages on while judicial treatment remains inconsistent.

Fortunately, developments in the philosophy of personal identity can help by providing a framework for more accurate and consistent thought when moral responsibility is in question. A little background is necessary to understand how such an arcane subject can be of assistance:

There is a common sense in which a "person" is just one of a group of people, a human being. But there are other senses which play specialized yet important roles in our thought.

The concept of person additionally provides society with the seat of responsibility, the proper object of blame, punishment and reward. This honor, or burden, is reserved for persons alone, and is essential to the very idea of person. Thus, there is something special about persons that makes them distinct from any form of animal and that makes the concept distinct from "human being". No animal is considered morally responsible for its actions, and, although all people are considered or treated as human beings, not all are, in this sense, treated as persons.

In a more subjective sense, the concept of person becomes our

concern whenever we become concerned with those facts about ourselves that seem essential and with which we identify when we are satisfied that we will survive to some future time. This sense of the word is not limited to the mere idea of a locus of perceptions. What we think of ourselves, what others get to know when they get to know us "as a person", includes values, attitudes, beliefs, something of our past history and of our future plans: character in a broad sense. This more subjective meaning of "person" is often a concern of theories of personal identity.

There are at least two senses of the term "personal identity": First, there is the narrow "numerical identity", exemplified by our survival concern that we be identical with (be the same person as) some future person. Second, there is a broader notion of our identity as persons, exemplified by the phrase "my identity". This sense of the term includes all general facts about persons, such as that we are responsible agents, as well as individuating facts such as those about our characters. This second "identity" is what theorists try to reach with criteria for numerical identity.

There is an obvious connection between numerical identity and moral responsibility (How can I be held morally responsible if I cannot be said to be the person who did the crime?). There is also an obvious connection between moral responsibility and the concept of person as essentially a morally responsible agent. There is a more subtle relationship between moral responsibility and our personal identity in the sense of what we identify with when we reflect upon ourselves as persons. I shall state it now and discuss it later: It is the very nature of our self-awareness that justifies persons reserving for ourselves and for no other entities the honor (or burden) of moral responsibility.

The purpose of this paper is to examine the insights that psychological theories of the nature and identities of persons can offer to our thought about moral responsibility. In addition, I hope to suggest a framework for making judgments about moral culpability in legal contexts. The approach seems promising: a discussion of the relationships between moral responsibility and persons, the entities with which moral responsibility is exclusively associated. I proceed as follows:

In the first section I show, quite briefly, the development of the

prevalent purely psychological theories of personal identity,[2] and, in particular, their criteria for numerical identity, i.e., criteria which allow us to say "*A* is the same person as *B*". As part of the discussion, I demonstrate how logical difficulties led this quest astray, resulting in a modern theoretical requirement that psychological experience be embodied in a physical entity.

In section II, I discuss how certain modern theories have nevertheless emphasized the spirit of the purely psychological theories by focusing on "what matters" about persons and by disclaiming the intrinsic importance of numerical identity. After it becomes clear that numerical identity has no intrinsic importance to persons, it (perhaps surprisingly) will become clear that numerical identity (e.g., "He is the guy who did it") is neither necessary nor sufficient for moral responsibility. I argue that, therefore, the law should hesitate to find moral responsibility based solely on the fact that the person in question is the same person as the one who did the crime. Rather, the law should be concerned with what matters about persons relative to moral responsibility.

Section III includes a discussion of various candidates for "what matters". The discussion yields "psychological connectedness" (closely related to our character and to our reasons for acting) as the essential person-characteristic for deciding questions about moral responsibility for actions. I describe how the strength of psychological connectedness can vary widely over time and circumstance and why we might wish to incorporate this into our thought about persons.

In section IV, I question whether the fact that psychological connectedness can vary in strength over time and circumstance ought to be applied to judgments about moral responsibility for actions. After examining the reasons why we exclusively hold persons responsible for actions in that special way we call morally responsible, I conclude that it should.

---

[2] I am indebted in this introductory section to the first seven sections of John Perry, ed., *Personal Identity* (Berkeley: University of California Press, Ltd., 1975) – hereinafter referred to as *P.I.*

I.

In this section I describe how modern theories regarding the nature and identities of persons have evolved. I begin with the purely psychological theory of John Locke and end with the more recent work of Derek Parfit and John Perry whose theories require that psychological experience be embodied in a physical entity. We will see that the embodiment requirement is used merely to correct logical difficulties associated with purely psychological theories of personal identity.[3] A later section will show how Parfit and Perry emphasize the spirit, if not the letter, of the purely psychological approaches by arguing that what matters about persons is purely psychological.

For John Locke, a person was a mental entity, not just a locus of perceptions, but, more importantly, "an intelligent being, that has reason and reflection",[4] and is conscious of itself having perceptions. It seemed to Locke that one must be aware of having an experience for the experience to rightfully be called one's own, "it being impossible to perceive without perceiving that (one) does perceive".[5] Locke defined a person and his experience only with reference to his "reflective consciousness", and, "as far as this consciousness can be extended backwards to any past action or thought, so far reaches the identity of that person".[6]

The reason that Locke put this temporal limitation on personhood was that for him, "person" primarily represented the bearer of responsibility for actions. It seemed unfair to charge a person with responsibility for a past action so remote that he could not remember it. Since the person could not become conscious of having done the action, it could have nothing to do with his present consciousness.

---

[3] I do not mean to slight the significant problems associated with physical identity. Those, however, are beyond the scope of this article and are not essential to its thesis.

