RICE UNIVERSITY

Locke's Theory of Personal Identity: A Causal Interpretation

by

Edward Fried

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

Master of Arts

APPROVED, THESIS COMMITTEE:

Mark Kulstad, Associate Professor Philosophy

Richard E. Grandy, McManis Professor Philosophy

Hanoch Scheinman, Assistant Professor Philosophy

HOUSTON, TEXAS

MAY 2005
UMI Number: 1425827

Copyright 2005 by
Fried, Edward

All rights reserved.

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI Microform 1425827
Copyright 2005 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346
Anxiety is not the disease. It is but the symptom. The disease is nothing else than Life itself, which binds all creatures aware they have a past and a future to certain indeterminate limits. The cause of the disease is God . . . and so is God the cure.

Anonymous
Abstract

In this thesis I argue for a novel interpretation of Book II, Chapter xxvii of Locke’s *Essay concerning Human Understanding* according to which Locke’s theory of identity is taken to be an Aristotelian four-causal theory. On this interpretation, we are the same person as anyone who efficiently causes us to remember being them. More particularly, two diachronically distinct persons x and y are the same person just in case either x causes y to remember being x or y causes x to remember being y; and two contemporaneous persons x and y are the same just in case for all diachronically prior persons z, z causes x to remember being z if and only if z causes y to remember being z.

While this interpretation suffers from the drawbacks that Locke nowhere articulated such a view and had a well-known antipathy for scholastic notions generally, I argue that it has certain advantages. In addition to being at least consistent with Locke’s dicta on causation, it is both formally and substantively adequate. Moreover, it reflects Locke’s unconscious enmeshment in scholastic notions, and is useful as a heuristic; it sheds light on some of the more obscure aspects of the text (e.g., §12 and the concept of “man”). I conclude with the hope that future research will provide even more support for this novel interpretation.
Acknowledgments

I wish especially to thank my advisor, Mark Kulstad, for his help, confidence and support; my readers, Richard E. Grandy and Hanoch Sheinman for many friendly conversations and helpful comments; my wife Angela, for her selfless love and devotion to me during this project; and fellow graduate student Will Smith for his able proofreading assistance. All of these were NESS conditions (i.e., causes) of this project coming to fruition. I alone, however, am responsible for all its flaws.
# Table of Contents

Abstract .......................................................... -iv-
Acknowledgments .................................................. -v-
Table of Contents .................................................. -vi-
Diagrams .......................................................... -viii-
Formulae ........................................................... -ix-
Citation Style ...................................................... -xi-
0. Introduction .................................................... -1-
   A. Layout of the Thesis ........................................... -2-
   B. Limitations of the Project .................................... -3-
I. Basic Concepts ................................................... -4-
   A. Concept of A Person ............................................ -4-
      The Ontological View ........................................... -7-
      The Positivist View ........................................... -8-
      Quasi-Positivist Views ......................................... -9-
      Recent Developments ........................................... -10-
   B. Identity ....................................................... -11-
      Personal Identity ............................................... -15-
      Relative Identity ............................................... -16-
      Personal Individuation ........................................ -18-
II. The Mainstream View ........................................... -24-
III. Commentary ..................................................... -27-
   A. Relative Identity ............................................. -27-
   B. The Concept of a Person ..................................... -34-
   C. Personal Identity ............................................. -36-
      Reflexivity ....................................................... -36-
      Symmetry ........................................................ -37-
      Transitivity ..................................................... -38-
      Possible Memory vs. Experienced Memory (\(R^i R^i\) vs. \(R^i R\)) ................................. -44-
      Corrected Memory vs. Uncorrected Memory (\(R^c R^c\) vs. \(R^c R^i\)) .......................... -45-
      \(R^i R\) vs. \(R^c R^c\) ............................................... -58-
   D. Non-Substance View .......................................... -63-
   E. The Concept of “Man” ........................................ -67-
IV. The Circularity Objection ..................................... -71-
   A. The Fuller Definition ......................................... -74-
   B. Contents of Consciousness: The Mackie Position ............ -80-
   C. Plain Consciousness: Atherton’s View ......................... -83-
   D. Self-Consciousness ............................................. -93-
V. The Causal Interpretation ..................................... -96-
   A. Detailed Consideration of the Causal Interpretation .......... -99-
Matter—Material Causation ........................................ -99-
Life and Soul—Formal and Teleological Causation .... -99-
Person—Efficient Causation ................................. -102-
Causal-Remembering ........................................... -102-
B. Objections ..................................................... -104-
    Thiel .................................................................... -107-
    The Atherton Objection to the Causal Interpretation -108-
    Circularity Redux ............................................. -109-
    Causal Interpretation of the Prince/Cobbler Case ... -111-
C. Heuristic Uses of Causal Interpretation ................. -113-
    §12 ..................................................................... -113-
    The Concept of "Man" ......................................... -114-
00. Conclusion ...................................................... -116-
Bibliography ......................................................... -117-
Appendix A: Varieties of Privileged Access ................. -121-
Appendix B: Circular Definitions .............................. -123-
Diagrams

A.

B.

C.

D.
Formulae

Part-Person Criterion
I. \((\forall x)(Px_0 \supset \neg(\exists y)(Py & (x_m R y & \neg x_n R y)))\)

MAINSTREAM VIEW

Definition of Person
II. \((\forall x)(Px \equiv (Cx & x_{t_2}R x_{t_1}))\)

Personal Identity
III. \((\forall x)(\forall y)(((Px & Py) & t(x) > t(y)) \supset (x = P = y \equiv x R y))\)

COMMENTARY

Definition of Person—Diachronic
IIa. \((\forall x)(Px_0 \equiv (Cx & \neg(\exists y)(x_m R y & \neg x_n R y) & (\exists z)(Pz & x_0 = z)))\)

Personal Identity—Diachronic
IIIa. \((\forall x)(\forall y)((t(x) \supset t(y) & x = P = y) \supset (x R y \lor y R x))\)

Personal Identity—Synchronic
IV. \((\forall x)(\forall y)((t(x) = t(y) & x = P = y) \supset (\forall z)(Pz & t(z) < t(x)) \supset (x R z \equiv y R z)))\)

Combined Criterion IIIa & IV
V. \((\forall x)(\forall y)(((Px & Py) \supset (x = P = y \equiv ((t(x) \supset t(y) \supset (x R y \lor y R x)) \lor (t(x) = t(y) \supset (\forall z)(Pz \supset (x R z \equiv y R z))))))\)

CAUSAL INTERPRETATION

Definition of \(\varepsilon-\)
VI. \((\forall x)(\forall y)(x \varepsilon \rightarrow y \equiv ((Py \& (tx < ty \& y Rx)) \supset Px))\)

Part-Person Criterion
Ia. \((\forall x)(\forall y)(Px_0 \supset \neg(\exists y)(Py \& (y \varepsilon \rightarrow x_m \& \neg y \varepsilon \rightarrow x_n)))\)

Definition of a Person
IIb. \((\forall x)(Px_0 \equiv (Cx_0 \& \neg(\exists y)(Py \& (y \varepsilon \rightarrow x_m \& \neg y \varepsilon \rightarrow x_n))) \& (\exists z)(Pz \& z \varepsilon \rightarrow x_0))))\)

Personal Identity—Diachronic
IIIb. \((\forall x)(\forall y)(((Px & Py) & t(x) \supset t(y)) \supset (x = C = y \equiv (x \varepsilon \rightarrow y \lor y \varepsilon \rightarrow x)))\)
Personal Identity—Synchronic
IVb. \((\forall x)(\forall y)(((P_x \land P_y) \land t(x) = t(y)) \Rightarrow (x = C = y = (\forall z)((P_z \land t(z) < t(x)) \Rightarrow (z \varepsilon - x = z \varepsilon - y))))\)

Combined Criterion
Vb. \((\forall x)(\forall y)(((P_x \land P_y) \Rightarrow (x = C = y = (t(x) \neq t(y) \Rightarrow (x \varepsilon - y \lor y \varepsilon - x)) \lor (t(x) = t(y) \Rightarrow (\forall z)((P_z \Rightarrow (z \varepsilon - x = z \varepsilon - y))))))\)
Citation Style

All citations to Locke's *An Essay concerning Human Understanding*, Book II, Chapter xxvii are by section number and page in the Nidditch edition. All other citations to Locke are by Book, Chapter and Section number in the same edition. All citations to Leibniz are in the Remnant/Bennett edition of *New Essays on Human Understanding* and are by page number. Other citations are by (author ([year]), [page #]).
0. Introduction.

This thesis is an extended meditation on Chapter xxvii of Book II of Locke’s *An Essay concerning Human Understanding*, in which Locke presents his original theory of personal identity. It is however, more than a mere “rhapsody” or tapestry of interrelated commentaries. It is oriented toward making two points. (i) The first is that although Locke’s theory appears at times to have counterintuitive implications, it is (with some minor remodelings) both coherent and practical. Since many commentators have thought it was neither (As Curley (1982), 309) puts it, “this account of personal identity is teeming with paradox. . . . Thirty years ago the tendency was to regard such objections [as Reid’s] as fatal to Locke’s theory”) a successful defense of these claims would be a worthwhile project just in itself. (ii) The second is an argument for a novel interpretation of Locke’s text, according to which Locke’s theory of identity is best understood as being an Aristotelian four-causal theory.

(i) The defense has certain novel elements. Of particular interest are: the diachronic definition of person; the “assisted memory” criterion; a new approach to the “circularity objection” which examines each significant word in the definiens of “person” to see whether it gets Locke out of the circle, concluding that they don’t succeed; a “continued self-consciousness” interpretation of the criterion of personal identity (rejected in passing); my support of Locke against both the Reid/Berkeley and Parfit transitivity objections with novel defenses; and finally, the causal interpretation which I suggest as a defense against the circularity criticism.
(ii) Briefly, the argument is as follows. Given the most likely reading of Locke's view on the nature of persons and his apparent claim to be a "relative identity" theorist, his criterion of personal identity is viciously circular. This renders the relative identity thesis formally inadequate as an interpretation of the Lockean text. A competing proposal, the "relative individuation" thesis, is rejected as substantially inadequate; it cannot model all the cases. The "principle of charity" suggests we give consideration to an alternative interpretation (the causal interpretation) on which we take Locke's theory of identity to be de facto an Aristotelian four-causal theory. My claim is not that Locke "articulated," "would have supported" or "was committed to" such a theory. Rather, it is that the interpretation (a) is formally and substantively adequate (it is not viciously circular and models all the puzzle cases), (b) it reflects Locke's unconscious enmeshment in Scholastic notions, and (c) it is useful as an interpretive heuristic, casting light on some of the darker portions of the text (e.g., §12 and the definition of the concept of "man").

A. Layout of the Thesis.

Section I discusses the important background concepts of "Person" and "Identity." Section II ("The Mainstream View") is a brief precis of the Lockean theory of personal identity, in five parts: (A) Relative Identity; (B) The Concept of "Person"; (C) The Criterion of Personal Identity; (C) Non-Substance View; and (D) The Concept of "Man." Section III is a set of commentaries on each of the five aspects of the mainstream view. Section IV is a detailed consideration of the circularity objection. First (A) I examine all the significant words in the definition of person, to show that none of them protects the Lockean concept of
“person”—as I interpret it—from being circular given the relative identity thesis; then I defend my interpretation of that concept by excluding three rival interpretations, (B) the “Contents of Consciousness” (Mackie), (C) “Plain Consciousness” (Atherton) and (D) “Self-Consciousness” interpretations.

Section V is my presentation of the “causal interpretation” with several objections. Finally, Section 00 Is a brief conclusion.

B. Limitations of the Project

Secondary Literature. The use of secondary literature in this project has been selective. Rather than go for completeness, I have examined secondary texts that were frequently cited, perspicuous, or helpful (or all of these). Reid’s transitivity objection is both famous and acute; Leibniz is often very insightful although not frequently cited in this context; Atherton is neither frequently cited nor especially acute, but examination of her work helps develop the argument; Chappell is both well-known and acute, and his position must be refuted for my argument to go through; and so on. It follows from this that I cannot be certain that what I believe to be innovations have never seen print before. However, I do think that some of those innovations are so worthy of note that they would have occasioned comment by other “toilers in the field” (such as Thiel (1998 a & b) and Stuart (2004) whose work is both more synoptic and of relatively recent date.

Non-Substance View. The commentary on this aspect of Locke’s theory has been truncated in the interest of reducing the length of the thesis. I merely formulate a position, which no doubt many will find controversial. I hope to have the opportunity to defend my view at greater length in a possible future version of
this work.

Rest of the Essay. Examination of the Essay is limited to Book II, Chapter xxvii with rare and limited exceptions (e.g., references to Book II, Chapter xxvi in the section on the causal interpretation). In part this is justified by the fact that many commentators are of the opinion we cannot expect to derive a completely consistent reading of Chapter xxvii and the other, earlier portions of the Essay. (The chapter “Of Identity and Diversity” was added to the second edition of 1694.) I don’t think that is quite satisfactory, however. In particular, I believe that an examination of Locke’s position on essences is requisite to a full understanding of his views on identity. However, material on essences proved to be excisable in toto from the work with no harm done to the argument; so I have suppressed it in the interests of maintaining the thesis at a more reasonable length. A possible future version of this work would include this material, other dicta on identity found elsewhere in the Essay, and perhaps other concerns as well.

I. Basic Concepts

A. Concept of A Person

One way of thinking of the word “person” is as a term with a group of disparate but coordinate uses centered around certain distinctive aspects of the individual human life. It will perhaps be helpful, in getting a feel for the terrain, briefly to examine the term in historical context. It comes from the Latin word “persona” which was used for the masks which actors wore in the theater. The
masks were "characteristic" in the sense that different plays would use the same character names. The masks would quickly tell the audience which part was being played by which actor; and so the term came also to mean a role or part. This is the way in which Cicero used the term in perhaps its first philosophical appearance (Thiel (1998b), 869). In De Officiis (On Obligations), he mentions five personae or features that distinguish characteristic aspects of human social life: (a) Possession of reason (§ 107); (b) endowment with individuating natural attributes (§ 107); (c) social status (§ 115); (d) chosen profession (§ 115) and (e) orientation to morality (§ 117).¹ (2000.)

More recently, the term has been used by animal rights activists reductively to distinguish those entities that are mere rights-holders (Regan (1983)) or whose interests should be counted in a utilitarian maximizing calculation (Singer (1993)). Whatever we may think of the means and goals of these philosophers, it is in any case evident that their approach to personhood is the polar opposite of Cicero's. Whereas the animal rights activists are at pains to show how we are like other animals, Cicero's concern for duties led him to emphasize that which is distinctive about human beings: "It is relevant to every aspect of obligation always to focus on the degree to which the nature of man transcends that of cattle and other beasts." (2000, § 105.)

These positions delineate the extremes—both temporally and philosophically—of the uses which can be ascribed to the term "person."

Although it would be reasonable to go on to consider the middle ground in some

¹ Thiel for some reason says there are four. (1998b, 869.)
depth, I believe it will be both more efficient and illuminating to take a different
tack. Rather than thinking of the term as one whose definition has changed over
time, I suggest we approach it as a partly determinate concept for which different
philosophers have offered diverse positive accounts in response to the
requirements of the time and their other philosophical commitments. Take for
example (overgeneralizing somewhat) Cicero, a Stoic, living in pre-Christian
Rome.\footnote{\textit{De Officiis} was written between 46 and 44 B.C. (2000, 12.)} Perhaps this time of brutishness and vice needed to be reminded of the
differences between man and beasts. On the other hand, Singer and Regan—a
Utilitarian and a Kantian, respectively—write in a time in which, by the exercise of
will, mankind has to some extent disciplined its instincts. Such a time perhaps
needs to be reminded of our similarities to other animals.\footnote{As Jung wrote: “Since the differentiated consciousness of civilized man
has been granted an effective instrument for the practical realization of its
contents through the dynamics of his will, there is all the more danger, the more
he trains his will, of his getting lost in one-sidedness and deviating further and
further from the laws and roots of his being. This means on the one hand the
possibility of human freedom, but on the other it is a source of endless
transgressions against one’s instincts. . . . [O]ur progressiveness, though it may
result in a great many delightful wish-fulfillments, piles up an equally gigantic
Promethean debt which has to be paid off from time to time in the form of
hideous catastrophes.” (1968, 162-63.)}

More to our purpose, anyone offering a positive account of the concept of
“person” in the 17\textsuperscript{th} century would have considered such an concept as subject to
the following five constraints (Thiel, (1998b), 869-70):

\begin{itemize}
\item[(1)] It would be useful in discussions of rights and/or responsibilities.
\item[(2)] It would be useful in discussions of human action and justice.
\item[(3)] It would be useful in discussions of divine justice (the individual
\& Last Judgments).
\item[(4)] It would be useful in discussions of the Trinity.
\end{itemize}
(5) It would be responsive to the intuition that all and only rational, self-conscious beings are persons.

According to Thiel, there were two prominent approaches to the concept of personhood in use at that time: (1) the substantial person or "ontological" view; and (2) the legal person, or (what I'll call) the "positivist" view.

The Ontological View

On the one hand, the ontological person was (after Boethius) a unitary rational substance which combined both soul and body. (Thiel, (1998b), 870; Boethius (1973), 10-11.) Although not committedly either Platonic or Aristotelian, the concept of the persona as combined substance seems prima facie more conformable to the Aristotelian view that soul is the form of the body (i.e., its formal cause). As Bambrough (1995) writes,

"Aristotle’s theory of soul places him, with Spinoza, in the . . . tradition of philosophical reflection on the mind-body problem, according to which the mind or soul and the physical body are not totally independent substances, but aspects, separable only in thought, of one and the same substance." (230.)

Indeed, this ontological view of the person predominated in Scholastic discussions throughout the middle ages and up to Descartes’s time. (Thiel (1998b), 870-71.) Descartes, building on the Platonic view that the soul as the "knower of the forms" is in "bondage to the body" (Melchert (1995), 130) and inherently separable from it, would go on the hold that the “person” is not a single rational substance, but rather a composite of substances. Indeed, he held that these two substances—res cogitans and res extensa—are “really distinct” (Meditation 6). (Descartes (1995), 330.) Moreover, in his view the concept of a person was subordinate to that of the “self” or “I”—“That by which I am what I
am” 4 which he identified with the soul and “res cogitans.”

The Positivist View

There is no obvious reason why the ontological view of the person in either Cartesian or Scholastic incarnations could not have been used in philosophical discussions of legal and political matters. But in practice it was mainly limited to metaphysical and theological discussions, for example, those related to the Trinity (Thiel (1998b, 870), the nature of the resurrected body, etc. On the other hand, the positivist view—that a person is just that entity which acts subject to its rights and responsibilities—predominated in social and ethical contexts. 5 For example, Hobbes, whose writings came out of the Natural Law tradition, 6 defines “person” (at xvi.1) as “he whose words or actions are considered either as his own, or as representing the words or actions of another man, or of any other thing to whom they are attributed, whether truly or by fiction.” In other words, a person is both an entity to whom acts can be attributed and to whom responsibility for such acts can be transferred (or claimed to be transferred) by acts of transference. This aspect of personhood (i.e., agency) was of particular importance to Hobbes, because in his theory the sovereign was appointed through the agreement of persons to transfer their natural rights to the commonwealth for their own good:

The only way to erect such a common power as may be able to defend them from the invasion of foreigners and the injuries of one

4 Discourse on the Method, Part IV. (Descartes (1964-74), VI 33.)


another, and thereby to secure them in such sort as that by their own
industry, and by the fruits of the earth, they may nourish themselves
and live contentedly is to confer all their power and strength upon one
man, or upon one assembly of men, that may reduce all their wills, by
plurality of voices, unto one will, which is as much as to say, to appoint
one man or assembly of men to bear their person, and every one to
own and acknowledge himself to be author of whatsoever he that so
beareth their person shall act, or cause to be acted, in those things
which concern the common peace and safety, and therein to submit
their wills, every one to his will, and their judgments, to his judgment.
(1994, xvii.13, 109.)

But this focus on agency naturally makes the term “person” unsuitable for
various other uses. Hobbes claimed that he used the term so that it was
consistent with traditional thinking on the trinity (xlili. 3, 334):

But a person (as I have shown before, ch. [xvi] is he that is
represented, as often as he is represented. And therefore God, who
has been represented (that is, personated) thrice, may properly enough
be said to be three persons (though neither the word Person, nor
Trinity, be ascribed to him in the Bible).

But there seems no way that a person as rights holder or transferer could
naturally survive death, nor any way that a “person” like the Leviathan: “This
done, the multitude so united in one person is called a Commonwealth, in Latin
Civitas” (xvii. 13, 109) could be “conscious” in the normal sense of the word. So
from our point of view, the Hobbesian usage feels somewhat eccentric.

Quasi-Positivist Views.

Locke was the first to originate a quasi-positivist view, viz., a view of the
“person” as rights/responsibility holder but based on neither substance nor legal
fiction, but some other factor (in this case, memory of being prior persons).
Locke’s quasi-positivist approach could be seen as a way of pulling together all
the disparate uses of the concept of “person” (with the exception of the Trinity).
(Thiel (1998b), 871.) However, from the point of view of his contemporaries, the
project had a fatal shortcoming. Although this approach freed Locke from
dependence on an abhorred Scholastic notion ("substance"), it left the retribution
for our deeds in an after-life a matter of God’s special and inscrutable attention,
beyond human comprehension except through faith in revelation.