[4] John Locke, *Essay Concerning Human Understanding*, 2nd ed., ch. 27, (1964), reprinted in John Perry, ed., *Personal Identity* (Berkeley: University of California Press, Ltd., 1975), p. 39.

[5] *Id.*, p. 39.

[6] *Id.*, pp. 39–40.

Since a person was exclusively identified with his present consciousness, then, it seemed, the action could have nothing to do with him.

For Locke, then, a person is the same person who performed a past action if and only if he can remember being reflectively aware of performing that action.

Locke's concept of person is simplistic, but it introduces several intuitions which, fleshed out one way or another, remain in the literature today: that persons are essentially non-physical entities, that psychological criteria are sufficient to identify them, that persons are temporally limited by psychological features, and, as has been mentioned, that persons are the rightful bearers of responsibility.

As it stands though, Locke's theory contains logical difficulties, the most famous of which was demonstrated by Thomas Reid with his Brave Officer Paradox.[7] As the story goes, a young boy was flogged for robbing an orchard. Later in life, as a brave officer, he took a flag from the enemy. Still later, in advanced age, he was made a general. The brave officer could remember being flogged, the general could remember taking the flag, but the general could not remember being flogged. Locke's theory tells us that the general is the same person as the brave officer, that the brave officer is the same person as the boy, and that the general is not the same person as the boy. Logic, on the other hand, tells us that if *A* is the same person as *B* and if *B* is the same person as *C*, then *A* is the same person as *C*. Therefore, the general is not the boy, yet must be the boy. Hence the paradox. As a theory of numerical identity, Locke's is in trouble since it violates the transitivity of identity.

Anthony Quinton solves this particular difficulty in 'The Soul'.[8] A "soul-phase" is a momentary aspect of a person. *A* is the same person as *B* if they are connected by a series of soul-phases, each of which is directly continuous with its immediate predecessor and successor. Two soul-phases are directly continuous if the latter contains a memory of an experience of the former. According to this analysis, the general

---

[7] Thomas Reid, 'Of Memory', ch. 6 (1785), reprinted in *P.I. supra* note 2, at 113.

[8] Anthony Quinton, 'The Soul' *The Journal of Philosophy* **59** (1962): 393, reprinted in *P.I., supra* note 2, p. 113.

does not have to remember being the flogged boy to be the same person. He needs only to remember being someone who remembers being someone . . . who remembers being the flogged boy.

Quinton's analysis has its own problems. With his demand that each soul-phase be connected to its predecessor by actually remembering an experience contained in it, Quinton's theory breaks down, for instance, if the person is in the habit of sleeping. For in sleep, there is reason to claim, the chain of soul-phases connected in the required way is broken. As John Perry points out, Quinton's analysis "resolves the Brave Officer Paradox, in fact, only on the assumption that the brave officer is an insomniac consumed by memories of his most recent past".[9]

H. P. Grice added another level of complexity to the purely psychological theory when he addressed Perry's type of objection. Grice incorporated the notion of "possible memories" into his account and removed Quinton's requirement that each soul-phase, or "total temporary state" as Grice calls them, be linked by memory to its immediate neighbors.[10] Here, *A* is the same person as *B* if and only if they are end points in a series of total temporary states, each member of which, given certain conditions (such as if the person were awake), either would contain a memory of an experience contained in some previous member, or contain as an element some experience, a memory of which would, given certain conditions, occur as an element of some subsequent member. This account avoids the Brave Officer Paradox and other common counterexamples of memory criteria theories. I can be the same person as the one who performed some action even if I don't remember performing it, and I can be the same person after a lapse in reflective consciousness as occurs in sleep.

That philosophers go to such lengths to patch up an idea so fraught with logical difficulties does credit to the strength of the original intuition. There is a strong attractiveness to the idea that bodily identity is not a necessary condition for personal identity. When

[9] John Perry, 'The Problem of Personal Identity', *P.I. supra*, note 2, p. 18.

[10] H. P. Grice, 'Personal Identity', *Mind* **50** (1941): 330, reprinted in *P.I. supra*, note 2, p. 73.

someone gets to know me as a person, the idea goes, they do not get to know me by learning about my body. They must instead form some notion of my true character, my beliefs, values, history, and so on. And, if I should find my thoughts emanating from another body, my character and memories remaining intact, then I should consider myself to be the same person.[11]

Suppose I have total and irreversible amnesia and forget all my past experiences, including those upon which I draw for my beliefs, values, and goals. Those who knew me might well feel that the person they came to know when they came to know me "as a person" bears little relation to the person existing in the wake of this amnesia. They seem to have reason to claim that they are not confronted with the same person even though confronted with the same live human body.

For the above sorts of reasons, philosophers have tried to appeal to psychological continuity as the criterion of personal identity. Upon reflection, it is felt, it becomes clear that what we are really interested in finding, when we are interested in finding the same person, is someone psychologically continuous with him. Since memory is the most accessible or verifiable psychological condition, and since persistence of memory generally guarantees the psychological relationships considered essential to a person, then psychological continuity has often been analyzed in terms of memory, and memory criteria have been offered as criteria for personal identity.