All the great Ends of Morality and Religion, are well enough
secured, without philosophical Proofs of the Soul’s Immateriality; since
it is evident, that he who made us at first begin to subsist here, sensible
intelligent Beings, and for several years continued us in such a state,
can and will restore us to the state of Sensibility in another World, and
make us capable there to receive the Retribution he has designed to
Men, according to their doings in this Life. (IV.iii.§6, 542.)

It was widely thought this dependance on revelation rather than reason would
leave the belief in God and the soul open to skeptical attack, which of course
proved to be the case.

Recent Developments

Most modern philosophers have completely jettisoned the theological
aspects of the Lockean project. The quasi-positivist approach remains popular,
however, with some ethicists who have developed the Lockean idea. (Peter
Singer, e.g., specifically quotes Locke for political purposes, to demand the
same protections for non-humans as given to humans. (1993, 87.)) They argue
that non-human animals deserve freedom from slaughter or indiscriminate use in
medical experimentation on the ground that they partake in the same qualities
that make humans “persons” and thus should be considered eligible for the
same civil protections as humans. Now, Locke would almost certainly have
supported this argument conditionally, for he wrote:

For were there a Monkey, or any other Creature to be found, that
had the use of Reason, to such a degree, as to be able to understand
general Signs, and to deduce Consequences about general Ideas, he
would no doubt be subject to Law, and, in that Sense, be a Man, how much soever he differ'd in Shape from others of that Name. (III.xi.§16, 517.)

The question of whether there is such a monkey (or a monkey who is self-conscious, as required by II.xxvii.9) is an empirical matter, of course; one that is the subject of much investigation. (See generally, Griffin (1992).) However, I feel sure Locke would have drawn an additional conclusion that animal rights activists avoid: that persons so constituted (even if non-human) are responsible agents. For he wrote,

Person, as I take it, 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. (§26, 346.)

There is much more here to be said on this topic, but it is really beyond the scope of the current project, so I will leave it for a more convenient time.

B. Identity

As Frege (Year) dryly observed in "On Sense and Reference" (Uber Sinn und Bedeutung), "[e]quality [by which he meant identity]\(^7\) gives rise to challenging questions which are not easy to answer." What questions might be difficult to answer? We could start with the problem of definition. As Perry and Bratman (1999) put it, "identity [is] that relation that obtains between each thing and itself and never obtains between a thing and anything else." (820.) This gives us some useful formal information—the identity relation is not only transitive and symmetric, it is reflexive. But it leaves the substantive question unanswered.

\(^7\) "I use this word in the sense of identity . . ." (56.)
What relation? Well, the obvious answer is, “sameness.” But again, not too helpful—sameness of what? The first major step towards understanding the identity relation comes with Leibniz’s answer to this question. Leibniz’s Law (the “indiscernability of identicals”) holds that identity is a relation of sameness of properties: if two numerically distinct objects are identical, then they have all their properties common. But now we start to get a glimpse of possible problems. For everything changes—indeed, if things didn’t change, we would have no interest in identity in the first place. It seems we can always find a way to deny the consequent of the conditional, with the result that there is no identity relation where we normally expect to find one. We wish to say (e.g.) that the Mississippi in ebb (“Me”) is identical to the Mississippi in flood (“Mf”). But how can they be one and the same thing if they differ in one property—viz., one is in ebb and the other is in flood? This seems like a problem; and induces Perry and Bratman (1999) to refer to what they call the “paradox of identity.” (820). The problem is only a “paradox” in the loose sense, however, because no reductio can be derived either from Mf ≠ Me or from its contradictory, Mf = Me. In the first case, there is simply nothing logically contradictory in denying identity through time. So Heraclitus thought, for he wrote:

8 Not to be confused with the identity of indiscernibles—“If two objects have all their properties common, then they are not two at all, but one”—which makes the substantive claim that all numerical distinct objects differ in some material property.

9 I take it that in order for statement A to be a true semantic paradox it must yield a double reductio; i.e., it must be possible to derive a contradiction from both Statement A and not-Statement A.
It is not possible to step twice into the same river, nor is it possible to touch a mortal substance twice in so far as its state (*hexis*) is concerned. But thanks to the swiftness and speed of change, it scatters things and brings them together . . . neither ‘again’ nor ‘later’ but *simultaneously*, it forms and it dissolves, and it approaches and departs. (1996, 55.)

In the second case, we can dissolve the apparent contradiction by the expedient of adding a time index to the properties of ebb and flood, so that both of them (ebb and flood) are properties of a single object: $M_{t-1}(\text{ebb}) = M_{t-2}(\text{flood})$.

Nevertheless, neither solution is entirely satisfactory. On the one hand, rejecting duration of objects is highly counterintuitive. On the other, time-indexing properties seems question begging. Something more in the way of explanation than an assumption of duration seems in order. An example will bring out the point. Consider the object “The occupants of Apartment B-4.” At $t_1$ they are Mr. and Mrs. Smith; at $t_2$ they are Mr. and Mrs. Brown; at $t_3$ they are Mr. and Mrs. Jones; at $t_4$ they are Mr. and Mrs. and Baby Jones; at $t_5$ they are Ed and Angela. Of course, the “occupants of Apartment B-4” is subject to change. But by time indexing the various inhabitants we can turn it into a single object which *always* has the five properties of being Mr. and Mrs. Smith at $t_1$, Mr. and Mrs. Brown at $t_2$, etc. As Odenberg (1993, 4) puts it, “[t]he problem with the notion [that Leibniz’s law provides a criterion of identity] is that community of property only obtains *in virtue* of identity’s obtaining and so cannot *itself* be a criterion of identity.” [Emphasis in original.] This brings out the important point that to be really helpful, Leibniz’s law must be backstopped by a theory of
objecthood, either general or particular.⁩

A further front in the war on Leibniz’s law was opened by Frege himself. As is well known, Frege was primarily concerned with the meaning of the identity relation—not with its metaphysics. He wondered how a relation with the formal properties described by Perry and Bratman could be both true and informative. He determined that the relation of identity is best thought of as being endowed with meaning in the context of discovery. That is, it is not a relation of names (in which case it is false), nor of objects (in which case it is uninformative), but rather of the modes of presentation of objects (which modes he called “senses”), which are public, abstract objects which constitute the meanings of words. We can realize that two distinct senses “pick out” the same object in the world; this realization is what makes the identity relation both true and informative.

Frege’s solution is not universally accepted. But it raises considerations which have convinced many philosophers that Leibniz’s view does not apply to properties in intensional (i.e., knowledge or belief) contexts. (Stoll (1967), 123.) A way of making this point clear is to try to substitute terms with identical referents into a belief context. I believe Mark Twain wrote Uncle Tom’s Cabin. Can we replace Mark Twain with Samuel Clemens and preserve the statement’s truth value? If not, it appears Samuel Clemens and Mark Twain differ in the

---

⁩There are several such theories (e.g., natural kinds, conceptualism, nominalism) in addition to Leibniz’s own theory of “monads,” on which it is unnecessary for us to dwell at length here.
property of “being believed by Ed to have written Uncle Tom’s Cabin.” But isn’t Mark Twain just a pseudonym for Samuel Clemens?

**Personal Identity**

A theatrical example will prove helpful here and show how the above concerns relate to the problem of personal identity. Arthur Miller likes to tell the story of meeting an old school chum (let’s call him Joe). Mr. Miller was standing on 47th Street in New York City, near the theater where his new play “Death of a Salesman” was playing to sellout crowds. Joe recognizes him and greets him, “Hey Artie—it’s been a long time! How ‘ya doing?” “Oh—hey Joe! What are you doing around here?” “Hacking for a living. I’m between fares. Say, you always wanted to write. Having any luck?” “Oh, a little bit,” Arthur Miller replied. “In fact, I have a play up right now.” “Oh yeah? Whereabouts?” Arthur Miller pointed at the nearby marquee, where his name was in all caps above the show’s title. Joe’s eyes and arms opened wide, and he said loudly in a tone of shock and surprise, “*Arthur Miller!!!*”

We may surmise that this response aptly demonstrates Frege’s point. For is it not obvious that Joe had made a discovery?—viz., the discovery that while the two names “Artie” and “*Arthur Miller!!!*” had divergent meanings—meanings available to anyone who knew Arthur Miller either when he was 15 or 40—that these two meanings pointed at not two entities, but a single entity, one and the same in diverse times and places? And this discovery was reflected perhaps not consciously but verifiably, in the form of a thought which took the form of the relation “Artie” = “*Arthur Miller!!!*” (Note that “Artie” and “Arthur Miller” differ in the property of “being known by Joe to have written ‘Death of a Salesman.’””)
Moreover, the example brings forth another (and for our purposes, the crucial) point. Why exactly does “Artie” = “Arthur Miller!!!”? It is this latter problem which Locke set out to solve in Book II, Chapter xxvii of his Essay.

**Relative Identity**

Given the limitations of Leibniz’s Law, perhaps it would be well to consider alternatives. One such alternative is the “relative identity” thesis. Relative identity is the view that there is not a single criterion of sameness (as Leibniz’s Law holds), but many, one for each sortal concept under which an object can fall. As Geach (who originated the recent controversy over the view and who has been its most vocal proponent) put it (1972, 238) identity “is an incomplete expression and it is short for ‘x is the same ‘a’ as y’, where ‘a’ represents some count noun understood from the context of utterance.” Identity and individuation are closely allied (as Geach points out, 238), we say “one and the same” (in German “*ein und dasselbe*”). And after Wiggins the thesis is usually divided into two: a thesis of sortal dependency of individuation (D) and a thesis of sortal dependance of identity (R):

(D): Given two contemporaneous apparent things (a and b) to both of which can be ascribed two different predicates (F and G) we will wonder whether they (assuming each one is one single thing) could be the same ‘F’ and not the same ‘G’ (or vice versa). (Wiggins (2001), 22.)

(R): Given two temporally distinct things (a and b) to both of which can be ascribed multiple predicates, the project of determining whether they are diachronically the same thing is different depending on which predicate they are considered under. We will thus wonder whether they can be the same ‘F’ and not the same ‘G’ (and vice versa). (Wiggins (2001), 23.)

Wiggins considers that (D) and (R) are logically distinct, since the answer
to (D) is always negative, but the answer to (R) is sometimes affirmative.
(Wiggins (2001), 24.) These three factors can be brought out by the following example: First, consider a forest of one hundred trees. We can call them “that forest” or “those one hundred trees.” If we wish to determine whether this thing, this forest (for the one hundred trees and the forest are one and the same thing) is the same as some other contemporaneous forest of one hundred trees (D), we compare them. Note that it is impossible for us to get different answers to the question, “Is this the same forest as that?” and “Is this the same one hundred trees as that?”

Now consider the same forest and one hundred trees again. This time our task is to compare them to some other diachronically distinct forest and one hundred trees. Could they be the same forest and not the same trees, or vice versa? Certainly, they could be. One hundred years from now, there could be different trees (imagine a fire destroying them all) but the same forest. Conceive of the trees all transplanted 1000 miles away into the midst of another forest; they would be part of that forest but still be the same trees.

The nice thing about the relative identity thesis is that you don’t have to have all the same properties to be the same thing; just those required by the sortal. Unfortunately, the tide of philosophical support has receded from the relativity of identity thesis practically since Geach first enunciated it, turning back towards Leibniz’s law. Wiggins asserts that all supposed examples of (R) actually depend on one of two kinds of errors: an ambiguity in the use of the copula or the deceptive use of certain referring expressions. (2001, 51-2.) In our example case, he would say that “the forest” and “the one hundred trees” do
not designate a single thing. (53.) It is not that we can distinguish them on the
ground that they can have separate fates and thus do not share all properties, as
Leibniz’s Law requires; rather, they have many distinct synchronic properties,
and the difference in fates results from these more fundamental differences. (A
forest is also the birds that nest in it, the underbrush, the stream that trickles
through it, the shadows these cast, etc.) Quine and many others have
articulated similar concerns; so much so that the relativity of identity thesis is
now (relatively) in abeyance.

Personal Individuation

Finally, we should give some consideration to the problem of
individuation. If Perry and Bratman are correct, a solution to the problem of
identity is also at least in part a solution to the problem of individuation, because
an individual is all those things with which we have the relation of identity, and
includes none of those things with which we do not have such a relation. But
without a theory of objecthood, we cannot know what an individual is any more
than we know what identity is.

This is probably clear enough as a general principle. Unfortunately, it may
not be immediately obvious when we speak of the individuation of persons,
because we so take our separateness for granted. But to properly understand
Locke, we need to give a little thought to the issue—since the normal criterion of
personal individuation (different body) is one he explicitly rejects.

An obvious solution nevertheless seems available. A person is—if
nothing else—a reflective consciousness, “I” to itself—and that “I” just simply is
different from and distinguishable from other “I’s”—if for no other reason than “I”
am considering the problem of individuation and other "I"s are not! Even if we reflect on the issue it does not seem terribly controversial. Each consciousness is eo ipso unique, therefore individual and distinguishable merely by acts of consciousness which can not be shared. Locke seemed to hold this view, for he wrote:

\[ \ldots \text{Since consciousness always accompanies thinking, and it is that makes everyone to be, what he calls self, and thereby distinguishes himself from all other thinking things, in this alone consists personal identity, i.e. the sameness of a rational Being \ldots} \]

(§9, 335.)

And if we feel the need to inquire further there will always be some more general criterion of individuation on which to fall back, e.g., Locke’s claim that:

From whence it follows that one thing cannot have two beginnings of Existence, nor two things one beginning, it being impossible for two things of the same kind, to be or exist in the same instant, in the very same place; or one and the same thing in different places. (§1, 328.)

This seems patently self evident, and in accordance with the very nature of thought. 11 Despite these dictates of common sense, I think we can motivate personal individuation as a problem in three ways: (a) technical concerns related to the individuation criteria themselves; (b) science fiction; and (c) real life issues.

(a) Technical concerns. We must consider first a charge of circularity. Consciousness has two aspects—that of subject and that of object. Now if “consciousness” for Locke is that of subject, we may be on safe ground. But if it is that of object, a charge of circularity arises, based on the following reasonable

\[ \ldots \text{[Y]et we cannot conceive but that they must necessarily each of them exclude any of the same kind out of the same place: Or else the Notions and Names of Identity and Diversity would be in vain, and there could be no such distinction of Substances, or any thing else one from another.”} \]

(§2, 329.) A very Kantian observation.

\[ 11 \text{“} \ldots \text{[Y]et we cannot conceive but that they must necessarily each of them exclude any of the same kind out of the same place: Or else the Notions and Names of Identity and Diversity would be in vain, and there could be no such distinction of Substances, or any thing else one from another.”} \]
principle: no object can be used as a criterion of individuation for some other object unless it itself has a clear criterion of individuation. Locke does not supply such a clear criterion, and so his account of individuation is circular. In "Locke’s Theory of Personal Identity," M. Atherton (1983) argues that Locke really has a "same consciousness" theory of personal identity. In her view, the consciousness in question in Locke’s theory is not that of subject but rather object (what she calls "plain consciousness"). (277-8.) Her approach to the individuation problem is to avoid circularity by referring each individual consciousness to its origin. But this criterion plainly has the same problem that the consciousness as object criterion does: for each "time and place" needs to be individuated from every other "time and place." It is not at all clear how that can be done, and Atherton does not even attempt the task. (Indeed, in Leibniz’s view specific "times and places" are individuated just by the presence or absence of certain objects from those locations. On this approach the criterion is irretrievably circular.) My point here is not directly to assert that Locke’s theory is circular. It arguably avoids the Sargent objection because for Locke, consciousness is that of subject, not object. (Later I will resuscitate the circularity objection on other grounds, however. See infra, p. 71.) Rather, it is to indicate that circularity is a pitfall waiting for the unwary interpreter who fails to keep the importance of individuation criteria in mind; it is a possible consideration that bears on the question of the interpretation of Locke’s intention in his use of

\[12\] “To Sergeant, consciousness refers to the self as the object in which acts of consciousness originate; therefore, the self must be an individual prior to the occurrence of acts of consciousness . . . .” (Thiel (1998b), 898.)
the term “consciousness”.

(b) *Science Fiction Examples*. There are some fictions that might raise questions about our pure reflective intuition that “one consciousness, one person.” They might be impossible—but they are not inaccessible and perhaps they bear consideration.

(1) *The Borg*. For example, what are we to make of the “Borg”—the collective with many members each of which is conscious at any particular time of what all the others are doing? The Borg appears to be made up of (what I’ll call) “co-temporal person parts,” and the question is whether the collective so constituted is also a person. I think we should say that the Borg is indeed a person, and on this ground—there is no prior person state of any of its co-temporal parts which is not available to the consciousness of all other parts of the Borg.

This leads to a criterion of individuation (the “person-part criterion”) as follows: given a group of presumptive person parts which both have individual consciousness and consciousness in common, the collective is a person iff there is no prior existing person (or part person) to which the consciousness of all the co-temporal person parts cannot reach. In other words, given a collective consciousness $x_0$ with person parts $x_{1-n}$ and $R = \text{“remembering being,”}$ then:

**Part-Person Criterion**

1. $(\forall x)(Px_0 \rightarrow (\exists y)(Py \& (x_m Ry \& \neg x_n Ry)))$

(For more on the “remembering being” criterion, see infra, p. 36.)

(2) *Vulcan mind meld*. On the above criterion, the Borg is a person. What about the Vulcan mind meld, in which two individuals share consciousness but
only for a period? One approach is to say that they are not a person because there is a *future* 'w' (or part person) to which one of the 'xₙ,s' will not have access. If we think this, we must adjust the part-person criterion to include either remembering or becoming. But it seems more reasonable simply to claim that the mind-meldees are one person at the time of mind meld, but separate both before and after.

(3) *Teletransportation.* The teletransporter case raises questions about the coherence of the Locke criterion because it apparently violates a transitivity criterion. (This will be more fully explicated, *infra*, p. 38.)

(c) *Real Life.* In real life, we do not run into problems of collective consciousness. What we do run into, however, is problems of collective unconsciousness. That is to say, personhood can be lost because individual consciousness is submerged in group action. Konrad Lorenz gives a frightening example:

I was once an unwilling witness of this sudden emergence and rapidly snowballing effect of this process of dehumanization, and if I was not drawn into its vortex it was only because, thanks to my knowledge of flocking behavior, I had seen the approaching danger sooner than others and had time to guard against my own reactions. To me there is small pride in the memory; on the contrary, no one can put much trust in his own self mastery who has ever seen men more courageous than himself, men fundamentally disciplined and self controlled, rushing blindly along, closely huddled, all in the same direction, with eyes protruding, chests heaving, and trampling underfoot everything that comes in their way, exactly like stampeding ungulates, and no more accessible to reason than they.¹³ (1963, 142-3.)

Lorenz calls this process "dehumanization" because it is for him an atavism—a

---

¹³ For Lorenz, the opposite of mass behavior is not individuated consciousness but personal relationship. To my knowledge the philosophical implications of this position remain yet to be explored.
recurrence to a state (flocking or schooling) that while normal for many lower
animals and always a possibility for us, is antithetical to "normal" human
functioning which we think of as individuated. Or consider these horrifying
vignettes from *Survival in Auschwitz* (Levi (1996)):

> He is Null Achtzehn. He is not called anything except that, Zero
> Eighteen, the last three figures of his entry number; as if everyone was
> aware that only a man is worthy of a name, and that Null Achtzehn is
> no longer a man. I think that even he has forgotten his name, certainly
> he acts as if this was so. When he speaks, when he looks around, he
gives the impression of being empty inside, nothing more than an
involucre, like the slough of certain insects which one finds on the
banks of swamps. . . . He is indifferent to the point of not even troubling
to avoid tiredness and blows or to search for food. He carries out all
the orders that he is given, and it is foreseeable that when they send
him to his death he will go with the same total indifference. (42-3.)

> On the other side of the road a steam shovel is working. Its mouth,
hanging from its cables, opens wide its steel jaws . . . then rushes upon
the soft clayey soil and snaps it up voraciously. . . . Then it rises, turns
half around, vomits backwards its mouthful and begins again. Leaning
on our shovels, we stop to watch, fascinated. At every bite of its mouth
our mouths also open, our Adam's apples dance up and down . . . we
are unable to tear ourselves from the sight of the steam shovel's meal.
(74.)

The issue seems of concern mainly to social scientists, but it is not their
exclusive concern. Some philosophers will not consider it a virtue of a theory of
personal identity that it does not (as Locke's does not) require us to consider
such entities as Null Achtzehn "persons" at the relevant time. They will say: "If
they are not persons, they have no secure rights, they may be treated any-which-
way with impunity." But if they are persons, how can we conceive of that which
seems to have turned them into less than persons as *eo ipso* a crime? Locke
might say (see *infra*, p. 27) that the example entities are the same man, but not
the same person—and something like this distinction seems worth preserving.
As Strawson (1991, 113) puts it:

When we think of such cases, we see that we ourselves, over a part of our social lives—not, I am thankful to say, a large part—do operate conceptual schemes in which the idea of the individual person has no place, in which its place is taken, so to speak, by that of a group. But might we not think of communities or groups such that this part of the lives of its members was the dominant part—or perhaps was the whole? It sometimes happens, with groups of human beings, that, as we say, their members think, feel and act "as one." The point I wish to make is that a condition for the existence, the use, of the concept of an individual person is that this should happen only sometimes.

More needs to be said about what goes on during that "sometimes," but it must await a more convenient time.

II. The Mainstream View

Having considered the significance of the concepts of "person" and "identity," we are now in a position to examine Locke's theory of personal identity. We will start with a brief overview which reflects a consensus on the part of scholars as to the main elements of the theory, the "mainstream view."