But even after this convoluted set of repairs to Locke's theory, there remained another charge, against memory theories of personal identity, first leveled by Joseph Butler: An account of personal identity in

[11] I do not mean to suggest in this paper that embodiment has nothing to do with character or personality. If, for example, I should awake to find my thoughts emanating from a paralyzed body or a body of the opposite gender, significant personality and self-perception changes might well soon follow. However, I would still awake convinced that it was me, and, at least initially, my character, values, attitudes, and so on would be the same. These cases seem analogous to rather drastic changes of environment, after which personality or character changes of one sort or another might be expected.

terms of memory is necessarily circular, because memory "presupposes and so cannot constitute, personal identity".[12]

The charge of circularity is often expressed in something like the following manner: There is no directly knowable difference between a real memory (where the person actually had the experience) and a merely apparent memory (where the person only believes that he had the experience). Since one can have apparent memories of other persons' experiences, apparent memory cannot serve as a criterion of personal identity. This means that memory theories must rely on "real" memory. But for a person to really remember a past action, he must *be* the person who performed the action. To claim that a person's memories are real memories then, we must first know that he is the same person who performed the action or who had the experience. Therefore, when we use memory criteria, we are presupposing personal identity.

Derek Parfit and John Perry submit similar accounts of personal identity that solve the problem of the circularity of purely psychological theories. They do this by requiring a causal link between successive psychological stages of a person. The causal link is provided by the embodiment of the successive psychological stages within the human brain. Embodiment is thought to solve the circularity problem since it allows persons to be identified along a spatio-temporal path, thus preventing persons who have exactly similar memories from becoming counterexamples to the psychological theory.

Although both Parfit and Perry rely on embodiment, they are concerned to emphasize the spirit, if not the letter, of the purely psychological theories, with discussions of "what matters" about persons and their identities that de-emphasize the importance of embodiment. Several interesting things emerge:

We are told some very plausible things, about the nature of persons, that have important implications for moral responsibility. We recapture something lost along the way as philosophers patched up Locke's

[12] Joseph Butler, 'Of Personal Identity', *Analogy of Religion* (1726), reprinted in *P.I. supra*, note 2, p. 73.

theory – the notion that events which are so remote in time as to have little to do with our characters and with the nature of our actions, have little to do with us. And we are provided with criteria for an entity which, in a relatively non-problematic way, bears moral responsibility for all its actions.

Although their views are very similar, I will be concerned first and foremost with Parfit, since his formulation of the issues is most useful for present purposes.

II.

The realm of personal identity discussion is a fantasy world, filled with brain transplants, clones, and exact cell-by-cell duplicates. The reason for these devices is to separate that which is necessary and/or sufficient for personhood from that which is merely contingent. If I am convinced that all my mental experience will emanate from a new body in just the same way it emanates from my present body, then I am convinced that I am distinct from any particular body even though I am used to associating persons (including myself) with particular bodies.

It is important to note that the common conception of person, although tested and perhaps more clearly defined, is not changed at all by these exercises. The reason, for example, that the brain transplant cases are intelligible is because one has trouble seeing any important difference between what is present in that fantasy case and what is present in the real world. As far as personhood and identity (that it is me) are concerned, nothing seems to be missing. What this allows us to do is to eliminate items from consideration, or to realize that certain facts are necessary for consideration, and to get closer to the essence of the matter.

The role of a puzzle case borrowed from David Wiggins figures strongly in the ideas of Derek Parfit.[13] He intends that it support an interesting thesis: that numerical identity, per se, does not matter. It is

[13] Derek Parfit, 'Personal Identity', *The Philosophical Review* **80** (1971): 3, reprinted in *P.I. supra* note 2, p. 200.

widely felt that unless certain questions about identity (is it /will it be the same person?) can be answered, important questions about survival, moral responsibility, interest in the future, and so on cannot be answered. These questions seem to presuppose a question about personal identity; e.g., if I cannot be said to be the same person as the one who committed the crime, how can I be held responsible for it? Parfit's contention is that identity seems important in such questions because in the "real world", or normally, judgments about identity imply what is *really* important.

In a garden variety brain transplant case, one's entire brain is removed and inserted into another body. It is widely accepted, in these thought experiments, that what is important about persons follows along. The original personality, memories, mannerisms, and so on emanate from the new body. It is also generally agreed that, were I to undergo such a procedure, I would survive and be numerically identical with the resulting person.

Wiggins's case is a little different. First, it is assumed that the brain causally over-determines mental states, so that an entire brain is not necessary for their embodiment. (In fact, people have survived the destruction of significant portions of their brain and, where no dysfunction occurs, it seems problematic to deny that the person resulting from such damage would be the same person as the original.) Wiggins's case also differs from the usual in the following way: After removal, the two hemispheres of *A*'s brain are separated and placed in separate brainless bodies, resulting in persons *B* and *C*. *B* and *C* are initially exactly similar in their relationship to *A*. Disregarding *C*, the relationship of *A* to *B* is said to be just like the relationship of the original to the resulting person in the "garden variety" case above. Disregarding *B*, the relationship of *A* to *C* would also be exactly similar. Parfit argues that if all that matters obtains in the relationship of the original to the resulting person in the first type of case, then all that matters obtains between *A* and *B* and between *A* and *C*. ("What matters" about persons is a central theme developed in the next section.)

The problem here is that there seems to be no answer to the question "Will I survive?" if survival is taken to imply (numerical)

identity, whereas intuition has it that what matters in survival *is* identity (i.e., I survive if and only if someone later exists who is numerically identical with me). Parfit claims that the question "What happens to me?" has but three possible answers: (1) I do not survive; (2) I survive as one of the two people; or (3) I survive as both. The problem with (1) is that (as Parfit assumes) we have already agreed that if there were only one such resulting person, I would survive, and ". . . how could a double success be a failure?"[14] The problem with (2) is that both *B* and *C* are initially exactly similar, so whatever reasons one could have to claim identity of *A* with *B*, one would also have to claim identity of *A* with *C*.