On the mainstream view, Locke's theory of personal identity has five aspects: (a) relative identity; (b) the concept of a person; (c) the criteria of personal identity and individuation; (d) a non-substantial view of persons; and (e) the distinction between the concepts of person and man. It is as follows:

(a) Locke espouses what is now called the relative identity thesis. That is, what constitutes criteria of identification of objects over time depends on the type (concept or sortal) under which the objects falls.

'Tis not therefore Unity of Substance that comprehends all sorts of Identity, or will determine it in every Case: But to conceive, and judge of it aright, we must consider what Idea the Word it is applied to stands
for: It being one thing to be the same *Substance*, another the same *Man*, and a third the same *Person*, if *Person*, *Man* and *Substance* are three Names standing for three different *Idea*s; for such is the *Idea* belonging to that Name, such must be the Identity . . . . (§7, 332.)

(b) It follows from this that to know what constitutes reidentification criteria for the person, we must have a clear idea of what a person is. For Locke, a person is a self-conscious, thinking thing that has a sense of its own continuity in time; that is, knows it has a past and a future. He writes:

This being premised to find wherein *personal identity* consists, we must consider what *Person* stands for; which, I think, is 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: It being impossible for any one to perceive, without perceiving, that he does perceive. (§9, 335.)

This can be put into logical notational more or less as follows:

Definition of Person

II. $(\forall x)(Px \equiv (Cx \& x_{t_2}R_{x_{t_1}}))$

where $t_2 > t_1$, $Px = x$ is a person, $Cx = x$ is conscious, $xRy = x$ remembers being $y$. The "remember being" criterion is not just any old remembering. It is the memory of actions, thoughts, perceptions and proprioception (*i.e.*, experience of having a body) which I present to myself as having formerly been my own.

(c) From (a) and (b), it follows that two diachronically distinct persons are the same person ("$=P=$") if and only if the later one remembers being the earlier one, as follows:

Personal Identity

III. $(\forall x)(\forall y)(((Px \& Py) \& t(x) > t(y)) \circ (x=P=y \equiv xRy))$

In addition, Locke provides a criterion of individuation. It holds that if
something has two beginnings, it is two things, and if it is two things, it has two beginnings. Such beginnings are distinguished by spatio/temporal location:

From whence it follows, that one thing cannot have two beginnings of Existence, nor two things one beginning, it being impossible for two things of the same kind, to be or exist in the same instant, in the very same place; or one and the same thing in different places. That therefore that had one beginning is the same thing, and that which had a different beginning in time and place from that, is not the same but divers. (§1, 328.)

Secondly, Finite Spirits having had each its determinate time and place of beginning to exist, the relation to that time and place will always determine to each of them its Identity as long as it exists. (§2, 329.)

(d) According to Locke, we cannot identify the continuity of consciousness which makes personal identity with continuity of immaterial substance, or "soul":

I say, in all these cases, our consciousness being interrupted, and we losing the sight of our past selves, doubts are raised whether we are the same thinking thing i.e., the same substance or no. Which however reasonable, or unreasonable, concerns not personal Identity at all. The Question being what makes the same Person, and not whether it be the same Identical Substance, which always thinks in the same Person, which in this case matters not at all. Different Substances, by the same consciousness (where they do partake in it) being united into one Person; as well as different Bodies, by the same Life are united into one Animal, whose Identity is preserved, in that change of Substances, by the unity of one continued Life. For it being the same consciousness that makes a Man be himself, personal identity depends on that only whether it be annexed only to one individual Substance or can be continued in a succession of several Substances. (§10, 336)

making the point that unity of substance does not make personal identity; and,

For as to this point of being the same self, it matters not whether this present self be made up of the same or other Substances, I being as much concern'd, and as justly accountable for any Action was done a thousand Years since, appropriated to me now by this self consciousness, as I am, for what I did the last moment. (§16, 341.)

making the point that one can have personal identity even without unity of
substance.

(e) Locke carefully distinguishes between the concepts of “person” and “man.” Whereas a “person” is a conscious, thinking thing that has a sense of its existence in time, a man is a rational creature of a certain form:

   And whatever is talked of other definitions, ingenious observation puts it past doubt, that the *Idea* in our Minds, of which the Sound *Man* in our Mouths is the Sign, is nothing else but of an Animal of such a certain Form: Since I think I may be confident, that whoever should see a Creature of his own Shape and Make, though it had no more reason all its Life, than a *Cat* or a *Parrot*, would call him still a *Man* or whoever should hear a *Cat* or a *Parrot* discourse, reason, and philosophize, would call or think it nothing but a *Cat* or a *Parrot*; and say, the one was a dull irrational *Man*, and the other a very intelligent rational *Parrot*. (§8, 333.)

and in many puzzle cases (e.g., “Nestor,” §14) he makes the point that the same consciousness in different bodies makes the same person and different men, while distinct consciousnesses in the same body make different persons and the same man (e.g., day/night man, §23, 344).

III. Commentary

Below are fuller commentaries on all five of the aspects of the mainstream view.

A. Relative Identity

Locke evinces support for the relative identity thesis in the strongest possible terms, both at the beginning and at the end of Chapter XXVII. Indeed, his last words of the Chapter are:

   For whatever be the composition whereof the complex *Idea* is made, whenever Existence makes it one particular thing under any
denomination, the same Existence continued, preserves it the same individual under the same denomination. (§29, 348.)

Moreover, as Thiel (1998a, 235-7) points out, both Boyle and Hobbes espoused the relativity of identity criteria to sortals. Nevertheless, although the relative identity reading of Chapter 27 may be considered standard,\(^\text{14}\) whether or not it actually reflects Locke's intentions has been a matter of some controversy. One problem (see supra, p. 16) is that relative identity is probably not a coherent theory. The principle of charity at least suggests that we should be open to alternative readings if they do justice to the letter and spirit of Locke's text.

There are, indeed, actually three ways of interpreting Locke's position: (a) The relative identity thesis; (b) the "limited commitment" view; and (c) the relative individuation (multiple entities) thesis.

Let us take (b) first. Under the limited commitment interpretation, Locke commits himself to no more relativity of identity than of the categories be actually names; not to a total relativity of identity for all concepts. As Thiel (id., 239) puts it, "On this reading, Locke's account of identity is merely nominalistic in tone, not in content." Locke seems to presuppose only one category which fits men, animals and plants, viz., life (positive thesis). Moreover, there is no suggestion in Locke's text that variations in our concepts of man would mean a change in our identity criteria, as we might infer if the relative identity view is correct (negative thesis).

Thiel (id., 239-40) rebuts both of these theses, pointing out (a) that Locke apparently distinguishes among different types of life, and (b) that some of__________

\(^{14}\) It is most recently defended by Stuart (2004).
Locke’s text does suggest that our ideas determine our criteria of identity, *e.g.*, “. . . supposing a rational Spirit be the *Idea of a Man*, then the *same Spirit*, whether separate or in a Body will be the *same Man*.” (*Id.*, 240.)

For the moment, we may consider his “limited commitment” view refuted; although in response to other problems with the relative identity view I will resurrect a version of it later (the causal identity theory) which is immune to these objections.

A second alternative reading is what we call the “relative individuation” (*id.*, 241) or “multiple entities” view. As articulated by Vere Chappell (1989), this interpretation holds that multiple distinct entities (each with their own individuation criteria) can coexist in the same place. Chappell’s argument may seem odd, but it is hard to refute (at least with regard to Locke’s early examples of sortals under which objects are identified). Briefly, his claim is that if we take, *e.g.*, a horse and the matter of the horse, they are not the same single thing because they don’t have the same beginnings (for Locke, the basic principle on which objects are individuated). Nevertheless, they can coexist because they are not of the same kind. Although this usage seems odd, Chappell does not argue that it is good metaphysics (although he thinks so). He only argues this is the way Locke himself viewed the example. “I do not think that the doctrine of double existence is bizarre . . . here my concern is only to establish what Locke thought.” (*Id.*, 75.)

Matthew Stuart (2004) points out (following Hoffman (1980) that the “one thing . . . one beginning” principle does not follow from the principle “no thing can exist at two places at the same time,” as Locke claims it does in §1. What does
follow is that one thing cannot have two simultaneous beginnings. Stuart also
objects that Locke's comments in II.xxvii.2,

... yet we cannot conceive but that they must necessarily each of
them exclude any of the same kind out of the same place: Or else the
Notions and Names of Identity and Diversity would be in vain, and there
could be no such distinction of Substances, or any thing else one from
another.

preclude any interpretation which holds that two substances can be co-located.

As to the first objection, although it seems correct, Chappell's argument is
immune to it. The reason is that Chappell is only claiming that Locke thought of
the situation in the way he describes: he is not asserting that his interpretation is
the most coherent theory. Since the non-sequitur is far from obvious and the
resulting principle anti-commonsensical, we have no reason to think it reflects
Locke's thought as he was writing.

The second concern is more troublesome, and points to a disagreement
on the interpretation of the relation of paragraphs one and two of §2. One way
out of the conflict is to accept the Uzgalis (1990, 287) reading, according to
which organisms are compound modes. Stuart rejects this reading on the basis

Stuart's critique here has some merit, but it does not seem completely to
determine the issue. So we may conclude that (at least insofar as the relation of
matter and living things is concerned) the Chappell reading remains a possible, if
perhaps somewhat less plausible, interpretation of Locke's text. However,

15 The point (for a relative identity theorist) is that one thing can have
multiple non-simultaneous beginnings corresponding to the various sortals under
which it fails.
Chappell’s argument has less appeal with regard to the (for our purposes) more important distinctions between persons and men and persons and souls. He himself admits that the earlier argument will not suffice, since although men and persons may have different beginnings, Locke’s position was that in the real world they are always coincident:

\[ \ldots \text{[I]t may be conceivable that the person a prince is should be associated with two different men at different times, but nothing like this ever occurs in actual fact. This means, then, that persons and their associated men do not in fact have different “beginnings of existence.” So it looks as if our main reason for distinguishing them, regarding them as two distinct entities, is lost. (1989, 79.)} \]

Therefore, he relies on a modal version (\(P^l\)) of Locke’s principium individuationis: “\( \ldots x \) and \( y \) are distinct entities if it is possible for one of them to have, or to have had, a different beginning of existence from the other.” (1989, 79.) I do not think Chappell’s argument (80) that it is plausible to suppose that Locke would have accepted this principle is very strong. It goes in part as follows:

It is hard to imagine any philosopher rejecting this principle, least of all Locke, who explicitly connected the identities of things with their spatio-temporal positions. To do so he would have had to countenance cases in which an \( x \) and a \( y \) are identical and yet could have begun to exist at different times. This hardly seems intelligible. (1989, 79.)

As to the first part of Professor Chappell’s claim, I do not think it means much to say Locke “explicitly connected” things with positions. Unless we wish to go so far as to say he was an essentialist about position, a highly dubious (and in light of his anti-essentialism of III.vi.4 unsupported) reading. As to the second part, while the conclusion may well be false, it does not seem unintelligible, and there is no reason why Locke might not have viewed the
matter in just this way; *i.e.*, he might have been a contingent identity theorist.

It is true that two things which are identical are (by a theorem of modal logic) necessarily identical.\(^{16}\) But it doesn’t follow from this and from his connection of things to their spatio-temporal beginnings that he would have believed it impossible for these things to be identical and begin in different times. He might have thought (incorrectly) that it was possible for two things that are in fact identical to be non-identical. In that case, even though a person could possibly begin at a different time from the man, s/he could still be the same thing as him/her (as the relative identity theorist would have it).

Undoubtedly, if being a relative theorist requires one to take identity statements as contingent that is a good reason not to be one. But since Professor Chappell claims to be concerned with what Locke thought, not what he should have thought, it is not a good reason to reject the thesis that Locke articulated a relative identity view.

Moreover, it does not seem to be the case that “Locke is committed to the distinctness of men and persons, quite apart from *Pl.*” (Chappell (1989), 80.) His concern is that if one thing can’t have two endings, and man and person are a single thing, then we can’t have a general resurrection of persons. This conclusion is based on the same faulty reasoning which led to Locke’s conclusion that one thing can not have *two beginnings, i.e.*, that one thing can’t have two endings. While the fact that Locke made a non-sequitur does not contradict the claim that he intended his thought to be interpreted in a certain

\(^{16}\) See Kripke (1980), 3.
way, it does undermine claims related to that to which Locke's position actually commits him. If Stuart and Hoffman are right, then all that Locke is committed to is that one thing can't have two simultaneous endings—which is at least as supportive of the notion of Christian after-life as Chappell's own quasi-Cartesian reading.

Finally, we should note that the conclusion of Chappell's argument that men and persons are distinct entities (even if Locke would accept it) is dependent on the claim that "nearly all the man-person puzzle cases he presents in Chapter 27" (1989, 79) presuppose that it is possible for men and their corresponding persons to begin in diverse times and places. Unfortunately, it is not the presupposition of the key cobbler/prince case. In this case, we are to understand the Prince qua both person and man and the Cobbler qua both person and man each as having respective non-simultaneous beginnings and endings.

We now find ourselves in a bit of a quandary. Both the "limited commitment" and "relative individuation" views seem to be substantively inadequate—they do not handle all the cases. This failure is disappointing, because we should have doubts about the formal adequacy of the relative identity interpretation (it seems right to say, with Wiggins, that the "person" and the "man" cannot be a single thing—they have do not have all their properties in common. Moreover, infra, p. 71 I will raise an additional objection—the circularity objection—that renders the relative identity interpretation totally untenable. But in the meantime it seems the best of a weak group of interpretive
options.

B. The Concept of a Person

The mainstream view of the concept of a person is correct as far as it goes. However, in my view, it is rather superficial and does not fully capture the nuances of Locke's thought. Here is the main definition:

This being premised to find wherein personal identity consists, we must consider what Person stands for; which, I think, is 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: It being impossible for any one to perceive, without perceiving, that he does perceive. . . . And by this every one is to himself, that which he calls self: It not being considered in this case, whether the same self be continued in the same, or divers Substances. (§9, 335.)

But though the same immaterial Substance, or Soul does not alone, where-ever it be, and in whatsoever State, make the same Man; yet 'tis plain consciousness, as far as ever it can be extended, should it be to Ages past, unites Existences, and Actions, very remote in time, into the same Person, as it does the Existence and Actions of the immediately preceding moment: So that what ever has the consciousness of present and past Actions, is the same Person to whom they belong. (§16, 340.)

Clearly, a person is a conscious rational being, as the mainstream view holds. And consciousness is not here "plain consciousness", as Atherton (1983, 227-8) holds—"consciousness" being that by which everyone "perceives that he does perceive."17 (§9, 335.)

The main problem, though, is that the mainstream view fails to take account of the diachronic element in the definition. As Locke says, a person is

17 Atherton, makes this claim virtually without textual support. This and many other otiose observations, make it clear she is revising Locke, not interpreting him—despite her assertions to the contrary.
that which "can consider it self as it self, the same thinking thing in different times and places." (§9, 335.) Prima facie, in order for there to be a person x at any time, there must also be a person y at another time whom person x can consider himself to be the same as.

Commentators tend to pass over this point or conflate the definition of person with that of personal identity.\(^{18}\) For example, Winkler writes:

A person, Locke believes, is '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' (E.II.xxvii.9:335). Personal identity extends as far as consciousness: a later person is the same as an earlier one just because the later person is conscious of some thought or action of the earlier." (1998, 149.)

Or Thiel:

The self as person is constituted by consciousness unifying thoughts and actions: "That with which the consciousness of this present thinking thing can join it self, makes the same Person." (1998b, 890.)

I have not made a full literature review, but given these comments it seems one may at least say the observation has not been made very clearly, if at all. (The reason for insisting on this point will be made clear when we discuss the circularity objection, infra, p.71.) A "person" considered at an isolated moment is not a person at all for Locke, but a mere "person stage."

An additional point is whether we should (contra Locke) allow for the possibility of spatially dispersed persons. Such consideration perhaps seems anachronistic. Nevertheless, I think it important that we do so. In part this is

\(^{18}\) Brody (1972, 332) does not even attempt to articulate Locke's view of person separately from his definition of "same person."
because it would correct an apparent oversight but—more importantly—because it allows us to add an explicit individuation condition to our identity definition (see next section). It will also provide us some Lockean machinery with which to deal with the Parfit transitivity objection (see infra, p. 42).

Finally, we may ask if the concept of consciousness is redundant in the definition of person, being subsumed under that of remembering—it being impossible to remember if not conscious. As we will see (infra, p. 44) remembering here is of a special type, viz., possible remembering. Since I see no easy way to distinguish the possible remembering I have when I am awake (when I am a person, according to Locke) from that I have when I am asleep (when I am not a person), the simplest way to do so is indirectly, by adding consciousness as a separate condition. This leaves us with the following definition of a person:

Definition of Person—Diachronic

IIa. \((\forall x)(Px_0 \equiv (Cx \& \neg(\exists y)(x_m Ry \& \neg x_m Ry)) \& (\exists z)(Pz \& x_0 = P = z)))\)

C. Personal Identity

A concept of identity should meet three formal criteria: it should be reflexive, symmetrical and transitive. We should consider how our theory meets (or may be adapted to meet) these formal demands, and whether such meeting has any substantive repercussions. In addition, the “remember being” criterion is unclear. We shall have to consider what precisely it might mean.

Reflexivity

What is reflexivity here? One might say, \(x = P = x\); but while this is true it is
rather uninformative. Moreover, since this is relative identity, reflexivity does not by itself provide us with a criterion of individuation; we need to know what property matters for the sortal (unlike for "="; where variation in any property individuates objects).

The property that seems to matter is having memories of formerly being persons. Two contemporaneous individuals are the same iff they have all the same memories of being persons, i.e.:

Personal Identity—Synchronic

IV. (∀x)(∀y)((t(x)=t(y) & x=P=y) ⊃ (∀z)(Pz & t(z)<t(x)) ⊃ (xRz ≡ yRz))

Symmetry

The mainstream view does not take account of the formal need for symmetry in the criterion of personal identity. That need can be met by making the remembering criterion two directional; that is, if x is diachronically the same person as y then either x remembers y or y remembers x.

Personal Identity—Diachronic

Ilia. (∀x)(∀y)((t(x)>t(y) & x=P=y) ⊃ (xRy v yRx))

This not only fulfills a formal demand, but also forestalls the objection that the definition of person is temporally recursive. If that were the case, it would be hard to answer the question "what exists before the first person in a series?" We would be (perhaps) unwittingly committing ourselves to a doctrine of eternal pre-existence, inconsistent with Locke's thought. Now the definition is still "recursive" in the sense that for there to be one person there must perforce be two; but this is no more than to say a person is only a person if s/he persists through time. This claim may seem odd but is not without precedent since the
same claim is usually made for "substance."

We can put the assertions of this and the previous section together to yield the following definition of =P=:

\[
\text{Combined Criterion IIIa & IV}
\]

\[
V. \forall x)(\forall y)((Px \& Py) \Rightarrow (x=Py) \equiv ((t(x) \& t(y) \Rightarrow (xRy v yRx)) v (t(x)=t(y) \Rightarrow (\forall z)(Pz \Rightarrow (xRz \equiv yRz))))
\]

**Transitivity**

Our memory criterion for “same person” is thus both symmetrical, and (with the addition of the individuation criterion) reflexive. Is it transitive? Apparently not; since while xRy and yRz, it may be that \(\sim xRz\). This problem was early noted (perhaps first in print by Leibniz in the *New Essays*) and often repeated since first publicly enunciated by Berkeley.\(^{19}\) Its canonical form is due to Reid, and goes as follows:

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 having been flogged at school, and that, when made a general, he was conscious of his having taken the standard, but had absolutely lost the consciousness of his flogging.

These things being supposed, it follows, from Mr. Locke’s doctrine, 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 so far back 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 (2001), 347.)

\(^{19}\) Jolley (1984, 128); see Berkeley (1948-51, 298-99.)
One can hardly think of a better solution than the one advanced by Leibniz (1997, 236) to the effect that we should simply deem ourselves to be the same person as all persons who we are the same as are the same as; so that if xRy and yRz, x=P=z whether or not s/he “actually” remembers being z.\(^{20}\) Mackie objects to this amendment, as follows: “[T]his is a revision, not an interpretation, of Locke’s account. Not only does he not say this, he commits himself explicitly to a different view.” (Mackie (1976), 181.)

Perhaps Locke thinks he has to bite this bullet, but I believe he is wrong; and Mackie is wrong in saying that Locke is *committed* to non-transitivity of personal identity. The Leibniz solution offered above is available to Locke, although perhaps he id not realize that; and it is consistent with his theory, while the text from which we can infer that he rejected it (§20, 342) is not consistent with his theory.

There are two reasons for this, corresponding to two types of cases. One is that of “distinct incommunicable consciousness” (Diagram A) and total amnesia (Diagram B). In these types of cases, there is no person whom I remember being who is the same as any person whom I do *not* remember being. (See Diagrams A and B.) In Diagram A, no person whom x remembers being \(t_1-t_2, t_3-t_4, t_5-t_6\) is the same as any person y remembers being \(t_6-t_7, t_4-t_5, t_2-t_3\); in Diagram B, no person after \(t_1\) has any memory of any person *before* \(t_1\). So transitivity is not violated.

Cases of the second type (Diagram C) involve partial amnesia. At \(t_3\) one

\(^{20}\) For a theory of personal identity based on transitivity of memory, see Grice (1941), 330-50.
might remember the period $t_1$-$t_2$, but at $t_4$ one might not, since at $t_4$ x remembers $t_3$, x is the same person as y and y (as y remembers being z) is the same person as z; but $\neg x = p = z$, since x does not remember z; so transitivity apparently is violated. However, it should be noted that we are here speaking of possible memory (in some yet to be determined sense of “possible”—see the next section). Otherwise, we would not even be the same person as the someone who was sitting where we are reading the previous sentence of this text. For Leibniz, who believed that traces of all memories remained as petites perceptions, it was a matter of principle that no memory could be completely lost.

[I]t is unreasonable to suppose that memory should be lost beyond any possibility of recovery, since insensible perceptions, whose usefulness I have shown in so may other important connections, serve a purpose here too—preserving the seeds of memory. (Leibniz (1997), 239.)

So for Leibniz, the transitivity move presented no problem; transitivity could not be violated if in principle a case like that in Diagram C could not happen. And we can take a similar approach. We can take the position that as a contingent matter of fact no memory is completely lost but only temporarily lost. The fact that various stimuli, from objects to electrical stimulation of the brain, can restore to us past memories we might have thought we’d forgotten supports this view, although of course it cannot prove it.

Mackie however, claims that this position is not available to Locke:

“[T]hough he [Locke] does not explicitly recognize this, he is committed to giving up the transitivity of personal identity.” (Mackie (1976), 182.) He cites two pertinent passages in support of this claim. One of them is from §24:
For whatsoever any Substance has thought or done, which I cannot recollect, and by my consciousness make my own Thought and Action, it will no more belong to me, whether a part of me thought or did it, than if it had been thought or done by any other immaterial Being any where existing. (345.)

This supports Mackie’s view only on assumption it is possible to forget. But our conjecture denies this assumption. The other text has a little more bite. It is a quote from §20:

But yet possibly it will still be objected, suppose I wholly lose the memory of some parts of my Life, beyond the possibility of retrieving them, so that perhaps I shall never be conscious of them again; yet am I not the same Person, that did those Actions, had those Thoughts, that I was once conscious of, though I have now forgot them? To which I answer, that we must here take notice what the word I is applied to, which in this case is the Man only. And the same Man being presumed to be the same Person, I is easily here supposed to stand for the same Person. (342.)

Now, this text does provide some support for the Mackie position, but only thus far: it shows that Locke was not willing explicitly to state that memory is ineradicable; for if he thought so he could have said so here. He might even have thought that memory loss is possible. But I don’t think this is enough to

21 He clearly thought that God could retrieve memories at the Last Judgment—

[A]t the Great Day, when every one shall receive according to his doings, the secrets of all Hearts shall be laid open.* The Sentence shall be justified by the consciousness all Persons shall have, that they themselves in what Bodies soever they appear, or what Substances soever that consciousness adheres to, are the same, that committed those Actions, and deserve that Punishment for them. (§26, 347.)

However, I am unwilling to give up on the usefulness of the memory criterion for judgments of culpability while one is alive, so I’m not depending solely on this text to make the point.

*1 Cor. 14:25 and 2 Cor. 5:10.
afford the inference that Locke is "committed" to the view that memory is eradicable. My reasoning is as follows: Locke is committed only to the logical consequences of his theory. Although he may think that his comment at §20 follows from the rest of his theory, in fact it does not; it is an error. So Locke is not committed to this position (although he did enunciate it), nor to any future inferences that supposedly follow from it. The reason it is an error is that the term "I" does \textit{not} refer to the "man". When "I" use the indexical "I", I am referring to \textit{myself} as a thinking thing—which is just what Locke means by "person." (See "The Circularity Objection," \textit{infra} p. 71.) So when I refer to "I" in the past, I am also referring to myself—as person—on pain of amphibole.\textsuperscript{22} Moreover, it must be so. Because on Locke's theory only persons can act—"person," not "man," is the "forensic" concept. Nowhere else in the Chapter does Locke refer to "man" as a concept denoting a responsible agent, and he is quite right not to do so, since that would largely (in practice \textit{completely}) collapse the distinction between "man" and "person." (See "The Concept of Man," \textit{infra} p. 27 for an explanation of why this is so.)

I do think the §20 text is evidence Locke did not think of Leibniz's solution to transitivity Diagram C. But I also think that text is inconsistent with Locke's theory as otherwise enunciated and separable from it. It is reasonable to think that he would have accepted the Leibniz suggestion as "friendly" and used it instead of the §20 make-shift had it been offered to him.

Parfit (1984, 199-200) offers an important variation on the transitivity

\textsuperscript{22} \textit{Cf.} Wiggins's complaint that many supposed cases of relative identity are cases of ambiguous use of referring expressions.
objection, what he calls the “branch-line case” (Diagram D). At $t_2$ a person enters a teletransporter which recreates his every physical state in the body of a new person on Mars. However, although normally the original body is destroyed in this procedure, on this one occasion it survives. It ($y$) still has all the memories, etc. which constitute it the same person as ($x$), as does her counterpart on Mars ($z$). Are $y$ and $z$ the same person?

Note they are both the same as $x$; and so by transitivity, it appears they are. Since we have agreed to entertain the possibility of dispersed persons, why not simply accept that $y$ and $z$ are the same? Then they are $x_1$ and $x_2$, person parts of the same person ($x_{1,2}$). The problem is that $y$ and $z$ have different memories of being persons. More specifically, $(\exists w)(x_2 Rw \land \lnot x_1 Rw)$; so (according to our contemporaneous personal identity criterion) $x_{1,n}$ is not a single person. Does this not violate transitivity? Yes, it does. However, Locke can block the example. His theory is by intention “topic-neutral” (see infra, p. 63). That is, it withholds judgment on the materiality or immateriality of the soul. But the example only works if the soul is material. The 17th century view was that immaterial things are indivisible. (Cf. Descartes’s holding that the “real distinction” between the mind and body is that the “body is by its nature divisible while mind is indivisible.” (Descartes (1995), 330.) If there are two separate persons $x$ and $y$, then there are two separate souls—but how did one immaterial soul become two? That is not possible. So Locke can say the example has as an essential premise an assumption he rejects—namely, the materiality of the
soul. I conclude that the transitivity problem does not present an insuperable obstacle to acceptance of Locke's theory of personal identity.

In addition to the concern over whether the remember being "xRy" criterion is transitive, there are also a variety of issues it raises which need clarification. We can think of these as occurring along three axes: possible vs. experienced memory (\^R vs. \*R); corrected vs. uncorrected memory (R^c vs. R^n); total vs. partial memory (R^t vs. R^p). The first raises the question of whether we can be persons when we are not remembering; the second raises questions of whether the criteria for having a memory are partly internal, and or to some extent external; and the third raises the question of how much of a past self's experience we need to recall to be considered the same person as that earlier person.

Possible Memory vs. Experienced Memory (\^R vs \*R)

Unless we allow that memory is in some sense *in potentia*, we are not the same person as any prior person, even the person we were a few seconds ago—unless we are actually remembering being that person. Moreover, by our definition of person, we are not even a person *at all* unless we are remembering; for it requires that for x to be a person, (\exists y)(xRy...). This is unacceptable; and so memory must count even though it is not actually being experienced.

The question we then need to answer is: in what sense *in potentia*? Two answers present themselves: a "capacities" analysis and a "possible worlds" analysis. At this point I am not even clear on what the issues are that might

\[23\] This discussion is the justification for adding caveat 2 to our "assisted memory" criterion R^a. See *infra*, p. 55.
separate these approaches. "Possible worlds" has the advantage of having a significant logical machinery available, but it is also more complex. Notoriously, it has difficulties distinguishing between the nearly possible and the distally possible. But it is not clear that a capacities analysis could do better. It has problems distinguishing between degrees of development (or deterioration) of capacity, as well as distinguishing between capacity as such and "the capacity to utilize one's capacity." On the other hand, the possible worlds approach seems anachronistic; nor do I think the capacities approach necessarily undermines topic neutrality (as it would if we thought of it as merely a physical capacity). For Locke himself refers to "operations of thinking and memory" as properties of thinking things. (§27, 347)

Leaving more detailed examination of these issues for a possible future work, I will just stipulate that when we speak of "memory" or the "R criterion" we are talking about a capacity of thinking things. That is, xRy means x²Ry, "x has the capacity to remember being y."

**Corrected Memory vs. Uncorrected Memory (R² vs. R")**

There is, next, disagreement over whether Locke's memory criterion is of the corrected or uncorrected variety. If memory is uncorrected, then you are responsible for actions of persons you remember being, even if "you" did not do them.²⁴ If corrected, then you are not responsible for such actions; on the other hand, you might be responsible for an action you are deemed to remember but

---

²⁴ The quotations around "you" are because, theoretically, you did do them if you are the same person who did them, and you are the same person as any person you remember being.
do not actually remember. The textual evidence tends to favor the view that Locke intended R\textsuperscript{o}, although there are both counter-arguments and arguments available for R\textsuperscript{c}.

Curley (1982, 311) brings out some of the most perspicuous evidence, i.e., Locke’s motivation for his moral theory. It has two elements:

(1) Personal identity needs to be grounded in something of which we are certain: “So that self is not determined by Identity or Diversity of Substance, which it cannot be sure of, but only by Identity of consciousness.” (Locke, §23, 345.)

(2) Moreover, justice of punishment and reward has as a necessary condition that we remember being the person whose action is being so required: “In this personal identity is founded all the Right and Justice of Reward and Punishment; Happiness and Misery, being that, for which every one is concerned for himself, not mattering what becomes of any Substance, not joined to, or affected with that consciousness.” (Id., §18, 341.)

Motivational thesis (2) is an attack on the Cambridge Platonists (Henry More, Thomas Glanville, and others) who held that the doctrine of metempsychosis explained differences in the condition of human beings; that is, our good or evil acts in past lives were requited by our estate in this one. (Thiel (1998b), 897.) Motivational thesis (1), on the other hand, is apparently an attack on the doctrine of innate ideas. But it may not be at first clear how these can co-exist or provide reason to hold R\textsuperscript{u} over R\textsuperscript{c}; since R\textsuperscript{c} is certain, but may not be something we experience; while R\textsuperscript{u} is something we experience, but is clearly not certain.

The key to unraveling this riddle is to separate our belief in the contents of our memory from the belief we are having a memory. Now, as Leibniz points out, we “can be mistaken” in our memories (in the sense that what we remember does not correspond to reality). But it is just because we can be mistaken that
the evidence of the two motivational theses favors R\textsuperscript{4}. We can think of memory as corrected memory (often called "recollection"). Recollection cannot be mistaken; it necessarily corresponds to reality—but it is not something an individual knows. It is deemed memory, not necessarily something that is part of a knower's experience as consciously held belief. On the other hand, although we can err in the content of our memories, one thing it might seem we cannot err in is the belief we are having a memory. That is an actual experience and uncorrectable. We have "privileged access" to it, and no one can say (no matter what happened in the real world) that we are mistaken in thinking or believing we remember.\textsuperscript{25}

In "Varieties of Privileged Access", William Alston delineates six types of epistemic privilege we have to our own mental states: (i) infallibility, (ii) omniscience, (iii) indubitability, (iv) incorrigibility, (v) truth-sufficiency, (vi) self warrant. (See Appendix A.) According to this taxonomy, our belief we are having a memory (R\textsuperscript{4}) is both infallible and omniscient; it is impossible for me to believe I am having a memory without that belief being both justified and true (infallibility) and it is impossible for it to be true I am having a memory without my also believing I am having one and being justified in that belief (omniscience).\textsuperscript{26}

\textsuperscript{25} Under the mind/brain "contingent identity" thesis of J.J.C. Smart (1981) and others, the belief we are having a memory could be correctable via neurologic examination. The contingent identity thesis is controversial, and certainly did not occur to Locke.

\textsuperscript{26} Alston would probably deny that we have omniscience with respect to our having a memory. I feel virtually certain Locke would assert it, since he writes that it is “impossible for one to perceive, without perceiving that he does perceive.”
But choosing $R^c$ rather than $R^n$ and asserting Locke meant by "memory" "recollection" would deprive our "having a memory" of any privilege higher than self-warrant (it is logically impossible that we hold a belief we are having a memory ($R^c$) without being justified in that belief); and considering how unreliable our memory actually is, arguably deprive our experience of our memories of any privilege at all. This would render motivational thesis (1) null; and justice (according to motivational thesis (2) beyond the reach of human judgment.

There is other textual evidence favoring $R^n$ as well. On the one hand, (Case A) Locke writes,

> But if it be possible for the same Man to have distinct incommunicable, consciousness at different times, it is past doubt the same Man would at different times make different Persons: which, we see is the Sense of Mankind in the solemnest Declaration of their Opinions; Humane Laws not punishing the Mad Man for the Sober Man's Actions, nor the Sober Man for what the Mad Man did, thereby making them two persons . . . . (§20, 342.)

So Locke clearly intends that personal identity is not restricted to the observable and verifiable, viz., one's continuing "man"hood. And on the other hand, (Case B) he writes;

> Matter without consciousness united to any Body makes the same Person. But let him once find himself conscious of any of the Actions of Nestor, he then finds himself the same Person with Nestor. (§14, 340.)

Making the same point from the opposite direction—our memory is not only a necessary but a sufficient condition of our being the same as a prior person.

But it is easy to feel sympathy with the proponents of $R^c$; for them, both cases A and B sound preposterous. They seem to invite sarcastic gibes: Leibniz writes, "I would have retained my rights . . . if I forgot my whole past . . . without
having to be divided into two persons and made to inherit from myself.” (236-37.) Even Locke himself observes,

    But that which we call the same consciousness, not being the
same individual Act, why one intellectual Substance may not have
represented to it, as done by some other Agent, why I say such a
representation may not possibly be without reality of Matter of Fact, as
well as several representation in Dreams are, which yet, whilst
dreaming, we take for true, will be difficult to conclude from the Nature
of things. (§13, 338.)

and concludes

    [t]hat it never is so, will by us, till we have clearer views of the
Nature of thinking Substances, be best resolv’d into the Goodness of
God, who as far as the Happiness or Misery of any of his sensible
Creatures is concerned in it, will not by a fatal Error of theirs transfer
from one to another that consciousness which draws reward or
punishment with it. (§13, 338.)

One way to argue for \( R^c \) is to flat out reject motivational thesis 2 for type A cases.

Certainly it is a necessary condition of justice in punishment that the same
person be punished as committed the crime; but it is not necessary that the guilty
party actually believe s/he is the same person. So for Leibniz: “[O]ne can be
accountable for what one has done, even if one has forgotten it, provided that
there is independent confirmation of the action.” (1997, 241.) Moreover,

    [i]f we are to speak quite generally of punishment, however, there
are grounds for questioning whether it is absolutely necessary that
those who suffer should themselves eventually learn why, and whether
it would not quite often be sufficient that those punishments should
afford, to other and better informed Spirits, matter for glorifying divine
justice. (Id., 246.)

\( R^c \) provides this, because it argues that we are the same person as who
committed a crime if our corrected memory makes us the same, even if we have
actually forgotten.¹²⁷ Thiel is also willing (1998b, 894-5) to bite this bullet. It may be impossible in principle to "distinguish between genuine and pretended lack of consciousness: since a person’s self-ascriptio of actions through consciousness is beyond the courts' knowledge, how can they ever 'distinguish certainly what is real, what counterfeit'?" (Thiel (1998b), 894.) Courts are justified in making judgments based on the R criterion corrected by our existence qua human: "In other words, they can judge only with regard to the identity of the self as human being, not as person." (Thiel (1998b), 894.) He believes Locke expects type A cases to be of concern only to God at the Last Judgment—when "'[t]he secrets of all Hearts shall be laid open', and 'No one shall be made to answer for what he knows nothing of.'" (Thiel (1998b), 895, from §22.)

As to type B cases ("paramnesia"), Mackie is dismissive. "Neither Locke's own theory nor the revision of it suggested above [to provide transitivity] need be embarrassed by cases of this kind." (Mackie (1976), 184.) His reasoning is that

> [g]enuine remembering is related causally to the experiencing or learning of what is remembered, and through there is some causal link between the actions of various men who really fought at Waterloo and the king's fantasies, and again between the real murderer's actions and the fantasies of those who confess, these causal links are of the wrong kind to constitute memory, since they go through oral or written reports, evidence left by the murderer, found by the police, and described in newspapers, and so on. (Mackie (1976), 184.)

Moreover, he points out that if Locke was really committed to R¹, then his plea that the goodness of God will not permit punishment to be meted out unfairly to sensible creatures who err in their memories would be unnecessary.

---

²⁷ Of course there may still be real cases of mistaken identity.
If, as Locke is maintaining, consciousness itself and nothing else constitutes the identity of the person, then in the imagined cases there would be no error, no injustice, that the goodness of God can be invoked to prevent. The actions of which someone thus became directly conscious would be as much his as anyone’s past actions are in any normal case. As Locke has insisted, it would not, on his theory, matter at all if these actions had been done by a different man or by a different spiritual substance. (Mackie (1976), 184-5.)

These counter arguments do not really constitute grounds for rejecting $R^u$ and accepting $R^c$. For one thing, most of these arguments do not have much force. Let us consider them in order;

Leibniz. Leibniz’s argument is not really the full frontal assault on motivational thesis 2 it appears to be. Leibniz believes that memories cannot be completely lost:

[It is unreasonable to suppose that memory should be lost beyond any possibility of recovery, since insensible perceptions, whose usefulness I have shown in so many other important connections, serve a purpose here too—preserving the seeds of memory. (1997, 239.) (Reprinted for your convenience.)

His comment on §26 (quoted above) is conjectural, and he does not mention any such “grounds”; he goes on to say that “[s]till, it is more likely, at least in general, that the sufferers will learn why they suffer.” (Iid., 246.)

Theil. Theil fails to distinguish between procedural and absolute justice. His argument seems in essence to be that if it is impossible in principle for human courts to use $R^u$ to determine culpability, then it cannot be a necessary condition of justice that they do so. But human justice is always held to two divergent standards—those of procedural and absolute justice. The question of what can or cannot in principle be known does not touch questions of absolute justice at all, but merely should influence questions of what constitutes
procedural justice. Although Thiel quotes Locke in a way that suggests that Locke concedes this point, I believe this is based on a misreading of the text, which has been taken out of context. When Locke writes “[b]ut in the great Day, wherein the Secrets of all Hearts shall be laid open, it may be reasonable to think, no one shall be made to answer for what he knows nothing of; but shall receive his Doom, his Conscience accusing or excusing him” (§22, 344) he does not, I think, mean to say that God will attend to what people actually did or did not do in passing judgment on them. Coming after the line “[s]ince ‘tis evident the personal identity would equally be determined by the consciousness, whether that consciousness were annexed to some individual Immaterial substance or no” (id., §23, 344), I believe the comment is better interpreted as implying that the Great Judge will determine what it was that was person actually remembered at the time of trial, and provide correction for any imperfections in human court procedure.

Moreover, there is additional textual evidence that Locke does not concede Theil’s point. Locke writes,

But if it be possible for the same Man to have distinct incommunicable consciousness at different times, it is past doubt the same Man would at different times make different Persons; which, we see is the Sense of Mankind in the solmnest Declaration of their Opinions, Humane Laws not punishing the Mad Man for the Sober Man’s Actions, nor the Sober Man for what the Mad Man did, thereby making them two Persons . . . .” (§20, 342.)

This implies that Locke believes it to be possible that a court could determine that “the same man [could] have [two] distinct incommunicable consciousness[es] at different times.” Given modern methods of interrogation
this does not seem to be out of the question, and Theil needs to give us more to
go on before we rule Locke’s position “out of court.”

_Mackie._ Despite Mackie’s confident dismissal, problems of paramnesia
are not to be swept aside so easily.  

28 Mackie apparently thinks that it is an easy
matter to distinguish causal links of “the wrong kind” from those of the “right
kind.” A little examination of the wider literature on causality reveals this
confidence to be misplaced. So much is it not the case, in fact, that I make the
ability to, model paramnesia an argument _for_ the causal theory of identity I present
_infra.,_ p. 98. If there is a causal theory which supports Mackie’s claim, I would
like to know what it is. Perhaps a suitable corrective condition could be found,
but the only obvious ones are (i) “only memories of actions done by the same
person are correctly causally linked,” and (ii) “only memories of actions done by
the same wo/man are correctly causally linked.”  (i) is objectionable as circular,
(ii) is objectionable as jettisoning Locke’s commitment to a distinction between
the concepts of “man” and “person.” If there is another more viable corrective
principle, we are still awaiting its enunciation.

As to Mackie’s concern over Locke’s remarks re “the Goodness of God” in
§13, I agree that the quote looks damning, undermining Locke’s entire enterprise
here. But there is another interpretation more friendly to Locke, and which brings
out an important point. Mackie’s complaint fails to distinguish between two
different ways in which memory can be corrected: (i) as to its content, and (ii) as

---

28 Parfit is especially thorough on the issue. _See_ (1984), 219, _et seq._
to whether it is a memory at all. I should say there is no reasonable basis for rejecting such a distinction, because the experience of having mistaken memories is a common feature of minds; the only way to collapse it is to assert that memory is *eo ipso* not memory if it is mistaken.

Locke’s comment, coming on the heels of a complaint:

> But that which we call the *same consciousness*, not being the same individual Act, why one intellectual Substance may not have represented to it, as done by some other Agent, why I say such a representation may not possibly be without reality of Matter of Fact, as well as several representation in Dreams are which yet, whilst dreaming, we take for true, will be difficult to conclude from the Nature of things. (§13, 338.)

suggests that it is best taken as a question of how to distinguish memories from dreams and imaginings.