Problems with (3) result whether we take it to mean "*A* survives as *each* of *B* and *C*" or "*A* survives as both *B* and *C*" (considering them to be one person with a divided mind). If we claim that (3) is the answer considering the former meaning, we claim our way into a violation of the transitivity of identity (where "survival" is taken to imply numerical identity): *B* would be the same person as *A*, *A* would be the same person as *C*, but *B* would not be the same person as *C*. If we consider the latter meaning and claim (3) to be the answer, we do severe damage, Parfit points out, to our concept of a person. "If they later met, they might even fail to recognize each other. It would be intolerable to deny that they were different people."[15]

Parfit concludes that there seems to be no plausible answer to the question "Will it be me?" in this case although it is clear that the question "Will I survive?" must be answered in the affirmative. Therefore, according to Parfit, we need a sense in which survival does not imply identity and in which one person can survive as two.

In fact, it seems that if we grant, as Parfit would, that survival does not necessarily imply identity, then the answer to the question "Will it be me?", taken as a question about numerical identity, will simply be "No". *A* will survive as each resulting person, but *A* will be identical with no resulting person. Regardless, Parfit feels that Wiggins's case makes the belief that the question "Will it be me?" must always have

[14] *Id.* p. 201.
[15] *Id.* p. 201.

an answer implausible. But more importantly, it makes it trivial, because it undermines the belief that identity is important.

The manner in which it is said to undermine this second belief is as follows: The relationship of *A* to *B* or to *C* contains all that matters in any normal case of survival. The relationship of *A* to *B* or to *C* does not contain identity. So identity cannot be what matters in any normal case of survival.[16] Parfit feels: I survive over time; what is important about me continues, normally, as one person. Thus, I am normally identical with some future person. But I could survive as two, each person containing all that matters in any normal case of survival. When I survive as two persons, the facts that imply identity with a surviving person do not obtain. But, so what?

It seems likely that if all that matters in any normal case of survival obtains in Wiggins's case, then all that matters about other concerns for which identity with some past or future person seemed the answer might also obtain. For example, in a garden variety brain transplant case, where I can non-problematically be said to be the future person, all that matters regarding responsibility for past actions exists. If I was responsible for a crime before the operation, then I would be afterwards too. Each branch of Wiggins's case is exactly like that case. So, Parfit would argue, how can the fact that I may or may not *be* some future person matter here?[17] The fact is that for many people, identity

---

[16] Note that simpler worms, cut in half, will regenerate each missing half, resulting in two worms. Clearly, the worm survives, but it cannot be said that the original worm is the same worm as either resulting worm. But, so what?

[17] There is, in fact, a flaw in Parfit's reasoning which I will mention but, since it does no damage for present purposes, I ignore it in the text: Suppose X → Y is the garden variety brain transplant case, and that

$$A \begin{array}{l} \rightarrow B \\ \rightarrow C \end{array}$$

is Wiggins's case. The fact that A → B is equivalent to X → Y in the absence of C and that A → C is equivalent to X → Y in the absence of B does not entail that these relationships in the *presence* of one another are equivalent to X → Y. In fact, A, like X, will have spent much of his life as a single entity and much of what matters to A will be connected with this fact. (wife, job, desire to be the best,

has seemed important. Parfit would explain this as follows: In the "real world", or normally, identity coincides with what matters and we find it convenient to imply what matters with identity statements. What matters and identity coincide normally because in the real world, survival is always one-to-one. So while Wiggins's case shows that identity does not really matter, it seems important, and in fact has a "derivative" importance because it normally can be, and is, used to imply what matters.

It seems very hard to deny the logic of Wiggins's case. If I consider a garden variety brain transplant case, I am hard pressed to discover anything missing of any importance in the relation of the person before the operation and the person that results from the operation. This gets even more difficult if my new body is similar to the old one. Perhaps it is a clone. Each branch of Wiggins's case seems just like that case. Given that we assume, as we should (at least for present purposes), that the brain causally over-determines its effects such that we could survive intact with half our brain, then Wiggins's case makes it hard indeed to insist that "Will it be me?" or some such identity question need be answered affirmatively before we can answer questions about moral responsibility and survival, or whether we should look forward to some future person's pain with horror or merely with sympathy.

What this suggests, in terms of moral responsibility, is that whatever reason we have for holding someone responsible for an act, it cannot be the fact that he is numerically identical with someone who did the act. All the reasons we could have to hold someone morally responsible for some act follow along through the brain transplant case. The brain transplant case seems to be mirrored exactly (in all of the relevant ways) by each leg of Wiggins's case. So if *A* did a morally blameworthy act in Wiggins's case, then *B* (at least) would be morally

---

unique, etc.). The best score for *all* that matters in Wiggins's case then would be one. But A will survive as the other as well. This objection is important regarding some of Parfit's later claims, but we part company before then. For present purposes, this problem I mention can be viewed as a change which "matters" but does not effect the continuation of survival, responsibility and so on.

blameworthy. But *B* is not the person who did it since *B* cannot be numerically identical to *A*. All that mattered in terms of moral responsibility followed along but the fact that "he is the guy who did it" did not.