This is a concern which needs to be taken seriously, because it suggests that our original conclusion as to the justification for taking R— that it has an epistemic privilege— was premature. If we can mistake dreams for memories, our belief we are “having a memory” is not infallible and is probably not omniscient; indeed, it is unclear that it has any epistemic privilege at all. Some correction of R seems justified which yet falls short of a full connection of content and does not fall prey to the concerns which rule out corrections (i) and (ii) (*supra*, p. 53). My intuition is that the general problem of distinguishing imagining from memory will eventually succumb to a functionalist analysis.

---

29 Leibniz, I think, can be taxed with the same error when he writes: “This belief that we have done something can deceive us if the action was long ago. People have mistaken their dreams for reality, and have come to believe their own stories by constantly repeating them. Such a false belief can cause perplexity, but it cannot make one liable to punishment if there are no other beliefs which confirm it.” (241.)
However, I do not propose undertaking such a project here. Rather, I will make a more pragmatic suggestion.

My suggestion is that instead of $R^a$, we deem Lockean remembering to be $R^n$; the memory we have when we consider our memory carefully and with the assistance of all available internal and external evidence. The memories we have of being a person after such a procedure are determinative, and not subject to further correction. This leaves it open to me to remember being Nestor, and thus to be Nestor, no matter how unlikely this might seem to others based on external evidence. And if my memory of being a past “self” is still not refreshed after such a procedure, then I am not that person and cannot justly be punished, even if “there is independent confirmation of the action.” (Leibniz, 241.) In addition, two caveats need to be mentioned at this point: (1) we cannot be the same person as some person who did not exist, and (2) we cannot be the same person as someone if, through transitivity, that would make it indeterminate which of two contemporaneous persons has a particular soul.

(1) is actually a condition built into our criterion already:

**Personal Identity—Diachronic**

IIIa. $(\forall x)(\forall y)((t(x) \neq t(y) \& x = P = y) \Rightarrow (x R y \land y R x))$

for the universal quantifier by convention normally ranges over only existents. (We might interpret this condition more or less liberally, but that is not an issue we need to settle here.) (2) is required by the topic-neutrality of Locke’s theory. Souls were conventionally thought indivisible (cf. Descartes, who thought indivisibility constituted the “real distinction” between thinking and extended things (Meditation Six)). Since Locke does not want to give up on the possibility
that consciousness is conveyed by soul, he would not entertain examples which preclude that possibility.

The switch to R⁺ (with the two caveats, R⁺⁺) has the unfortunate effect of weakening the epistemic advantage our belief we are having a memory has over other beliefs. It is no longer infallible (since we might believe we have all available evidence and the caveats are met when they are not) nor omniscient (for we might not believe they are met when they are). Nor is it incorrigible (since others might be in a position to tell us our belief we are having R⁺⁺ is wrong). Nevertheless, I think such beliefs are still advantaged, insofar as they are both indubitable and truth-sufficient.

Many of these factors can be brought out in the following example based on a true story. A woman (let us call her Xena) has been missing for many years. Her family hires a private investigator to search for her. He has the idea of looking up all her former classmates. Among these classmates is Yena. He finds Yena living a thousand miles away, and to his surprise discovers it is really Xena who has been living for years under the identity of Yena. Xena had lost her memory in an accident; her amnesia was total, except for the name and many of the incidents of the life of Yena. From these fragments she had concocted a coherent past life, and was living comfortably without any thought of the life of Xena.

Were Xena and Yena the same person? No, because once Xena was confronted with the evidence of her past life by the detective, she realized what had happened and remembered her accident and much of her life as Xena. It is important to note that at no time from t₃ (see Diagram E) on was she the same
person as Yena; because it was always possible for her to remember being Xena and not Yena. What if no amount of evidence or remonstration could convince her that she was Xena and not Yena? Then, according to $R^a$, she would be the same person as Yena. However, according to $R^{a\ast}$, she would not be; because her being the same person as Yena would mean two contemporaneous persons (Xena and the “real” Yena) would be the same person as Yena (say at time $t_2$); and since there is only one soul of Yena (if there are souls), it would be indeterminate which of the two Yena’s at time $t_5$ possesses it.

However, if Yena had died at $t_2$, then Xena’s $R^{a\ast}$ at $t_5$ that she is Yena would make her the same person as Yena. Note that this conclusion (that Xena has an $R^{a\ast}$ of being Yena) would be both indubitable and truth sufficient: for (i) everyone who might be in a position to doubt Xena’s belief she is having an $R^{a\ast}$ of being Yena has (or should have) provided evidence to her that she is not $=P=$ as Yena (this per hypothesi is the meaning of “all available evidence” in the $R^a$ condition): and (ii) given the truth that Xena has $R^{a\ast}$ it is impossible for Xena not to believe she has such an $R^{a\ast}$, since $R^{a\ast}$ has built into it (again, per hypothesi) a justificatory procedure (investigation and due consideration).

I think the $R^{a\ast}$ criterion overcomes the shortcomings of $R^a$; while it provides some correction of our memories (both their content and the belief we are having a memory) it still allows for the possibility we “could be Nestor” if we remember his actions in the appropriate way. Some people may find this counterintuitive, but I feel comfortable that Locke was willing to “bite this bullet” and I am too.
R\textsuperscript{1} vs. R\textsuperscript{p}

In “Locke on the Identity of Persons,” B. Brody argues that:

[I]t seems clear that it is [R\textsuperscript{p}] and not [R\textsuperscript{1}], that Locke was advocating. To begin with, [R\textsuperscript{1}] is totally implausible since it guarantees that no person is identical with anything that existed just a few hours ago. (Brody (1972), 332.)

On the other hand, (Mackie (1976), 183) seems to be arguing (if I read him correctly) that only R\textsuperscript{1} is a strong enough criterion to get us a theory of personal identity and not merely a theory of “action appropriation.” For identity of the person, we need R\textsuperscript{1}. If both of them are right, then Locke’s identity criterion for persons is useless. I believe that Mackie is wrong (see infra, p. 80). R\textsuperscript{p} is enough to get us personal identity on the diachronic concept of person I advocate supra., p. 34. But that does not by itself tell us that Brody is right. His claim so far—and Mackie’s as well—is about what Locke should have said, not really what he meant (and much less what he wrote). Sensitive to this, Brody continues by quoting the case of the “mad” and “sober” man:

But if it be possible for the same Man to have distinct incommunicable consciousness at different times, it is past doubt the same Man would at different times make different Persons; which, we see is the Sense of Mankind in the solmnest Declaration of their Opinions, Humane Laws not punishing the Mad Man for the Sober Man’s Actions, nor the Sober Man for what the Mad Man did, thereby making them two Persons . . . . (§20, 342.) (Reprinted for your convenience.)\textsuperscript{30}

and goes on to conclude,

[t]he sober man and the madman are not the same person because they have incommunicable consciousnesses, i.e., each is not conscious of anything the other is conscious of, and not merely

\textsuperscript{30} Locke the uses “humane” for “human” throughout the Essay.
because one does not remember some one the thing of which the other is conscious.

But this is not enough to clinch the argument. It merely states that total lack of consciousness of a past self is a sufficient condition for being a different person. But this is consistent with the possibility that some lesser lack of consciousness is also sufficient to make one a different person. (i.e., only total recall makes you the same.) To confirm that total forgetfulness is also a necessary condition for being a new person, we would like to see a case in which partial lack of consciousness is not sufficient to make one a new person; from which we could infer that partial consciousness is sufficient for being the same person. For example, when Locke considers the anti-parallel case of the drunk and sober men, he might have said “justice punishes the sober man for what he did when drunk because he must have at least some recall of the incident.” He did not say that; rather, in §22 he wrote:

But is not a Man Drunk and Sober the same Person, why else is he punish’d for the Fact he commits when Drunk, though he be never afterwards conscious of it? Just as much the same Person, as a Man that walks, and does other things in his sleep, is the same Person, and is answerable for any mischief he shall do in it. Humane Laws punish’d both with a Justice suitable to their way of Knowledge: Because in these cases, they cannot distinguish certainly what is real, what counterfeit; and so the ignorance in Drunkeness Sleep is not admitted as a plea. (343-4.)

Moreover, Locke seems at times explicitly to endorse R⁴. For example, in §10 he writes:

[F]or as far as any intelligent Being can repeat the Idea of any past Action with the same consciousness it had of it at first, and with the same consciousness it has of any present Action; so far it is the same personal self. For it is by the consciousness it has of its present Thoughts and Actions, that it is self to it self now, and so will be the
same self as far as he same consciousness can extend to Actions past or to come . . . . (336.)

In §26 he writes,

This personality extends itself beyond present existence to what is past, only by consciousness, whereby it becomes concerned and accountable, owns and imputes to itself past Actions, just upon the same ground, and for the same reason that it does the present. (346.)

And in §16 he writes,

[Had I the same consciousness, that saw the Ark and Noah’s Flood, as that I saw an overflowing of the Thames last Winter, or that I write now, I could no more doubt that, that I write this now, that saw the Thames overflow’d last Winter, and that viewed the Flood at the general Deluge, was the same self, place that self in what Substance you please, than that I that write this am the same my self now whilst I write (whether I consist of all the same Substance, material or immaterial, or no) that I was Yesterday. (341.)

In the latter quote, Locke leaves us a little wiggle room: for there he writes, “[h]ad I the same consciousness, I could no more doubt . . . .” This could mean he considers R1 a sufficient but not necessary condition for personal identity. Likewise §26 is equivocal—“same ground” may just mean memory (Rn) which makes our past actions part of our current consciousness. But §10 is hard to finesse; for he writes—“As far . . . so far . . . .” suggesting that less far than the same consciousness would not be sufficient for continued personhood.

On the other hand, he often writes as if less far is sufficient—for example,

That with which the consciousness of this present thinking thing can join itself, makes the same Person, and is one self with it, and with nothing else; and so attributes to it self to itself, and owns all the Actions of that thing, as its own, as far as that consciousness reaches, and no farther; as everyone who reflects will perceive. (§17, 341.)

But let him once find himself conscious of any of the Actions of Nestor, he then finds himself the same Person with Nestor. (§14, 340.) and others.
But such evidence is also equivocal. As to the second quotation (and other references to “action”, e.g., in §16) he may mean by “conscious of any of the actions” full consciousness of such action, as one had at first—and though the first suggests that memory in the normal sense will suffice (as that is how consciousness actually joins itself to the past), Locke does not explicitly say that.

And yet we have great sympathy with Brody’s concern that R1 commits Locke to an empty theory. Locke, for his part, clearly does not intend his criterion to be empty; for he writes,

Nothing but consciousness can unite remote Existences into the same Person, the Identity of Substance will not do it. For whatever Substance there is, however there is, however framed, without, without consciousness, there is no Person: And a Carcase may be a Person, as well as any sort of Substance be so without consciousness. (§23, 344.)

and even more explicitly, in §16:

But though the same immaterial Substance, or Soul does not alone, where-ever it be, and in whatsoever State, make the same Man; yet ‘tis plain consciousness, as far as ever it can be extended, should it be to Ages past, unites Existences, and Actions, very remote in time, into the same Person, as well as it does the Existence and Actions of the immediately preceding moment. (340.)

Here he seems to assert that short term and long term memory work the same. Another, more reasonable, interpretation, is that “as well as it does” in §16 just means that memory serves to make the person “me” from long ago as much the same person as the current “me” as the person “me” from a moment ago does. But in either case, he clearly intends that his criterion does work and make diachronically distinct person stages a single person. And so also in §25: “And thus, by this consciousness, he finds himself to be the same self which did such or such an Action some years since, by which he comes to be happy or
miserable now.” (344.)

This leaves us in a bit of a quandary. The textual evidence clearly favors taking Locke to be espousing \( R^1 \). Yet this is contrary to common sense and even Locke’s own desire for a theory which will support claims about the justice of punishment. There are, I think, two positive solutions, both of which go somewhat beyond the text but are in the spirit of Locke’s motivation for the theory.

(1) Taking a cue from §§16 and 25, perhaps what we need is \( R^1 \) of \textit{action}, not of the total contents of consciousness at a given time. This would allow us to “unite remote existences and actions” into a single person and justify punishment or reward, which requires consciousness of the action being requited. (We might also view this as a “partial-personal identity” criterion, taking our cue from “as far as . . . so far.”) However, this has the drawback that it lends some credence to the Mackie concern (\textit{infra}, p. 80) that Locke’s theory is not a theory of personal identity but rather of “action appropriation” (Mackie (1976), 183).

(2) Alternatively, we can take our cue from §16 and Locke’s claim that he “could no more doubt” that he was the same person who saw the “general deluge” than that he was the same person that saw the Thames overflow if he had the same consciousness of each. Putting emphasis on the word “doubt,” we could say that what he should have said is that whatever degree of remembering is enough to provide us with \textit{certainty} that we had the earlier experience is that degree which makes us the same person as originally experienced it. In combination with the \( R^{a*} \) criterion (\textit{supra}), \( R^{d} \) opens a little space between \( R^1 \) and what is sufficient to make us the same person as some other person in the past.
whose acts and consciousness we remember.

One word in one line of a single section may seem like a thin reed on which to base an interpretation of an entire chapter. However, the $R^d$ criterion is supported by some more general considerations.

As we have seen (supra, p. 46) Locke’s theory of personal identity is to some extent a response to a skeptical concern. In some respects, Lockean skepticism is more radical than Cartesian, for Locke not only wishes to refrain from making knowledge claims, but also sometimes claims not to know. But here, it seems typically Cartesian: if I can doubt I am the same person, then I am not the same person. Note that the $R^a$ and $R^d$ criteria are not redundant; we can be unsure of even our best memory, and we can be sure of memory which has little support, and indeed, is false. I do not think this solution is ideal; Locke’s claims on this topic are too problematic to make an ideal interpretation possible. I do think it is somewhat better than taking $R^p$ against the textual evidence just on the need to make Locke’s theory non-empty.

D. Non-Substance View

How deep does Locke’s commitment to the non-substantiality of personhood go? If we judge by the comments quoted to justify the non-substantial view, it goes very deeply indeed; for he says whether we are the same substance or not “concerns not personal identity at all.” (§10, 336.) But this makes Locke’s criterion of personal identity wildly counter-intuitive. How could a consciousness “reach so far” if there is no connection other than just the thought? Isn’t this just foolishness? As Leibniz said,
Organization or configuration alone, without an enduring principle of life which I call 'monad', would not suffice to make something remain numerically the same, i.e. the same individual . . . [A]s for substances which possess in themselves a genuine, real, substantial unity, and which are capable of actions which can properly be called 'vital'; and as for substantial beings, quae uno spiritu continentur as one of the ancient jurists says, meaning that a certain indivisible spirit animates them: one can rightly say that they remain perfectly 'the same individual' in virtue of his soul or spirit which makes the I in substances which think. (231-2.)

The soul is the source of real identity of persons: “And that is how I distinguish the incessancy of a beast’s soul from the immortality of the soul of a man: both of them preserve real, physical identity . . . .” (236.) Perhaps, we might conjecture, Locke meant to say only that whether we are the same immaterial substance is what does not concern personal identity; i.e., he was a materialist. One might infer that, perhaps, from such comments as this (at §25, 346):

Thus any part of our Bodies vitally united to that, which is conscious in us, makes a part of our selves: But upon separation from the vital union, by which that consciousness is communicated, that, which a moment since was part of our selves, is now no more so, than a part of another Man’s self is a part of me; and ‘tis not impossible, but in a little time may become a real part of another Person. And so we have the same numerical Substance become a part of two different Persons; and the same Person preserved under the change of various Substances,

where he seems to suggest material substance can be “a real part of a person.”

In the end, however, I wish to reject both of these readings. What I think is going on is that Locke accepts that consciousness and soul do always go together; however, what makes for the identity of persons is not soul, but consciousness. In other words, we might have thought (as the ontological view of personhood holds) that the identity of persons and souls mutually entailed each other:

A. x and y are the same person if they have the same soul.
But Locke makes the important distinction between the soul and the consciousness. The soul is there, but it is not the soul that does the philosophical work; it is consciousness that does it. Again, this is an attack on the Cambridge Platonists (see *supra*, p. 46), according to whom the soul gets expunged of all memories of its prior life before being reincarnated as a new being. Locke is saying that soul without consciousness cannot make a "same person." He does not, on this reading, mean to say that same consciousness without any soul at all can make a same person; he expects people to fill in the blank with the common understanding that consciousness is conveyed by a soul.

So now we have:

B. x and y are the same person if they have the same consciousness (*i.e.*, one remembers being the other); and, moreover,

C. If either x and y are the same person or x and y have the same consciousness, then they have the same soul.

replacing the mutual entailment of soul and person of the ontological view with a one-directional entailment.

Thus we can account for such comments as this: "But let him once find himself conscious of any of the Actions of *Nestor*, he then finds himself the same Person with *Nestor." (§14, 340.) Coming on the end of a long disquisition on why having Socrates's soul without his memories would not make you the same person as Socrates, Locke does not mean to suggest that the memories *without* the soul make you the same person as Socrates. Rather, he means to say that you need the memories along with the soul to make the same person "as *Nestor*."
The view also accounts for his choice of the following words for the “Prince/Cobbler” case: “[S]hould the soul of a Prince, carrying with it the consciousness of the prince’s past life . . .”, which makes no sense if we take the non-substance view to mean consciousness has no relation to substance at all.

While I think Locke meant to imply that consciousness had a one-way entailment relation to substance qua soul, I also think that Locke did not wish to commit himself on the question of whether the soul was material or immaterial. Although he writes, “I agree the more probable opinion is, that this consciousness is annexed to, and the affection of one individual immaterial substance” (§25, 345) this is as much to point out that this dictum is not a matter of certain knowledge. And in fact, his theory in modern parlance is “topic neutral”—it does not take a position on the materiality or immateriality of mental properties. We take “the soul of man, for an immaterial Substance” but are really “in the dark concerning these Matters.” (§27, 347.) This very neutrality was a matter of irritation to some of his contemporaries, who would much have preferred him to take the position that the soul is immaterial:

Since Locke’s theory does not commit him to either the materialist or the immaterialist account of the soul, it could in principle, be accommodated to both accounts. Yet as a matter of fact, although at least some materialists referred favorably to Locke’s theory, defenders of the immateriality of the soul tended to reject it, partly because it leaves open the very question of immateriality. The most prominent theological critic was Edward Stillingfleet, Bishop of Worcester, with whom Locke entered into a long controversy over the Essay as a whole. (Thiel (1998b), 897.)

But Locke really did not think this was necessary—for he wrote,

‘[t]is not of such mighty necessary to determine one way or t’other, as some over zealous for, or against the Immateriality of the Soul, have been forward to make the World believe (IV.II.6, 542)
on the ground that our Maker can resurrect us in whatever form He pleased and make us fit to receive His judgment.

E. The Concept of “Man”

Most commentators are agreed that Locke’s distinction between the concepts of “man” and “person” is presented rather imprecisely. “Locke seems not in Chapter 27 to have had one steady view of what a man is.” (Chappell (1989), 76.) There are really two concerns: (1) what is the definition of “man”? and (2) how does that definition interact with that of person? Since the second concern may place some constraints on answering the first, it will be well to begin with it. There are, I think at least four possible ways in which the concept of man might be related to that of person.

(a) The man is the outer visage of the person (whose identity is otherwise known to no one except the individual involved and God).

This interpretation perhaps helps us make sense of certain puzzle cases (as, e.g., the alcoholic vs. the sleepwalker (in §22, 343-4) and the Prince and the Cobbler (in §15, 340) which make reference to the difficulty outsiders have in making judgments as to personal identity based on consciousness. It is an interpretation favored by both Thiel (1998b, 894-95) and Atherton (1983, 274-75).

(b) A person might be both a rational and a self conscious being, whereas a man is merely a rational being.

Locke refers to both “man” and “person” as “rational beings”. However, although he insists that a person is self-conscious (in a way articulated more fully
infra, p. 93) he nowhere makes the same claim for man. Against this thesis we could cite the evidence of III.xi.16, where Locke writes,

For were there a Monkey, or any other Creature to be found, that had the use of Reason, to such a degree, as to be able to understand general Signs and to deduce Consequences, about general Ideas, he would no doubt be subject to Law, and in that Sense, be a Man, how much soever he differ’d in Shape from others of that Name. (517.)

suggesting that the person (here "man") is someone who is (merely) rational, not self-conscious. But this text does not reflect Locke’s most considered thought on the matter (as demonstrated by the fact it does not distinguish “man” from “person” as the later chapter on identity does); so perhaps it is permissible to discount it.

(c) Person is the action-owning, responsible aspect of the human individual.

The wo/man includes rationality, but not accountability. This makes Locke’s claim that “person is a forensic term” contrastive, the idea being that “person is a forensic term and wo/man is not a forensic term.”

(d) The wo/man is by nature limited to a particular physical instantiation and form of body, but the person is not.

Although persons usually (for all we know, always) are associated with particular wo/men, in theory there is no obstacle to one person inhabiting different wo/men or one wo/man being associated with more than one person. In addition, it seems likely that Locke would accept the possibility of persons of other forms than human, but only humans can be men:

‘[T]is not the Idea of a thinking rational Being alone, that makes the Idea of a Man in most Peoples Sense; but of a Body so and so shaped joined to it; and if that be the Idea of a Man, the same successive Body
not shifted all at once must as well as the same immaterial Spirit go to the making of the same Man. (§8, 335.)

Summary: we must note at first that conditions (a) and (c) are contradictions; for it cannot both be that a man (a) is the responsible agent for practical purposes, and (c) that the person is the responsible agent for all purposes. I am inclined to think that the reasons weigh more on the side of keeping (c); for Locke in at least one case (the mad/sober man case (§20)) foresees the possibility of human judgment based on change of person. Moreover, the cases in point do not depend exclusively on (a) for their meaningful interpretation; but can be made sense of in other ways. As to the rest, they appear to be non-exclusive; indeed, each may capture some aspect of Locke’s thought.

We must now consider the possible interpretations of the concept of “man” itself, as well as whether the above considerations place any constraints on our choice. Locke lays out these possible alternatives fairly clearly in §21, where he writes:

First, It must be either the same individual, immaterial, thinking Substance: In short, the same numerical Soul, and nothing else. Secondly, Or the same Animal, without any regard to an immaterial Soul. Thirdly, Or the same immaterial Spirit united to the same Animal. (343.)

Of these, the first seems unlikely. Nowhere in the text does he advance it as his own view. Moreover, he explicitly rejects it at §6, where he writes that this “way of speaking must be from a very strange use of the word Man, applied to an

31 For interpretation of the prince/cobbler case, see infra p. 111. The interpretation of the alcoholic/sober man case I will postpone to a possible future stage of the work.
Idea, out of which Body and Shape is excluded . . . ” (332.)

That leaves the second and third possible interpretations. As Chappell (1983, 82 fn. 9) notes, Locke seems to endorse both of these at different points in the text, and at other times mentions both without indicating a preference.32 On the interpretation of the distinction between man and person given above, we might give the edge to the combination view; for a body (presumptively) has no rationality and cannot meet requirements (a) and (b). However, if we incline to the animal only view, we might wish to amend (c) by dropping the claim that man is rational; and this would further provide a justification for the claim that only person is a forensic term, since it may not be clear why the rationality of the man (even without self-consciousness) should not entail responsibility.

Below, I will argue that the causal interpretation resolves this conundrum by taking this “form of man” to be the rational soul. But meanwhile we can respond to the attack on condition (c) as follows.

Although Locke’s view is topic neutral, he frequently denies that we can know that immaterial substance is self-identical over time. “So that the self is not determined by Identity or Diversity of Substance, which it cannot be sure of, but only by Identity of consciousness.” (§23, 345.)

But one of the conditions of being the same man is having “the same immaterial spirit.” Since a condition of the justice of punishment is that we know that the entity being punished is the same as the doer of the deed, it follows that

---

32 Animal only: §§6, 8 and III.xi.20 (332, 333 and 518f.); Spirit and Animal, §§8 and III.vi.26 (335 and 453); no preference, §§21 and 29 (343 and 348).
the “man” is not a fit object for punishment—because unlike the “person,” we cannot know for certain whether s/he is the same as some earlier entity of the same kind.

IV. The Circularity Objection

The circularity objection arises in the following way. Consider our definition of the term “person”: “x is a person iff x is conscious and there is some y who is the same person as x.” The definiens for the definiendum “person” contains as one of its parts the term “same person as,” and the definiens for that term (ipso facto) contains the term “person.” This can be represented graphically as follows:

```
    0   Person
  1.   .   .   . Same person as
  2.   .   .   . Person.
```

Line 0 is the original definiendum; each additional line is one of the terms of the definiens of one of the terms of the line above. (Other terms of each line are suppressed for clarity.) Now, as I have argued elsewhere (Fried (2000)), excerpted in non-formal part as Exhibit B to his thesis), not all circular definitions are vicious. But we can probably agree that for a non-empirical definition, the reappearance of the definiendum within two steps is both necessary and sufficient for a definition to be termed “viciously” circular; it is not sufficiently

33 In my paper I call them “nodes.”
informative.\textsuperscript{34}

The same thing happens if we start within "same person as":

\begin{center}
\begin{tabular}{ll}
0 & Same Person as \\
1 & . . . . . . Person \\
2 & . . . . . . Same Person as.
\end{tabular}
\end{center}

Again, we have the reappearance of the original definiendum within two steps: a vicious circle. So Locke has not provided us with formally adequate specifications of the concept of "person" or "personal identity." Lacking adequate definition of either of these expressions, Locke's project in Chapter 27 must be considered fatally flawed.

Now of course, it may just be that we have misinterpreted Locke. It is worthy of note that circularity objections of various kinds have been raised many times before, \textit{e.g.}, by John Sergeant. As Thiel puts it:

Moreover Sergeant misrepresents what Locke has to say about consciousness constituting personal identity when he claims that for Locke, consciousness of \textit{Identity} constitutes \textit{Identity}. Had Locke said this, his theory would have been quite obviously circular. But Locke does not hold that consciousness of sameness constitutes sameness: rather, he holds that consciousness of thoughts and actions constitutes the person and its \textit{Identity} over time. (1998b, 898.)

\textsuperscript{34} Note that an additional circle may be read out of the definitions, as follows:

\begin{center}
\begin{tabular}{ll}
0 & Person \\
1 & . . . . . . Same person as \\
2 & . . . . . . Remember being \\
3 & . . . . . . Person
\end{tabular}
\end{center}

"Same person as" has part of its definition the term "remember being," which has part of its definition the term "person." According to our criterion, however, this circle is not vicious—it takes three steps for the original definiendum to reappear.
Thiel’s view apparently places him close to Mackie’s position, which, however, we will refute *infra*, p.80. A similar circularity charge is heard from Butler, who wrote that “Consciousness of personal identity presupposes and therefore cannot constitute personal identity, anymore than knowledge in any other case can constitute truth, which it presupposes . . . etc.” (Butler (1736), 100.) Thiel goes on to say, “[e]ven though the charge of circularity against Locke fails, it has been repeated many times since Sergeant up to the present day . . . .” (1998b, 899.)

As charges of circularity have been made and countered so often, it is appropriate for us to approach the matter with some care.\(^{35}\) Moreover, we should be committed to giving as charitable an interpretation of Locke as we possibly can; and Locke’s defenders clearly can find textual support for their position. We must be sure we have examined all the possible alternatives before recurring to the pessimistic conclusion. We will take two basic approaches to establishing the charge of circularity. In the next section, we will examine the definition more closely. Even though our version seems to be the plain import of his definition of person, perhaps we have overlooked some third element in terms of which “person” is defined which breaks the circle. So we will give a fuller definition which refrains from suppressing any term that might possibly be helpful in getting Locke out of the circle, and then show for each that it fails to do so.

In the following section, we will consider three more radical interpretive

\(^{35}\) I say “charges”, for such objections have been raised on many distinct grounds. (Thiel (1998b), 898.) I will not consider others in any detail here.
efforts: the Mackie position (a person is the contents of consciousness, and personal identity sameness of such contents); the Atherton position (a person is a center of plain consciousness and personal identity sameness of such consciousness); and finally, the self-consciousness position (a person is a center of self-consciousness and personal identity sameness of self-consciousness). I will demonstrate that none of these positions is a reasonable interpretation of the Lockean text.

A. The Fuller Definition

Let us turn, then, to the offending passage and see what we can do. It reads as follows:

This being premised to find wherein personal identity consists, we must consider what Person stands for; which, I think, is 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: It being impossible for anyone to perceive, without perceiving, that he does perceive. (§9, 335.)

Perhaps our reduction of this definition to “person is someone who is conscious and the same person as some other person” was a bit too hasty. We should consider a more fully fleshed version, which might contain significant additional information which breaks the circle, as follows: “A person is that thing (and only that thing) which is conscious and can consider itself to be the same thinking thing over time.” The points which need clarification seem to be as follows: (a) does “consider “ do any work? (b) what is a “self”? (c) what is the role of “consciousness”? and (d) what is a “thinking thing”? We shall consider them in turn.
(a) "Consider". This might suggest a divergence between thinking and being. The mere fact that you consider yourself the same person might not mean you are the same person. Since our definition of personal identity is of being the same person as some other, not usually merely thinking yourself such, on this interpretation the definition would be non-circular.

Unfortunately, as our discussion of the remembering being criterion (see R\textsuperscript{c} vs R\textsuperscript{u} supra, p. 45) makes clear, for Locke there is no space between the concepts of thinking and being the same person. It is enough to be the "same person" that you think yourself the "same person as." So this defense fails.

(b) Self. The most obvious reading of self is just the reflexive reference to person. It is hard to read out any alternative; and indeed, Locke himself writes:

Any substance vitally united to the present thinking Being, is a part of that very same self which now is: Any thing united to it by a consciousness of former Actions makes also a part of the same self, which is the same both then and now. Person, as I take it, 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. (§25, 346.)

These comments make it clear that "himself" means "his person." And so references to "self" and to "itself" in §9 cannot be to something other than "person" and thus do not take us out of the circle.

(c) Consciousness. Consciousness is that through which a person conceives of him/herself as being the same person. "Which it does only through that consciousness, which is inseparable from thinking . . . ." (Locke §9, 335.) Consciousness is thus requisite to being a person and thus to being the same person as some other, but having consciousness is not what constitutes one a person. Now, this much must still be said about the use of the term
“consciousness.” It cannot be removed from the definition; indeed it is a part of our stated definition of person. So it has the effect of making the definition more complex than it would be if it were just a “person is someone who is the same person as another person.” But that by itself is not enough to remove the circularity of the definition from viciousness. What is needed to do that is another “node”—something that displaces the term “same person as.” (See Appendix B and our fuller discussion of comparative circularity at “Circularity Redux,” p. 109.) It does not seem that consciousness (either plain consciousness or self consciousness) can fulfill that task. (See “The Atherton Position” and “Self-Consciousness,” infra pp. 83 and 93.)

Now it is true that Locke sometimes writes as if he means something else by consciousness, and perhaps we can infer a different definition of person from remarks such as this:

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 as it was then; and ‘tis by the same self with this present one that now reflects on it, that that Action was done. (§9, 335.)

As Atherton puts it, “Locke says it is thanks to consciousness that a person is a person, and there is no obvious reason to understand him here as saying that memory creates a present self.” (Atherton (1983), 277.) But I think that all that we can properly infer from such comments is that consciousness is the glue (or bridge) with which distinct persons are joined as one. “[B]ackwards to any past action or thought” means that there must be some diachronically distinct person who acts or thinks to which such consciousness is extended. So this text does not gainsay my interpretation that, for Locke, there must be two diachronically
distinct persons who are the same for there to be one diachronically continuous person. Consciousness “extends”—but it must extend and extend between two “persons” for a person to exist at all. “It is the same self for a person now as it was then.” (For “self”, see supra p.75.) His remarks at §17 underscore the point that “consciousness reaching” is a joining action and not a constituting action.

(d) Thinking thing. What does Locke mean by the term “thinking thing” in the phrase “consider itself as itself, the same thinking thing in different times and places?” His remarks at §10 clarify the matter somewhat: “I say, in all these cases, our consciousness being interrupted, and we losing the sight of our past selves, doubts are raised whether we are the same thinking thing; i.e. the same substance or no.” (336.)

So Locke means by same thinking thing same substance. This usage should not come as a surprise to us. This is the familiar Cartesian usage under which an “I” is a res cogitans. Locke obviously had read Descartes, and there is no evidence he means anything else here. Here, at last, is a reason to think the circle may be broken. For if the definition of a person does not require the use of the terminology “same person,” but rather instead the usage of “same substance,” there is no immediate (and thus vicious) circularity in it. Since, as we have seen, for Locke person is not a substantial entity, person and substance are not synonymous. But just that lack of synonymy should give us pause. For the Cartesians, the thinking thing was the soul. It seems this is also Locke’s usage, for he writes at §27: “[T]hat the same soul may, at different times be
united to different Bodies, and with them make up, for that time, one Man . . . .

(347.) We will use the terms interchangeably as we go forward. But we do not need to make this connection to see that throughout Chapter xxvii Locke is at pains to distinguish “person” from this “thinking substance” or “soul.” For example, as we continue the above quote, he writes: “Which however reasonable, or unreasonable, concerns not personal Identity at all. The Question being what makes the same Person, and not whether it be the same Identical Substance, which always thinks in the same Person, which in this case matters not at all.” (§10, 336.)

Locke’s considered position appears to be that same substance does not make the same person, and the same person does not require the same substance. (i) One person can be of many substances, and (ii) one substance be of many persons. On that basis, how can we grant him the license to substitute the term “thinking thing” for ”person” in the definition? Point (ii) brings home the question acutely; if the same thinking thing can be many persons, how can considering oneself the same thinking thing make one a single person?

I think the best way to approach the dilemma is to consider that “same thinking thing” is ambiguous between two readings. In one reading, there are diachronically distinct thinking things, $RC_1$ and $RC_2$ (etc.) which are identified by some rule through which we discover their identity. On a second reading, there is only a single diachronically continuous thinking thing, $RC_0$ which we refer to as the same in different times, although it is the same in and of itself, not on

\footnote{See also §14 and §16, where Locke uses “same immaterial substance” and “same soul” interchangeably.}
account of any external identifying principle. Now, only the second reading gets Locke out of the circle. The first one still requires us to embed a concept of identity—"same thinking thing as"—into our concept of person. And as the identity criterion for persons is equivalent to "identity of thinking things"—there being no conceptual space to distinguish what a thinking thing is from what a person is—we have merely recapitulated it with substitution of equivalent terms. But while the second one gets us out of the circle, it would commit Locke to what he specifically denies—that one thinking thing must constitute one and only one person.

For this would no more make him the same Person with Nestor, than if some of the Particles of Matter, that were once a part of Nestor, were now a part of this Man, the same immaterial Substance without the same consciousness, no more making the same Person by being united to any Body, than the same Particle of Matter without consciousness united to any Body, makes the same Person. (§14, 339.)

I submit that Locke is not entitled to the second reading, and so his use of the locution "same thinking thing" does not appear to absolve him of the charge of circularity as to his definition of "person" or "same person as."

These four points go far to demonstrate the circularity objection as against Locke's definitions of "person" "personal identity." Perhaps they do not determine the conclusion; but if they are not fully probative, it is because we have missed some very non-obvious factor; and it is up to the defender of the relative identity interpretation of Locke to tell us what it is.

Given that the obvious definition of "person" leads to a circularity given the relative identity reading there are two ways we could go. One is to jettison the relative identity interpretation; the other is to keep it and cast about for some
other substantively adequate interpretation of the concept of “person” (other than the diachronic one) which does not enmesh Locke in circularity problems. Given that the relative identity reading is the plain and obvious interpretation of the text, it makes sense at least to consider alternative versions of Locke’s concept of “person.” Below, I will consider three such attempts and show them all to be either formally or substantively inadequate as interpretations of the Lockean text.

B. Contents of Consciousness: The Mackie Position

The first alternative view we will consider is one proposed by John Mackie, to the effect that personhood is constituted by the contents of consciousness and sameness of person is sameness of such contents. In Problems from Locke (1976), John Mackie raises several objections to Locke’s theory of personal identity. The most original and potentially damaging of these is that Locke’s theory is not a theory of personal identity but might better be described as a theory of action appropriation. (183.) As we have seen, the problem has to do with Locke’s memory criterion. Mackie thinks that what constitutes a person is simultaneous consciousness of certain synchronic contents; actions, perceptions, thoughts. These all together are what make a person. But if this is so, then there can be no prior person with whom I can be identified as “same”—it being a contingent truth about memory that is does not provide us with all the detail of prior conscious states, but only selectively reproduces contents with which I am currently concerned. Mackie writes,

Since a man at \( t_2 \) commonly remembers only some of his experiences and actions at \( t_1 \), whereas what constituted a person at \( t_1 \) was all the experiences and actions that were then co-conscious,
Locke’s view fails to equate a person identified at $t_2$ with any *person* identifiable at $t_1$. (183.)

It follows from this, Mackie thinks, that Locke has not provided a criterion for identifying persons over time:

Locke seems to be forgetting that ‘person’ is not only ‘a forensick term, appropriating actions and their merit’, but also the noun corresponding to all the personal pronouns. (183.)

Commentators have dealt shortly and I think, unfairly with the Mackie objection. Curley (1982) derides Mackie’s claim that Chapter 27 fails to give us an account of the use of personal pronouns. “It is, I submit, a mistake to think that it is Locke’s intention to give us a theory of personal identity if that requires giving an account of ‘the noun corresponding to all the personal pronouns.’” (311.) And he goes on to quote Locke to the effect that personal pronouns are sometimes used to refer to the “man” only. (§20.) Now Locke here could (contra Curley) easily be interpreted as trying to correct our use of the personal pronouns. But that is not the main point. Rightly or wrongly, Mackie intends personal pronouns to mean what Locke means by “person” and Mackie is really concerned with the adequacy of Locke’s identification criterion, not his usage of indexicals.

Jolley (1984) has a different complaint. Picking out Mackie’s claim that Locke has provided a mere theory of action appropriation, he goes on to assert that a theory that just appropriates responsibility for prior actions is an adequate theory of the person, all Locke intended and all we should expect of him:

There are signs that Locke is less interested in analyzing our actual use or personal pronouns (which in any case, he would regard as confused) than in solving problems of moral and legal accountability. (129.)
It may then be simply misguided to complain, as Mackie does, that ‘person’ is not just a forensic concept, and that Locke offers us a theory of action appropriation rather than a theory of personal identity. (129-30.)

I have much more sympathy for Jolley’s view. In a way, he is right. I also agree with Jolley’s comment that,

[t]he essential point is that Locke wants to show that our ignorance on key metaphysical issues such as the nature of our mind is no bar to deciding problems of moral responsibility, even when these extend into the afterlife. (130.)

However, I think that Jolley too, fails to hit the mark. Locke is interested in more than a mere positivist account of person qua legal entity. He is interested in combining the positivist with the substantial account. And if Mackie is correct, Locke’s project has failed.

The real force of Mackie’s concern can be put in the form of dilemma. *Per hypothesis*, Locke’s theory of personhood tells us the person should be identified with the full contents of his/her consciousness. Given that, we have two choices. If we accept R₁ then the theory is empty, since we never have R₁. If we accept R₂, we have a workable theory but it is not a theory of personal identity but rather of something much weaker. It is no fit substitute for the substantial Aristotelian and Cartesian theories and it would no longer be topic-neutral (it being clear, I think, that the ability to reproduce an earlier conscious state in memory could be taken as a necessary condition of *res cogitans* being an immaterial substance). Since neither of these options is acceptable, the whole constitutes a *reductio* of the Lockean approach to personal identity.

I think the best way to counter the Mackie position is to deny the “given condition” that personhood is constituted by the experience of a consciousness
with all its present contents. This is just what out diachronic definition of personhood denies. It merely asserts that personhood is the self-conscious perception of something, plus one’s possible memories of being a former person. R-d is a satisfactory interpretation of Locke’s position on this view, because what distinguishes us from others is our memories of being persons, not the contents of our consciousness (as might be necessary if our theory allowed for instantaneous personhood).

C. Plain Consciousness: Atherton’s View

Next we will consider the Atherton view. At first glance, the Atherton position (that Locke’s theory holds that personal identity is identity of plain consciousness) seems to be the reverse of Mackie’s claim the Locke offers us merely “a theory of action appropriation.” But that, as we have seen, is based on a misreading of Mackie; and we are better served, I think, by taking her argument to suggest a close relative of the Mackie view and an alternative if we find it unsatisfactory. It may be seen as answering the question: “So it makes no sense to take Locke as proposing a theory under which the contents of consciousness identify and individuate persons. Might it not be possible, then, that Locke is proposing a theory under which consciousness without consideration of any of its contents does these things?”

If she is right, then there is a viable alternative to the diachronic definitions of personhood and personal identity; and there may be no need to tax Locke with circularity and resort to the causal interpretation I will be proposing.

Her project has two sides; a negative side and a positive side. The
negative side is to hold (a) that Locke does "not offer a way of telling how one should reidentify other people"; and (b) that "his account does not unequivocally depend upon memory." (276) As she puts it, "it turns out that [Locke] never did, in fact, hold a memory theory . . . ." 37

On the positive side, (c) she wishes to assert Locke's real view is "that a person is a single center of consciousness" (274); and what she means by this (d) is "plain" consciousness (277): "... we are to interpret Locke's term 'consciousness' . . . to mean awareness . . ." (278).

Note the close relation of (a) to (c) and (d). Atherton needs to establish (a) for (c) and (d) to be credible, since a plain "consciousness" theory of identity cannot be used for external reidentification. Atherton claims that Locke is not interested in providing a criterion for reidentification of people who have done acts—

Locke's criterion of personal identity is intended only to explain how it is that I am the same person that I was before, not how other people make this judgment, for he says that the prince's soul in the cobbler's body alone is able to recognize that he remains the same person as the prince. Only I can encounter the bearer of my personhood . . . . (275.)

Her evidence for this claim is based on a highly selective reading of the text. She refers only to the prince/cobbler case, where Locke says, "he could be the same cobbler to every one besides himself." But to maintain her position she must ignore many comments by Locke to the contrary; e.g.:

37 "...[I]n this phenomenal sense." But Atherton's argument is not that on Locke's view our memory of being a past person is correctable; rather, it is that Locke's theory is not a memory theory at all.
Person as I take it, 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. (§26, 346.)

In Locke’s famous phrase, “person . . . is a Forensick Term.” What else could he possibly mean than that the concept provides a sortal which can be used as a method for reidentifying other people? As he puts it, “there . . . another may say is the same person.” (346.) Even if Mackie is right and Locke has forgotten “it is more than a” Forensick Term he has probably not forgotten it is at least that!