Wiggins's case suggests that identity has the same derivative importance for moral responsibility that it has for survival. It is important *when* it implies what matters but it is not a necessary condition for what matters. Moreover, there are many examples that show that identity with "the person who did it" is also not a *sufficient* condition for moral responsibility. For example, hypnotism, insanity, coercion, etc., are all cases where "the person who did it" is not considered morally responsible.

That identity is neither necessary nor sufficient for moral responsibility, but has only a derivative importance, is not a fact of use only if split-brain transplants become possible. It seems that there are many situations where moral guilt is in question and where our intuitions are not clear. Is the virtuous family man to be condemned for his acts as an irresponsible eighteen-year-old? Should the insane stand trial when they are again capable? Are the totally reformed still deserving of punishment? Often such questions are answered in the affirmative, justified solely by the fact that "he did it". Because it often implies the reasons that matter, identity is taken for a reason itself. But if numerical identity has only a derivative importance, then to avoid injustice, questions about moral responsibility ought to be answered with reference to "what matters".

Not much has been said about "what matters" except that it appears to survive theoretical brain transplants. The nature of "what matters" and its further implications for moral responsibility shall be examined next.

III.

When I am concerned about my survival, I wish my mental life to continue. In addition, I wish it to continue with substantially the same character, complete with memories and goals.

Bernard Williams writes that the idea of an individual character,

broadly conceived, is essential to our concept of a person. He identifies a person with a character which includes broad "categorical desires" and "ground projects" closely related to his existence, from which spring much of his reasons for acting, and, "to a significant degree, give meaning to his life".[18] To those who know me as a person, as well as to myself, my individual character is as important as is the fact that I am a self-conscious thinking being. I wish my mental life to continue, but I wish it to be connected, through character, to my present self, its past and its projects for the future.

What makes the brain transplant cases intelligible and convincing is that the person prior to the surgery is psychologically continuous, in the above way, with the person resulting from the surgery. And the psychological continuity involved is apparently just the same as is our psychological continuity with our own non-problematic future selves. This is what obtains in Wiggins's case though numerical identity with a future person does not. *A* is psychologically continuous with *B* and with *C*. The mental life of *A* continues and we have all that seems to matter for survival.

Not only does the mental life continue, but there are all the same connections of character as in any non-problematic case of survival. Derek Parfit would say that *A* is just as "psychologically connected" with *B* and with *C* as we are with our non-problematic future selves and this is all that should matter.

For Parfit, what matters is not psychological continuity per se, but the component, or subset of psychologically continuous relationships, called psychological connectedness. Psychological *connectivity* is defined as "the holding over time of particular 'direct' psychological relationships"[19] such as the relation between an intention and an action, or between an experience and the memory of that experience. We are psychologically connected with that temporally extended stretch of

---

[18] Bernard Williams, 'Persons, Character and Morality' reprinted in Amelie Oksenberg Rorty, ed., *The Identities of Persons* (Berkeley: University of California Press, 1976) – hereinafter: *I. of P.* – p. 209.

[19] Derek Parfit, 'Lewis, Perry and What Matters', *I. of P. supra*, note 18, at 98.

our psychological continuity with which we are connected by memories, present experiences, future "projects" and so on. By way of contrast, psychological *continuity* is defined as a "chain of overlapping 'direct' psychological relationships".[20] We are psychologically continuous with ourselves as infants, but we are not, to any significant degree, psychologically connected to ourselves as infants.

Psychological connectedness is offered as the important aspect of psychological continuity because it is a necessary condition for our past experiences to contribute to our present experience and for our present experience, via projects extending into the future, to be connected with the future. Our future selves represent to us the completion of our projects, the satisfaction of our interests, desires, and intentions. Outside of this future connectivity, which defines what of our future selves is within the range of our interest, there is, for Parfit, nothing about this future self that matters to us. Those periods of our psychological continuity with which we are not "connected" are felt to be, in Bernard Williams's phrase, "beyond the horizon of our interest".

John Perry agrees with Parfit that the importance of numerical identity is derivative and that what matters in the continued existence of a person (and so *about* the person) are various "special relationships": the relationships of psychological connectivity. He elaborates on the notion that connectivity rather than continuity is what matters in his analysis of the interest we have in our own futures. This is expressed in the following terms: What reasons do we have to act now to promote our having or not having a certain property in the future? He writes:

> A person has a reason to act, if he wants some event to occur, and believes his performance of that act will promote the occurrence of that event. I shall call any events a person at a given moment wants to occur in the future his projects (at that moment).[21]

Perry defines an event broadly, so that they can include processes,

---

[20] *Id.* p. 98.

[21] John Perry, 'The Importance of Being Identical', *I. of P., supra*, note 18, p. 71.

states, and so on. It is a fact about persons, says Perry, that we are reliable; we are likely to have much the same interests, desires, and projects tomorrow that we have today. It is also a fact that there will normally be no better candidate for the completion of our projects tomorrow than ourselves. We seem to have a special concern about our identity with some future person, but, according to Perry, it is derivative from our interest in our present projects and their future completion. Just as we have no regrets about a past beyond the range of our memories, we have no interest in a future beyond the range of our projects.

The parts of our stream of consciousness with which we are significantly connected recaptures, but with allowance for interest in the future, Locke's idea that only things which can form a part of our present consciousness can be meaningful to us as persons. Locke would have it that we are not morally responsible for any action beyond the range of our memories because it could have nothing to do with us.

Accordingly, we have more "direct" psychological relationships, (we are connected to a greater degree) with ourselves in the recent past and the near future than with ourselves in the distant past and distant future. Memory fades, characters change.