One possibility is that it is useful but only for questions related to ultimate responsibility, i.e., for disposition of the person at the Last Judgment. Atherton might find this reading congenial, since the fact God could know what is in our minds would not disagree with her claim that for practical purposes identification must be conceived of as being based on the kind of evidence (fingerprints, eyewitnesses accounts, etc.) which are concomitant only to bodies (Locke’s “Man”); Locke does occasionally sound like this might be his intention;

For though punishment be annexed to personality, and personality to consciousness, and the Drunkard perhaps be not conscious of what he did, yet Humane Judicatures justly punish him; because the Fact is proved against him, but want of consciousness cannot be proved for him. But in the great Day, wherein the Secrets of all Hearts shall be laid open, it may be reasonable to think, no one shall be made to answer for what he knows nothing of; but shall receive his Doom, his Conscience accusing or excusing him. (§22, 344.)

And therefore conformable to this, the Apostle tells us, that at the Great Day, when every one shall receive according to his doings, the secrets of all Hearts shall be laid open.38 The sentence shall be justified by the consciousness all Persons shall have, that they

38 1 Cor. 14: 25 and 2 Cor. 5: 10.
themselves in what Bodies soever they appear, or what Substances soever that consciousness adheres to, are the same, that committed those Actions, and deserve that Punishment for them. (§26, 347.)

However, in the end this “exclusive” reading of the text cannot be supported. In an important argument, Locke writes:

But if it be possible for the same Man to have distinct incommunicable consciousness at different times, it is past doubt the same Man would at different times make different Persons; which, we see, is the Sense of Mankind in the solemnest Declaration of their Opinions, Humane Laws not punishing the Mad Man for the Sober Man’s Actions, nor the Sober Man for what the Mad Man did, thereby making them two persons . . . .” (§20, 342.)

Now, admittedly, this example has some controversial aspects. But Locke’s intention here is not one of them. In contrast to the “drunkard” case the §22 “mad man” case foresees the type of occasion where the machinery of human justice can determine that some entity which is the “same man” as a diachronically distinct other is not the same person; in such cases, Locke claims, legal culpability should follow the person, not the man.

At this point, I believe Atherton’s claim (a) is fully refuted. Moreover, if Locke means his concept of person to be “Forensick” in the obvious sense, Atherton’s claim that his theory is not a memory theory but a continuity of plain consciousness theory cannot be accepted. Consciousness as phenomenological experience without reference to any contents is completely inaccessible to external reidentification. So it should not even be conceivable on the Atherton interpretation that a person could be found innocent of a crime because he was not the same person as the one who committed it, although s/he was the same

39 The courts actually refrain from executing madment because the take mental illness to be an excusing condition.
Although the Atherton interpretation would present Locke as having an impractical theory—and, moreover, one that does not meet standards he himself sets for a theory of personal identity—it could still be an accurate reflection of his intention. That is, although the theory “read out” of his text is incoherent, he may not have realized it at the time he was writing—in other words, he blundered. Now it is possible that such a thing might have happened (we argue that something like that occurred with his theory of individuation, supra p. 29). But in order to be convincing, such an argument must eliminate obvious alternative interpretations which are not “non-sequiturs.” Here, the obvious alternative interpretation is the memory theory. Moreover, it is the standard interpretation of the Chapter. Now, although it is not necessarily an easy matter to determine what a person remembers, it is not impossible in principle to do so. So although it is true that “Locke’s solution and his way of conceiving the problem have been found to be problematic” (Atherton (1989), 273) they are not as problematic as Atherton’s own “more nearly coherent” (274) view.

To make a case for her interpretation, Atherton must eliminate the “memory criterion” as a viable competitor. She writes (276) that,

> his account does not by any means unequivocally depend upon memory. There are two longish passages in this chapter on personal identity containing summary statements that I want to quote in their entirety. The first is at 2.27.9, where Locke tells us first what he takes a person to be and then how we can derive from it the idea of personal identity.

After quoting Locke at II. xxvii. 9 and II. xxvii. 17, She writes,

> “The first thing to notice about these passages is that Locke nowhere uses the words ‘memory’ or ‘remember’. In fact, this is
generally the case throughout his discussion of personal identity. He speaks, as he does here, only of consciousness." (1989, 277.)

While the word memory does not appear in her selective reading of the text, that “this is generally the case . . . he speaks . . . only of consciousness” is pure chaff.

Here are some of the passages in which Locke uses the terms “memory” or “recolletion” (in this context, equivalent) in relation to his criteria of personal identity:

But that which seems to make the difficulty is this, that this consciousness, being interpreted always by forgetfulness, there being no moment of our Lives wherein we have the whole train of all our past Actions before our Eyes in one view: But even the best Memories losing the sight of one part whilst they are viewing another, and we sometimes, and that the greatest part of out Lives, not reflecting on our past selves, being intent on our present Thoughts, and in sound sleep, having no Thoughts at all, or at least none with that consciousness, which remarks our waking Thoughts. (§10, 336.)

But yet possibly it will still be objected, suppose I wholly lose the memory of some parts of my Life, beyond a possibility of retrieving them, so that perhaps I shall never be consciousness of them again; yet am I not the same Person, that did those Actions, had those Thoughts, that I was once conscious of, though I have now forgot them? (§20, 342).

If there be any part of its Existence, which I cannot upon recollection join with that present consciousness, whereby I am now my self, it is in that part of its Existence no more my self, than any other immaterial Being. For whatsoever any Substance has thought or done, which I cannot recollect, and by my consciousness make my own Thought and Action, it will no more belong to me, whether a part of me thought or did it, than if it had been thought or done by any other immaterial Being any where existing. (§24, 345.)

Did we know what it was, or how it was tied to a certain System of fleeting Animal Spirits; or whether it could, or could not perform its Operations of Thinking and Memory out of a Body organized as ours is: And whether it has pleased God, that no one such Spirit shall ever be united to any but one such Body, upon the right Constitution of
whose Organs its Memory should depend, we might see the Absurdity of some of those Suppositions I have made. (§27, 347.)

For granting that the thinking Substance in Man must be necessarily suppos'd immaterial, 'tis evident, that immaterial thinking thing may sometimes part with its past consciousness, and be restored to it again, as appears in the forgetfulness Men often have of their past Actions, and the Mind many times recovers the memory of a past consciousness, which it had lost for twenty Years together. Make these intervals of Memory and Forgetfulness to take their turns regularly by Day and by Night, and you have two Persons with the same immaterial Spirit, as much as in the former instance two Persons with the same body. (§23, 344-5.)

I think these passages make the point that Locke's use of the term "memory" is frequent. Moreover, the last quotation (from §23) makes absolutely clear that Locke means his "consciousness" criterion to be a memory criterion:

Make these intervals of Memory and Forgetfulness to take their turns regularly by day and by night, and you have two persons with the same immaterial Spirit, as much as in the former instance two persons with the same body. (§23, 345.)

What clearer evidence could we ask? At the very least, it is not flagrantly misinterpreting Locke to take his criterion of personal identity to be a memory criterion. Moreover, as Atherton herself admits, "memory is a species of consciousness." (279.) And so references to "consciousness" in the text cannot be taken unequivocally to exclude "memory" as a possible interpretation. Indeed, many times when Locke speaks of consciousness and not memory, it is effectively nothing else but memory, a bridge between the present and consciousness of the past:

But let him once find himself conscious of any of the Actions of Nestor, he then finds himself the same Person with Nestor. (§14, 340.)

And thus, by this consciousness, he finds himself to be the same self which did such or such an Action some Years since, by which he comes to be happy or miserable now. (§25, 345-6.)
And to punish Socrates waking, for what sleeping Socrates thought, and waking Socrates was never conscious of, would be no more of Right, than to punish one Twin for what his Brother-Twin did, whereof he knew nothing, because their outsides were so like, that they could not be distinguished; for such Twins have been seen. (§19, 342.)

But though the same immaterial Substance, or Soul does not alone, where-ever it be, and in whatsoever State, make the same Man; yet 'tis plain consciousness, as far as ever it can be extended, should it be to Ages past, unites Existences, and Actions, very remote in time, into the same Person, as well as it does the Existence and Actions of the immediately preceding moment. (§16, 340.)

and many others. So the evidence that Locke is proposing a memory criterion is not “equivocal” at all (as Atherton implies). That position can only be maintained by resorting to a highly selective reading of the text.

Moreover, if Locke’s theory is of consciousness as such—as we have seen, a highly unlikely reading—it could not be a theory of plain consciousness. While it is true that Locke does sometimes refer to the person being conscious sans phrase; e.g., §21: “[I]t is impossible to make personal Identity to consist in any thing but consciousness; or reach any farther than it does” (343), or:

Self is that conscious thinking thing, (whatever Substance, made up of whether Spiritual, or Material, Simple, or Compounded, it matters not) which is sensible, or conscious of Pleasure and Pain, capable of Happiness or Misery, and so is concern’d for it self, as far as that consciousness extends. (§17, 341.)

in all but a few cases, he specifies that consciousness is consciousness of actions; and it is that which makes us the same as some past person; E.g., § 17:

That with which the consciousness of this present thinking thing can join it self, makes the same Person, and is one self with it, and with nothing else; and so attributes to it self, and owns all the Actions of that thing, as its own, as far as that consciousness reaches, and no farther; as every one who reflects will perceive. (341.)
And §§26 and 25, respectively:

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. (346.)

And thus by this consciousness, he finds himself to be the same self which did such or such an Action some Years since, by which he comes to be happy or miserable now. (345-6.)

Or, in the “little finger” example, which Atherton quotes:

For as it is evident in the instance I gave but now, if the consciousness went along with the little Finger, when it was cut off, that would be the same self which was concerned for the whole Body Yesterday, as making a part of it self, whose Actions then it cannot but admit as its own now. Though if the same body should still live, and immediately from the separation of the little Finger have its own peculiar consciousness, whereof the little Finger knew nothing, it would not at all be concerned for it, as a part of it self, or could own any of its Actions, or have any of them imputed to him. (§18, 342.)

and many others. All things considered, and given the overwhelming evidence, I think we are entitled to read “consciousness” the few times it is used without qualification as “memory of prior thoughts and actions,” as the standard view holds, not along the lines of Atherton’s eccentric interpretation.

Does Atherton’s position then merit consideration as something other than a historical interpretation? Perhaps it bears some comparison to Parfit’s monumental Reasons and Persons (1984) as an independent, neo-Lockean theory of personal identity. To demonstrate that doesn’t merit such consideration goes beyond current need, for we have already shown it to be substantially inadequate as an interpretation. However, it might be worthwhile briefly to digress and consider the question.

Independent of the Lockeian text, the Atherton position is formally
inadequate because it fails to handle the Parfit transitivity problem. She considers the problem (at 289), calling it “person fission.” She takes it as an argument against a causal theory of identity (which it is not—see infra., p. 108); however, it certainly is an argument (if we take such things seriously) against her own position.

A copy of consciousness, she claims, would begin to exist at the time of fission. (This distinguishes it from the original.) But the reach of that consciousness ($C_2$) extends just as far back as that of the original ($C_1$). This means it either existed before it began or it had two beginnings. The former case is an obvious contradiction in terms; in the latter case, it implies the consciousness of $C_2$ is not one but two; undermining the unity of consciousness requisite for a theory of plain consciousness to be able to individuate persons.\textsuperscript{40} Neither the Parfit solution nor the Locke solution are available to her. “Simple awareness” cannot be in two places at the same time (the Parfit solution); nor can she block the case by an appeal to topic neutrality (she makes no appeal to the soul’s possible immateriality).

Finally, we might wonder why Atherton could not switch directions and make her theory one of continuous self-consciousness (on my reading, more in accord with Locke’s text). We shall turn to that possibility next.

\textsuperscript{40}This is different from the case (which she also considers) (290) of the severed corpus callosum. In that case, there are discernibly distinct incommunicable consciousnesses. The putative coexisting “persons” in C2 are not distinguishable except (or indeed, two in any way other than) by “their” origins.
D. Self-Consciousness

A third possible alternative to our concept of persons and the memory criterion is “same self-consciousness”: “A person is a center of self-awareness and is the same person as any prior (or future) person who is the same center of self-awareness.” To my knowledge, this view has never been defended as an interpretation of Locke’s theory of personal identity. There is a reason for this. Following Hume, who denied any consciousness of self, later empiricists (those most concerned with the proper interpretation of Locke) have held it doubtful that there is such a thing as a “self-conscious self.” And not only empiricists have so held—as Chisholm (1994, 94) noted,

“The two great traditions of contemporary Western philosophy—‘phenomenology’ and ‘logical analysis’—seem to meet unfortunately at the extremes. The point of contact is the thesis according to which one is never aware of a subject of experience.”

Indeed, this scepticism is limited neither to the west nor to philosophies; for the doctrine of “anatman” (“no-self”) has been a central tenet of Buddhism since its inception. As Hui-neng (sixth patriarch of the Zen school of Buddhism) wrote,

There is no bodhi [enlightenment] tree
Nor stand of a mirror bright.
Since all is void,
Where can the dust alight?
(Price (1990), 72.)

The stanza is a response in verse to that of a competitor for title of Patriarch,

Shen-hsiu, who had written:

Our body is the bodhi tree,
And our mind a mirror bright.

---

41 Russell (in The Problems of Philosophy) maintained that it is merely probable that we have acquaintance with self. (1959, 51.)
Carefully we wipe them hour by hour,  
And let no dust alight.  (Id., 70.)

Of course, not everyone has shared this skepticism. Among the non-skeptics we could reasonably list Locke, who averred that “thinking substances . . . cannot think without being conscious of it” (§13, 338), i.e., that there is a consciousness inseparable from thinking—a “perception that one is perceiving.” (§9, 335.) And so it might seem at least conceivable that the Lockean theory of personal identity could be interpreted as one of continued self-consciousness. We should therefore consider what such an interpretation would entail and how it could be supported or refuted.

There are many different views of what self-consciousness is, and so it would be well first to consider more specifically what it is for Locke. Locke would, I think agree that we are aware of “self” in the sense he often uses the term, i.e., as a thinking substance (either material or immaterial.) However, he would distinguish two kinds of self consciousness: consciousness of one’s own mind’s substance and consciousness of one’s own mind’s operations. So much, I think, is clear from this text:

When we see, hear, smell taste, feel, meditate, or will anything, we know that we do so. Thus it is always as to our present Sensations and Perceptions: And by this every one is to himself, that which he calls self. It not being considered in this case, whether the same self be continued in the same, or divers Substances. For since consciousness always accompanies thinking, and ‘tis that, that 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. the sameness of a rational Being . . . (§9, 335.)

In this text, the “consciousness that accompanies thinking” is that by which everyone is self to himself; and since he elsewhere calls self the “thinking thing”
(§9, 335), consciousness here seems to mean "consciousness of the operations of the thinking thing [self]." But it is not consciousness of soul or our substantiality as an entity (as, e.g., the Hindus might claim). In addition to the text quoted above, he writes that we are "in the dark concerning the Soul of Man." (§27, 347); and at many places in the Essay he denies that we can know what substance is in itself, independent of its affections. So self-consciousness, for Locke is consciousness of the operations (hearing, feeling, smelling, tasting, meditating and willing inter alia) of the mind, but not consciousness of the substantial foundation (if there is one) of these operations. Now, it is interesting that Locke (in asserting self-consciousness) and Hume in denying it are not entirely in disagreement. For Hume writes,

For my part, when 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 never can observe anything but the perception. (Hume (1978), 252.)

Notable among Hume’s "percepts" are "love or hatred, pain or pleasure"—which are also on Locke’s list! The point both seem to be converging on is that we are aware of our mind’s operations, but not of its inner nature.

Whether or not we find this particular view of mind to be an accurate one, I think it fair to say it is Locke’s position. But as such it presents poor prospects for being the centerpiece of a theory of personal identity. On the one hand, not only our perceptions but our perceptions of our perceptions are constantly changing. The sun is out today, it was behind a cloud yesterday; sometimes I’m glad of it, other times annoyed. Where is the continuity in this? (I hope we’re past the point of thinking that because I use an indexical term that guarantees
identity. Cases of ambiguous usage of indexicals are legion.) Even the operations of the mind change, sometimes in fundamental ways; I am aware that my mind works differently now than it did before I started to study philosophy. I am aware that my feelings on being cut off on the highway are different from what they were when I first started to drive. I have these “awarenesses” now, I didn’t have them yesterday—so in what sense is my consciousness of my own mental operations the same?

On the other hand, we might take the view that there is something underlying these individual judgments which is constant (call it soul, substance, “thinking thing”, or what you will). But however true this might be, that there exists such a something which is constant in us is unacceptable as an interpretation of Locke. Time and again Locke denies that continuity of substance is necessary to identity of person.

Nothing but consciousness can unite remote existences into the same Person, the Identity of Substance will not do it. For whatever Substance there is, however framed, without consciousness, there is no Person: And a Carcase may be a Person, as well as any sort of Substance be so without consciousness. (§23, 344.)

So that self is not determined by Identity or Diversity of Substance, which it cannot be sure of, but only by Identity of consciousness. (§23, 345.)

These, and many other comments in like vein, make it clear that for Locke the concepts of “person” and “substance” are completely distinct and independent notions.

I conclude that the difficulties facing a “continuing self-consciousness” interpretation of Locke’s theory of personal identity are daunting. Such a theory would have to posit a continuity that was neither of our perceptions (which
apparently have none) nor of anything *underlying* those perceptions (which apparently Locke would deny could be identified with the concept of person). I have grave doubts that any such project could succeed.

Consideration of the Mackie, Atherton and "self-consciousness" interpretations of Locke's theory of personal identity has eliminated the three most obvious competitors to our "diachronic" criterion of personal identity. This eliminates the only apparent plausible grounds for rejecting the claim that the relative identity interpretation of Locke's theory of personal identity is formally inadequate because it enmeshes the theory in a vicious circularity. At the very least, it shifts the *onus probandi* to those who still wish to defend the interpretation against the charge.

That said, it must be noted that the circularity objection is not necessarily fatal to Locke's project. The circle arises conditionally, on the assumption that Locke's theory is a *Wiggins R* relative identity theory. Personal identity on this view is "same P as" identity, and since "same P as" is part of the definition of "P," we get a circle. But if personal identity is not "same P as" identity, then the charge of circularity may be avoided. What other options are there? As we have noted, there are three possible positions to take with regard to the examples and arguments in sections one to eight of Chapter 27; (a) the relative identity position; (b) the relative individuation position; and (c) the limited types position. As I have just argued, the relative identity view is formally inadequate; as I have argued (*supra*, p. 33) the "relative individuation thesis" is substantially inadequate (it fails to account for all the cases). That leaves us with the "limited types" interpretation. Below, I will propose a variant of that interpretation I call the
"causal" interpretation.

V. The Causal Interpretation

Given that neither the relative identity thesis nor the relative individuation thesis provides both a formally and substantially adequate interpretation of the Lockean text, the principle of charity would seem to require that we give closer consideration to the third alternative, the "limited commitment" view under which Locke is committed to no more categories of identity than the ones he actually mentions. In this spirit, I wish to propose the "causal interpretation" of the text. On this interpretation, Locke's theory of personal identity is best viewed as a theory which utilizes the four Aristotelian causes—material, formal, teleological and efficient—to reidentify objects over time and individuate them. Whatever has all the same causes is the same thing. The causal interpretation does not enmesh Locke in circularity concerns (or at any rate, to the same degree) as the relative identity thesis; and unlike the relative individuation thesis, it properly models all Locke's puzzle cases.

In offering this interpretation, I am not suggesting we entirely set aside the relative identity and relative individuation theses. I believe Locke intended his view to be interpreted as one of these (probably the former). Moreover, Locke does not mention the four causes anywhere in the Essay and had a well-known antipathy for Scholastic notions generally. Nevertheless, I do think there is more to say for the causal interpretation than that it is merely both formally and substantially adequate, while the others are not. I believe Locke was deeply influenced by the Scholastic views; this influence unconsciously shows itself at
many points in Chapter 27. At the same time Locke wanted to set the old
notions aside he was enmeshed in them and could not entirely succeed in
freeing himself. Moreover, as we will see below, the causal interpretation has
great value as a heuristic, casting light on obscure aspects of the text, e.g., the
definition of man and the meaning of §12. I believe time and further research will
reveal the heuristic value of the conjecture in other places as well.

Now let us consider the causal interpretation in more detail.

A. Detailed Consideration of the Causal Interpretation

Locke mentions either three or four types of things: Matter (substance),
life, person (in §7), and spirit (in §15). These can be mapped onto the four
causes, as follows:

**Matter—Material Causation**

In Locke’s view, the matter of an object is the fixed quality of minimal
component parts which go into its making.

And whilst they exist united together, the Mass, consisting of the
same Atoms, must be the same Mass, or the same Body, let the parts
be never so differentially jumbled: But if one of these Atoms be taken
away, or one new one added, it is no longer the same Mass, or the
same Body. In the state of Living Creatures, their Identity depends not
on a Mass of the same Particles; but on something else. (330.)

Anything added to or subtracted from the “mass” changes the body. We can
without much controversy call such addition or subtraction material causation.