This gives rise to the idea of "Parfitian Persons": A "person-stage", (like Quinton's "soul-phase") defines the entire composition of a person at a given moment. Where two person-stages are sufficiently disconnected (the story goes), we might, if we wish, consider them to be stages of different persons. So a life, given radical changes in character, can be viewed, if we wish, as a series of selves.

The notion of a Parfitian Person is only partly metaphorical. The criteria Parfit gives for personal identity is psychological continuity with a normal cause, so strictly speaking, a person would last a lifetime. What matters about persons, though, generally would not. Parfit and Perry both believe that remote portions of a persons's life will have little to do with his or her character, motivations, or anything about them that seems important.

What is interesting about psychological connectedness, as opposed to continuity or identity, is that whereas those two relations are all-or-nothing, connectedness admits of degrees. Since what matters about

persons through a weakening of direct relationships over time admits of degrees, Parfit is concerned that this should be reflected in moral thought.

Bernard Williams does not believe that moral thought adapts well to the varying degrees of relationships that hold between stages of the same person removed in time and character from one another. As an example, he discusses "promising".

> Suppose that I promise to A that I will help him in certain ways in three years time. In three years time a person appears, let's say A*, who's memories, character, etc., bear some, but a rather low, degree of connectedness to A's. How am I to mirror these scalar facts in my thought about whether, or how, I am to carry out my promise?[22]

Williams believes, first of all, that his promise applies not only to *A*, but to the potential recipient of the help as well, because the only act that could count as honoring that promise would be an act that would help *A**. How is he to mirror *A**'s "scalar" relations to *A*? He finds only three ways in which they could be so mirrored, none of which he finds satisfactory: First, the action promised itself might have some scalar dimension that could be brought into accordance with the proximity of *A** to *A*. But it seems silly to Williams that if he, for example, promised to pay $50, that he would be obligated morally only to pay $35. Second, varying with the degree of connectedness of *A** to *A* could be the stringency of the obligation itself. Williams finds this a more serious suggestion, but still silly (e.g., if the promise was to marry, he would still have a moral obligation to marry *A** but an obligation "which came lower down the queue").[23] Lastly, varying with the degree of connectedness could be "doubt or obscurity as to whether the obligation (of fixed stringency) applies or not." Williams finds this more familiar to our thought but at the expense of not "embodying the scalar facts; it is a style of thought appropriate to uncertainty about a matter of all-or-nothing."[24]

---

[22] Williams, *supra*, note 18, p. 98.

[23] *Id.* p. 203.

[24] *Id.* p. 204.

This does not seem to be fair to Parfit. First, by insisting that a promise to A is necessarily a promise to A^*, Williams loses sight of the fact that people are morally obligated for *reasons*, and that a promise to A may depend, for its moral force, on facts about A which are not true of A^*. Suppose that I promise to pay A some money because A is needy, A is my friend, and one should help out one's friends (and suppose that A knows this). But A^* manages to win the lottery and runs off with my wife. To insist that my promise was to A^* is to miss some facts that seem highly relevant to my moral obligation.

Second, in the above situation, I may consider my obligation all-or-nothing, but that does not mean that the diminished connectedness of A to A^* is not embodied in my thought. Clearly, the more un-A-like A^* is in the relevant aspects, the more sure I will be that the obligation does not apply.

It also does not seem fair to charge Parfit with the belief, as Williams seems to do, that all facets of one's character are acquired at the same time and fade with the same rapidity, or that all facets enter into a moral obligation or responsibility with equal force.

It seems more reasonable to interpret Parfit as meaning that when, due to changes in character, the character-reasons we had to be morally obligated, to be held morally responsible and so on, are gone or are significantly diminished, then we ought to take that into account. This does not seem to be such a difficult thing to do, and common language is well adapted to reflecting differing degrees of connectedness: "That was a long time ago", "but I was only sixteen" — many such common phrases are quite able to imply mitigation of the moral force of an obligation or responsibility.

Before concluding that this *ought* to be done, however, in the next few pages I examine the reason why moral responsibility attaches to persons in the first place, because it is possible that the reasons for persons having this special sort of responsibility would make accounting for these "scalar facts" an unreasonable thing to do.

IV.

One thing missing from the discussion so far is any psychological

characteristic, such that when an entity has it, that entity is a person. Also, no good reason has been given for why it is so common to include the fact that persons are responsible agents in the concept of a person.

In *Conditions of Personhood*,[25] Daniel Dennet speaks of six common themes applied to personhood: (1) That they are rational; (2) that they are conscious; (3) that they can be considered a person by others; (4) that they are capable of reciprocating #3; (5) that they are capable of verbal communication; and, (6) that they are conscious in some special way. With the possible exception of (4), the first five items in this list could arguably also be said of apes.[26] But Harry Frankfurt, in *Freedom of the Will and the Concept of a Person*,[27] describes a feature of human consciousness that does indeed seem to be limited to persons and that goes a long way toward explaining why responsibility for actions is attributed to persons.

Human beings, writes Frankfurt, share with animals the possession of a certain class of desires that motivate them towards action. Such desires might or might not be the result of deliberation; they might arise from hunger, discomfort, fear, the sub-conscious, or any of a variety of sources. They might be good, bad, morally neutral, of varying strength and simultaneously incompatible with one another. Where "to X" refers to an action (omission, etc.) then "want to X" is a "first-order-desire".

Since first-order desires can be competing, of varying strength, impractical, and so on, then although they motivate, they do not all move us to action. Those first-order desires that *do* (or will or would, when or if we acted), that are therefore "effective", define the "will".

Where "to X" refers not to an action but to a desire of the first order, e.g. "I want (to want to X)" then the desire is a "second order desire". A subclass of second order desires is the class of "second order volitions". A person has a second order *desire* when he wants either "simply to have a certain desire, or when he wants a certain desire to

---

[25] *I. of P., supra*, note 18, p. 175.

[26] Much recent work has been done with higher apes which suggests that they are capable of the creative use of sign language.

[27] *Journal of Philosophy* **68** (1971): p. 5.

be his will".[28] Second order *volitions*, though, are limited to those cases where a person wants a certain desire to be effective, to be his will. To use Frankfurt's example, a therapist might want to be moved by the desire to use drugs in order to better understand his patients' problems, and yet have no desire to actually take the drug. The therapist has a second order desire, but does not have a second order volition, with respect to a desire to take the drug.

According to Frankfurt, what is special about the consciousness of persons is the capacity to become reflectively aware of, and evaluate, first-order desires. Through this capacity persons determine whether they want those desires to be effective, i.e., to constitute their will. "No animal other than man . . . appears to have the capacity for reflective self-evaluation that is manifested in the formation of second-order desires",[29] but it is having second order volitions, and not second order desires generally, that he regards as essential to being a person.

Since persons possess a character in the broad sense described above, certain first-order desires will be acceptable and others will not. A person "identifies" with one desire and "withdraws" from another through the formation of second-order volitions. He wants to want the one, and wants to not want the other. In effect, he does not want to be the sort of person who acts on the latter desire. Hence comes the sense of the notion that an unwilling addict (to use Frankfurt's example) does not take the drug of his own free will. He has identified with the conflicting desire not to take the drug, and has withdrawn from the desire to take it, through the formation of the second-order volition to want to not want to take the drug. Since he is in fact moved by the desire to take the drug, then he does not have the will he wants.

The connection with moral responsibility is this: Were it true of persons, as it is true of animals, that they could not form second-order volitions, then they would be stuck with the will they have. They would be "helpless bystanders to the forces which move them".[30] But

---

[28] *Id.*, p. 10.
[29] *Id.*, p. 7.
[30] *Id.*, p. 16.

the structure of the consciousness of persons is such that they can be reflectively aware of the desires that move them, can evaluate them and can form preferences or second-order volitions with respect to them. Since they can affirm or deny their desires, they can be responsible for their will, and so for their actions, in a way that animals, infants, and mentally defective adults cannot. For Frankfurt, what is special about persons is that they uniquely have the capacity for freedom of the will.[31]

An interesting implication of this point of view is that a person can possess all sorts of reprehensible first-order desires and yet in no way have a reprehensible character. The character would not be reprehensible unless the person identified with the reprehensible desires and formed second-order volitions that they be effective. Thus the person who must continually struggle with temptation is not necessarily any more blameworthy than a saint.

The capacity to form second-order volitions is a useful person-characteristic to discover. It provides a feature of consciousness that distinguishes persons from non-persons (and immature or defective persons). By being a feature unique to persons and by being a necessary condition for moral responsibility, it shows why part of the concept of a person has been that only they are morally responsible agents. And it helps explain why moral responsibility attaches to persons in the first place – through self-awareness, evaluation, and second-order volitions, persons uniquely have the capacity for freedom of the will and so can be responsible for actions in a way that non-persons cannot.

Here, again, is the emphasis on the importance of character. As the example of the unwilling addict shows, persons are not always free to have the will they want, but to the extent their will is free, it is a true reflection of character. It is a true reflection of character because the nature of evaluations, and therefore of the will, is going to depend on character. It seems that if anything is properly blameworthy, it will have much to do with character.

At the end of the last section, I left open the question of whether

[31] *Id.*, p. 14.

significantly diminished degrees of psychological connectedness ought to be taken into account in questions involving moral responsibility. Now it seems that the answer to this should be "yes". Even if persons, in fact, last a lifetime, what is reprehensible or blameworthy about them does not, or might not. The reasons we have for holding persons responsible in that special way called morally responsible are reasons of character. And after significant change of the right kind, we can be left with something to which all the reasons we had for moral condemnation no longer apply. We would not have the sort of person who would do the reprehensible act.

This idea is intuitively appealing. There are, within many of our memories, I imagine, acts that we have done for which we were morally culpable, especially if we can remember back to the days of our irresponsible but accountable youth. We are psychologically connected to these persons, but we are connected to a much lesser degree than we are with our more recent pasts. It seems quite natural and in fact quite just that this fact be taken into account if we are to now be accountable for these acts, because we are not now the same sort of person who did these acts. It seems absurd to suggest that, for example, a mature family man of respectable character presently has the same degree of moral culpability for shoplifting or stealing siphoned gasoline as he did when he committed the act as a youth of eighteen. It seems natural to say that he is not as bad a person. It seems that what makes one morally responsible, rather than, say, strictly liable, is just because the act springs from a reprehensible character, a character that would intentionally do the morally wrong act.

Just how degrees of psychological connectedness should be taken into account is the subject of another inquiry, but a few observations are in order.

By taking degrees of connectedness into account, it would be ridiculous to mean that if a person were, say, nine-tenths changed, then he necessarily loses nine-tenths of his blameworthiness. It could be that he had only one persistent and quite horrible character flaw and it remains intact.

But suppose, for example, that a remorseless, sociopathic, habitual

criminal contracted a rare disease that left him with irreversible amnesia, to the extent that his mind became a virtual blank slate. He was thereafter nurtured, properly socialized and educated; he now has all the characteristics of a solid citizen. In this case it seems that no matter what unspeakable crimes this person committed before, he is now absolutely not deserving of moral blame. Here, the notion of Parfitian Person makes sense; there is no psychological connectivity at all, and we might as well consider him to be a different person.

We are inclined to find such persons blameworthy and deserving punishment even under the above science fiction circumstances. The inclination is even stronger in the following cases:[32] (a) *A* brutally kills *B*, knowing and relying on the fact that the horror of the experience will radically transform his character in all the appropriate ways; (b) *A* brutally kills *B* and then cleverly chooses to undergo the proper sort of radical character reconditioning in order to avoid punishment and accountability. We are inclined to feel that *A* should not be off the hook in either case. No doubt much of this inclination stems from the "derivative" importance of identity. To the extent it does, the inclination is understandable but baseless. The above are extreme examples of cases where it is not true that "he is the one who did it (some reprehensible act)" yields a reprehensible person. The facts about the persons before their character changes are especially galling, yet the end result is that we have little or no psychological connectivity and we are left with persons to whom the reasons we had for moral condemnation no longer apply. If we punish, we punish someone whose character is so far removed from that of the original person that we might as well be punishing a different person. In a sense, we would be punishing the son for the sins of the father.

This does not mean that our inclination to punish is entirely baseless. We have reasons for imposing criminal sanctions besides moral desert. We might well want to make a social statement and denounce these former behaviors, or we might want to punish and thereby deter others. Arguably, if we do not impose sanctions in such cases, we allow people generally to think they can get away with such

[32] Suggested to me as problematic.

schemes. Then, of course, there is retribution, or revenge (also understandable but inappropriate under the facts of these cases). But all these are reasons for which moral responsibility is irrelevant. Under the right circumstances, theoretically at least, they could be satisfied by a clever story, by the invention of a guilty party, or by falsely claiming to punish and setting the perpetrator free.[33]

In reality, the above sort of case will be rare or nonexistent. Normally, we are likely to be closely connected in the relevant ways to our recent pasts and less connected to our distant pasts. By the suggestion that we should take degrees of connectedness into account, it seems reasonable to simply mean something like this: When we have lost a significant amount of the character-reasons why a person was morally culpable, then we should take that into account in our thought about justice, responsibility, and punishment. This would not call for radical changes in our behavior or thought. We already recognize that justice demands this. Evidence of positive character change is already a mitigating factor in myriad situations. Evidence of character impairment is often a mitigating or even an exonerating factor. Defenses such as "he is a changed person", that was a long time ago", or "he was not himself" have always had intuitive appeal but have been offered, received, and applied inconsistently. Perhaps we have lacked a consistent rationale. An analysis of personhood and moral responsibility provides an appropriate rationale for taking character-reasons for moral condemnation into account: Since numerical identity is not a reason for moral condemnation, then besides character reasons, there are no reasons.

---

[33] Some say that we cannot satisfy these goals with a clever story. They basically argue that such ruses would be found out, the authority of the state would be undermined, and these social goals would suffer. Rather than ignore this objection, I fortify my statement as follows: Theoretically, all we would need is a *very* clever story. In any event, the objection is irrelevant. Distinctions between moral desert and other grounds for criminal sanctions can be validly shown by theoretical possibilities. Practicability is not necessary. Similarly, it was unnecessary for split-brain transplants to be practicable for Wiggins's case to be instructive.

CONCLUSION

The work that has been done on the topic of personal identity yields a variety of interesting notions.

Among them is the idea that the body is not as important as it seems. It might be important *to* persons, but it is not an important thing *about* them. They are not less of a person after losing an arm and a leg. And if their thoughts emanate from another body, with their characters, memories, etc. remaining intact, then it makes sense to say that they are the same person.

That they are the same person turns out to have no intrinsic importance, because it is not a necessary or sufficient condition for things that really matter about persons but is only a usual guarantee that these things will obtain.

Next, as a candidate for what matters, psychological continuity, a common subject of personal identity theories, turns out to be far too broad since we have more psychological continuity than we can use. We are continuous with our embryonic stages and with distant, dying stages maintained by heart and lung machines, but for most of the reasons we or anyone else would be interested in ourselves as persons, these stages are far "beyond the horizon of our interest".

What we are (or should be) interested in, seems to be those stages of ourselves with which we are closely enough connected that they bear some significant relationship to our present characters, to our reasons for acting, to what we are as persons.

After examining the reasons why persons exclusively are morally responsible agents, the above appears true about questions of moral responsibility as well. For moral responsibility as well as for other subjects of concern about persons, numerical identity has no intrinsic importance. We are left only with character in the broad sense of psychological connectivity, and the relationship it bears, through the formation of second-order volitions, to our will, and therefore to our actions.

A person may become disconnected, or connected to a significantly diminished degree, from the person he was when he was the "person

who did it". If so, then all, or many, of the reasons of character we had to hold this person responsible, in that special way we call morally responsible, no longer apply. It seems right that this fact should be taken into account.

The ideas illustrated in this paper do not call for radical changes in our thought about justice, responsibility, and punishment. Rather, they express the reason or pattern behind intuitions we already have. The intended result is articulation and understanding, and new mental tools toward consistency and justice.

Stoel Rives Boley Jones & Grey,
Portland, Oregon,
U.S.A.