**Life and Soul—Formal and Teleological Causation**

According to the causal identity theory, plant and animal identity is a
combination of *formal* identity and *teleological* identity. How does that work? For
the Aristotelian, the soul is the form of living beings of all kinds.
On the Aristotelian view, all bodies are composed of form and matter; the matter is common to all bodies and remains constant in change, while the form explains why a given body has the particular properties it has at a given time. In the special case of living bodies, the soul was taken to be that form. (Garber (1998), 760.) Plants were distinguished from animals on the ground that, while plant souls have only the vegetative faculty, animal souls have both vegetative and motive faculties. (Id., 760.) The vegetative faculties include the ability provide oneself nutrition and to reproduce, while the motive faculties include both movement and the emotional life. Man is also a living being and as such partakes of formal causation. S/he is also different from other living things. Part of that difference is caught in Aristotelian terms by the notion of “rational faculties” of the soul. (Id., 760.) These include will, intellect and self-reflection. The soul of a plant was called a “vegetative soul,” of lower animals a “sensitive soul,” that of a human a “rational soul.” “The rational soul, unique to humans, was called mind (mens) in the strict sense.” (Id., 760.) Locke eschewed talk of occult faculties, and clearly thought we did not know what the soul is in itself, for we are “in the dark concerning these Matters.” (§27, 347.) But I do not think that precludes a commitment to a notion of formal causation which could be filled out by more acceptable explanatory methods (e.g., experiment). We do often speak of systems in ways reminiscent of the notion of formal cause, e.g., respiratory system, reproductive system, neuromuscular system, endocrine system, all of which are backstopped by the powerful explanatory apparatus of modern science. We also speak of “thinking processes,” which (depending on your outlook) might be a special kind of physical system, a non-physical system or even an ordinary physical system.
Life also has a telos, or purpose, and this is part of how it is caused to be. That is, in some sense, it “looks to the end”—has its own end in view as it accomplishes its day to day activities. The end is not only an identifying but also an individuating principle. It holds any single life together as a whole, characterizing it through its span of days. As Locke points out, this is true even of plants:

For this Organization being at any one instant in any one Collection of Matter, is in that particular concrete distinguished from all other, and is that individual Life, which existing constantly from that moment both forwards and backwards, and backwards in the same continuity of insensibly succeeding Parts united to the Living body of the Plant, it has that Identity, which makes the same Plant, and all the parts of it, parts of the same Plant, during all the time that they exist united in that continued Organization, which is fit to convey that Common Life to all the Parts so united. (§4, 331.)

However, it is even more obviously true of animals:

The Case is not so much different in Brutes, but that anyone may hence see what makes an Animal, and continues it the same. Something we have like this in the case of Machines, and may serve to illustrate it. For Example, what is a Watch? ‘Tis plain ‘tis nothing but a fit Organization, or Construction of Parts, to a certain end, which, when a sufficient force is added to it, it is capable to attain. If we would suppose this Machine one continued Body, all whose organized Parts were repair’d, increas’d or diminish’d, by a constant Addition or Separation of insensible Parts, with one Common Life, we should have something very much like the body of an Animal . . . .” (§5, 331.)

The telos of living things, including humans, is not for Locke a fully individuated fate. Rather, it is “specifick” in the old-fashioned sense, that is, of a type fitted to a kind. Machines are made after a pattern to be like others with the same purpose, despite variations in appearance. So perhaps a good scientific
correlate of Lockean telos is the genetic "blueprint" by which we unfold through
the arc of our life form birth through old age and death.

**Person—Efficient Causation**

The basic idea here is that each person is caused to be who they are by
the choices and actions of prior persons. There are two variants on the thesis.
One is the remembering-causal thesis and the other is the causal-remembering
thesis. More specifically, a person is an entity which is conscious and can
remember being caused to be a person (remembering-causal) or caused to
remember being a person (causal-remembering) by choices and actions of some
other person; and personal identity consists in remembering being caused to be
some prior person. For purposes of brevity, I will discuss only the causal-
remembering thesis, leaving consideration of the remembering-causal thesis to a
possible later stage of the work.

**Causal-Remembering**

On this view, we are the same as some other diachronically distinct
person when they made some choice and did some action which caused us to
remember being the person who made the choice and did the action.

**Examples:**

(1) I am writing this in room 415 at the Fondren Library. Although it does
not actually occur to me, I could (if I wanted to) remember that I had made a
choice to drive to Rice University to get a change of scenery. I am the same
person as the one who chose to do this, because that choice and action are
partial causes of my remembering being the person who did them.

(2) This week I am playing the tune of Gerald Finzi’s "Thou Did Delight My
Eyes” over and over in my head. I remember that last week I committed it to my memory in preparation for performance with the Choir at service Sunday. I am annoyed to realize that I am the same person as the one who chose to commit the tune to memory, since my actions last week caused me to remember being the person who learned the tune.

(3) I am waiting on death row for my execution. I remember answering the doorbell that announced the policeman who arrested me. I remember defending myself pro se in court. I am the same person as the people who did these things because their actions caused me to remember being them.

Since we are now working with causal and not relative identity, we need new formulae. The basic changes are that instead of =P= (for “same person as”), we will use =C= (for causally equals); and instead of R for “remembering being” we will use “ε−” (for causal remembering). Revised versions of our formulae are as follows:

VI. Definition of ε−

(∀x)(∀y)(y = ((Py & (tx<ty & yRx)) ⊃ Px))

Ia. Part-Person Criterion

(∀x)(∀y)(Px₀ ⊃ ¬(∃y)(Py & (yε−x₀ & ¬yε−xₙ)))

IIb. Definition of a Person

(∀x)(Px₀ = (Cx₀ & (¬(∃y)(Py & (yε−xₙ & ¬yε−xₙ)) & (∃z)(Pz & zε−x₀))))

IIIb. Personal Identity—Diachronic

(∀x)(∀y)(((Px & Py) & t(x)≠t(y)) ⊃ (x=C=y ≡ (xε−y v yε−x)))

IVb. Personal Identity—Synchronic
\[(\forall x)(\forall y)(((P_x \land P_y) \land t(x)=t(y)) \supset (x=C=y) \equiv (\forall z)((P_z \land t(z) < t(x)) \supset (z x = z x = z y)))]\)

Vb. Combined Criterion
\[(\forall x)(\forall y)(((P_x \land P_y) \supset (x=C=y) \equiv ((t(x)=t(y) \supset (x \in y \lor y \in x)) \lor (t(x)=t(y))) \supset (\forall z)((P_z \supset (z x = z x = z y)))]\)

B. Objections

There will, of course, be many objections to the causal identity thesis.

One major question is whether there is a concept of causality weak enough to allow Nestor's person, choices and actions to possibly be a cause of anyone remembering having been Nestor (cf. §14—"But let him once find himself conscious of any of the Actions of Nestor, he then finds himself the same Person with Nestor." (340.). I believe the answer to this question is yes. Many philosophers conceptualize causation in terms of sets of necessary and sufficient conditions.\(^{42}\) Two tests exist for whether a given condition may be considered such a cause: whether it is a necessary element of a sufficient set of conditions ("NESS") to bring about an effect, or whether it is an insufficient but necessary element of an unnecessary but sufficient set ("INUS") of such conditions.

On either NESS or INUS, Nestor's person, choices and actions may be

\(^{42}\) Richard Taylor (1967) in his "Causation" Encyclopedia of Philosophy article mentions H.L.A. Hart, A.M. Honore, R.G. Collingwood and A.J. Ayer. The motivations for this position include the facts that single causes are hardly ever enough to bring about effects, and many conditions which are "constantly conjoined" with effects are not necessary to bring them about. Many conditions are necessary to make a match light, but gravity (which is always present) does not seem to be one of them.
considered causes of someone remembering being Nestor. They are "conditions sine quibus non"—without which such remembering cannot occur. Of course, in cases like these—where the cause is at such great remove from the effect—it would seem that so much depends on the rest of the necessary conditions that it hardly makes sense to think of Nestor's actions as being even partial causes of one remembering being Nestor. We may pass over, I think, concerns about transmission; for an unbroken chain of events can bring us thus far from Nestor's actions to our becoming aware of them. The problem is that it seems something must have gone wrong in us for such a "mistake" to occur. It must be that we have significant problems in reasoning (editing and evaluating of memory), propriosense and/or in our environment. Put another way, the normal safeguards which restrict my consideration of my possible prior "persons" to those co-existent with my actual "manhood" must for some reason be absent.

However, it must be noted that if it is a mistaken memory—i.e., we are not "really" Nestor; then our R*+ condition will forbid our taking it as real and thus we won't be Nestor. Nevertheless, it is possible that the R*+ conditions will be met by what is to all external appearances a mistaken memory; we have all the available evidence and our reason is working properly. In such cases, there is no ground (on the Lockean view we have articulated) for doubting that we are Nestor.

An example may make this point more plausible. In an actual recent case, Buddhist monks from a Tibetan temple went abroad in search of a child who might be the reincarnation of their deceased Rinpoche (holy leader). Eventually, they found a child of the right age who had voiced memories of a
previous monastic life. With his mother's permission, they took him back to their
temple, where he identified many items of the Rinpoche's property as his own
("That's my book," "that's my bed," "that's my chair."). He's being raised now by
the monks, and is convinced (as are they) that he is the reincarnation of their
beloved spiritual leader. Note that the R* condition and caveats are met in this
case. All possible information pro and contra has actually been given to the boy;
the prior person existed and there are no competitors for his soul. Although the
example might also be used to cast some doubt on the assisted-memory
criterion—since it seems we might be assisted to remember things that didn't
happen to us—I feel confident that Locke would accept it as a case of "same
person," as the causal-remembering thesis suggests (since it could not happen
unless the Rinpoche had been a person).

This is not the same thing as saying, however, that Locke would have
endorsed the view that necessary conditions are equivalent to efficient causes.
That is a modern innovation. Moreover, Locke never espoused a four causal
view. His comments in II.xxvi ("On Cause and Effect and Other Relations") are
limited to what might be called "efficient" causation, and they are all we have to
go on as far as his intentions with respect to causation are concerned. Locke's
articulated views, however, are not so antithetical to the concept of cause as are
those of Hume, who rejected the notion of necessary conditions in toto.

For Locke, "cause" is "that which makes any other thing, either simple
Idea, Substance, or Mode, begin to be; and an Effect is that, which had its
Beginning from some other thing." (II.xxvi.2, 325.) By "began to be" he does not
mean necessarily cause to exist de novo. Rather, he means three things—either
rearrangements of matter into things by natural principle ("generation") or by artificial intention ("making"), or changes in these things ("alteration"):

When any simple *idea* is produced, which was not in that Subject before, we call it *Alteration*. Thus a Man is generated, a Picture made, and either of them altered, when any new sensible Quality, or simple *idea*, is produced in either of them, which was not there before; and the things thus made to exist, which were not there before . . . . (II.xxvi.2, 325.)

It seems quite reasonable, I think, to take it that "memory" is something that can be "caused" in Locke's sense. Moreover, Locke does not place conditions on alteration (as Hume does) which might rule out memories being caused in the ways we have suggested *supra*. So Locke's dicta on causation are at least consistent with one of our causal types, person/efficient causation.

On the other hand, Locke would undoubtably have rejected the more archaic notions of material, teleological, and formal causation. But I do not consider that fact determinative, since my claim is not that he articulated them or would have accepted them as sources of the identity of things. Rather, it is that he was unconsciously influenced by them, that they adequately model his cases, and are heuristically valuable. There is nothing in Locke's dicta on (efficient) causation that contradicts this claim.

Next we will consider two objections to the causal interpretations—the Thiel and the Atherton objections—which are not significant, but which must be included for reasons of completeness (and to avoid accusations of "hand-waving").

**Thiel**

Thiel objects to the limited commitment view both by attacking its positive
claim and defending the relative identity view. His objection to the positive claim is that

[E]ven though the identity of a man is secured by the same very general principle as is the identity of plants and animals, namely life, it is not the same sort of life in each case. Locke distinguishes between different sorts of life: horse-life is (obviously) not the same as oak life, and that is, horse-identity is not the same as oak-identity . . . . (1998a, 239.)

and so on for “man” as well: “Similarly, if our idea of man is that of a ‘vital union of Parts in a certain shape’ (i.e., in a human shape), then ‘as long as that vital union and shape remains . . . it will be the same Man.'” (Id., 240.) But as we have seen, the causal view takes this concern into consideration; indeed, the ability to distinguish between precisely these types of life is a motivating factor for choosing the causal interpretation.

In defense of the relative identity interpretation, Thiel points out that Locke considers a variety of possible definitions of the concept of “man,” suggesting “that our concepts determine what is required for the identity of objects over time.” (1998a, 240.) This we can grant without damage to our thesis, since we approve of the idea that Locke enunciated a relative identity view, but reject it as an interpretation on formal grounds.

The Atherton Objection to the Causal Interpretation

Atherton objects to a causal account of personal identity, as follows:

It might be tempting to try to develop an account of the continuity and individuation of each consciousness in terms of the causal relations among experiences that lead to a distinctive outlook. But, as an account of what individuates a person, this will ultimately conflict with Locke’s claim that identity requires a common origin. Consider a case of fission, by which it seems that at least one of the resulting entities does not share a beginning with the original. Nevertheless, the
causal relations and distinctive outlook will, at least at the time of the
fission, be the same in both. (Atherton (1983), 289.)

It's really hard to give a charitable interpretation of this argument. In a case of
fission, it makes no sense to say "at least one of the resulting entities does not
share a beginning with the original." Lacking some condition leading to
asymmetry (as when an embryo fissions to make "identical" twins) either both
share a beginning with the original or neither do! Moreover, what does "at least
at the time of fission" mean? Only at the time of fission—when there are not yet
two entities—will the "outlook be the same in both." After that the outlook differs.

The most charitable reading is that Atherton—in a rather garbled
fashion—is trying to articulate the question, "How does the causal thesis handle
the Parfit transitivity problem?" There's really no difference in the way the causal
thesis handles it from the way the relative identity thesis handles it. Both y and z
are caused to remember being x by x, so they are the same person as x—but z
is caused to remember being w whereas y is not, so they are not the same as
each other. But by caveat 2 to R*+, neither y nor z can remember being x—since
that would make it indeterminate which of the two possesses the "x" soul (if there
is one). So transitivity is not violated.

**Circularity Redux**

We showed that on the relative identity thesis Locke's criterion of personal
identity is circular, because the definition of person both contains the criterion
and is subsumed under it. However, it is not clear that we have avoided
circularity with the causal interpretation. The reason is as follows. The definition
of person has changed; it now contains the criterion of "efficiently causes."
“Efficiently causes” in turn contains a reference to “remembering being.”

“Remembering being” however, is of a special kind: it is not any old kind of remembering, but that of “being a person.” But to be a person you have to be the same person as another: you have to be efficiently caused by that person, thus remember being that person. The remembering must be of being a person; just “remembering” without that condition cannot make us the same person as something that isn’t a person.

In my paper “Digraphs and Definitions” (Fried (2000)), I propose four criteria by which we might judge the relative informativeness of circular definitions: (1) Distance at which the circle occurs from the apex (or definiendum); (2) Empirical content; (3) Length of path; (4) Complexity of definition. (A brief non-technical summary of the paper’s main findings is attached as Appendix B.)

Our suggested criterion grades significantly better than the original one on three of these criteria. First of all, it has a longer path; it takes three “steps” instead of two to get back to the original definiendum, as follows:

0 Person
1 . . . . . . efficiently causes
2 . . . . . . remembers being
3 . . . . . . person

Put another way, the introduction of the notion of “causation” makes the definition significantly more informative. Moreover, “cause” has itself a complex definition; including terms like “necessary conditions,” and which we have articulated above. Finally, it appears to be a partially empirical circle, since it
models actual processes of memory and self-reidentification. (Our original circle, on the other hand, was a purely formal one.) For all these reasons, the causal interpretation is not viciously circular; it constitutes a significant improvement in this regard over the relative identity interpretation.

**Causal Interpretation of the Prince/Cobbler Case**

A final objection to the causal interpretations brings out the important distinction between the claim that Locke intended or was committed to a causal theory of personal identity and the claim that he was deeply influenced by Aristotelian views, that that influence is apparent at many places in the text and can be used heuristically as an interpretive tool in many others. The objection holds that the causal theory is not adequate to handle the Prince/Cobbler puzzle case. That case (from §15), is as follows:

> But yet the Soul alone in the change of Bodies, would scarce to any one, but to him that makes the Soul the *Man*, be enough to make the same *Man*. 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 Cobbler as soon as deserted by his own Soul, every one sees, he would be the same Person with the Prince, accountable only for the Prince's Actions: But who would say it was the same Man? The Body too goes to the making the Man, and would, I guess, to every Body determine the Man in this case, wherein the Soul, with all its Princely Thoughts about it, would not make another Man: But he would be the same Cobbler to every one besides himself. (340.)

A peculiarity of the text not always noted is that the final clause refers to the "same cobbler." Is "same cobbler" the same man, the same person, or some other sortal category of "cobblers"? Although the last possibility seems antithetical to the causal hypothesis, we can pretty well discount it; "cobbler" doesn't seem like a very useful category for purposes of reidentification.

The real problem lies elsewhere. Locke means by Cobbler either "person"
or "man." (i) If he meant "person", then the force of the "but" in "[b]ut he could be the same Cobbler to every one besides himself" is contrastive; in the sense that the soul in the Cobbler's body does not make another "man" but does make another "person (from the prior Cobbler)." The Prince in the Cobbler's body is the same person to himself and a different person to others. (ii) If he meant "man," then the force of the "but" is concessive; although the Prince's soul would not make another man to others, it would make one to himself (i.e., he's different from the prior man, the Cobbler). On this interpretation, "every one" means "everybody else."

Interpretation (i) is the killer. If Locke meant to say that when a body has a change of soul it is still the same man, then the causal interpretation is blown out of the water: since it holds that the identity of man is in part assured by formal causation, which effectively is continuity of rational soul. Change of soul must mean change of man on the causal interpretation. On interpretation (ii), however, we are still in business; since it allows that although the change in soul (and man) would be undetectable to some, it would nonetheless be a change in man.

Between these two interpretations there is not much to choose. There is only a slight inference which should lean us towards (ii). Interpretation (i) is consistent only with the view of the "man" that makes him/her out to be body only; which we have rejected (supra, p. 27) because (inter alia) it implies that although our personhood cannot be verified by external means, our manhood can; and this nullifies the mad/sober man puzzle case.

This is not a very forceful ground for choosing one interpretation over the
other; as noted above, some commentators do take the (to my mind, inferior) alternative view and cite this passage in support of it. But even if we can not be brought to agreement on this point, it does no damage to my thesis. On the Aristotelian view, a change of soul in any living thing is an impossibility because it is the form of the thing itself. I believe this accounts for the indecision with which Locke approaches the conclusion that the body would determine the man in this case.

C. Heuristic Uses of Causal Interpretation

At the very least, using the causal interpretation heuristically has been useful in bringing the difficulties of this passage into focus. There are two other places where the interpretation is a powerful heuristic tool. One is §12; the other is the concept of “man.”

§12

It is a matter of some interest, I think, that the causal heuristic approach sheds some light on Locke’s opaque comments in §12. He writes,

But the Question is, whether if the same Substance, which thinks, be changed, it can be the same Person, or remaining the same, it can be different Persons. And to this I answer first, this can be no Question at all to those, who place Thought in a purely material, animal, Constitution, void of an immaterial Substance. For, whether their Supposition be true or no, ‘tis plain they conceive personal Identity preserved in some-thing else than Identity of Substance; as animal Identity is preserved in Identity of Life, and not of Substance. And therefore those who place thinking in an immaterial Substance only, before they can come to deal with these Men, must shew why personal Identity cannot be preserved in the change of immaterial Substances, or variety of particular immaterial Substances, as well as animal Identity is preserved in the change of material Substances, or variety of particular Bodies: Unless they will say, ‘tis one immaterial Spirit, that makes the same Life in Brutes; as it is one immaterial Spirit that makes the same Person in Men, which the Cartesians at least will not admit, for fear of making Brutes thinking things too. (§12, 337.)
His point seems to be as follows: materialists clearly think that personal identity can be preserved in change of substance—for them the question whether or not it can be so (with which he starts §12) is a non-issue. It is not a non-issue for the immaterialist, however, who identifies person with a substantial soul. Clearly for someone who believes in a thinking substance (like the Cartesians do), it must be the same substance to be the same self (as when Descartes used the term "personne" to refer to a body/soul complex).

Moreover, it is this sameness of thinking substance which preserves selfhood in persons. So what preserves selfhood in animals? It can't be sameness of matter; but Descartes would not allow that it is sameness of thinking substance (= immaterial spirit) which makes animal sameness either, because that would mean animals have souls.  

So a problem for the Cartesian arises unless s/he is ready to insist that animals are of the same matter throughout their lives—which seems an indefensible position. Now note that this problem is not at all a concern for the non-Cartesian Scholastic. As noted above, a person was usually (after Aristotle) understood to be a substantial entity with a special kind of formal cause, viz., a rational soul. But that soul is not itself substantial on this view. So someone prepared to uphold the doctrine of four causes and the non-Cartesian Scholastic view of person has no dilemma where the Cartesians do.

**The Concept of “Man”**

Another place where the heuristic value of the causal interpretation proves

---

43 Interestingly, it is not the case that all Cartesians deny souls to animals; Leibniz at any rate did not. (Leibniz (1997), 235.)
extremely helpful is in clarifying a confusion over what Locke means by the concept “man.” As Chappell (1983, 76 & 82 fn. 9) notes, Locke in §§6 and 8 seems to endorse to view that “man” means a body and nothing else:

He that shall place the Identity of Man in any thing else, but like that of other Animals in one fitly organized Body taken in any one instant, and from thence continued under one Organization of Life in several successively fleeting Particles of Matter, united to it, will find it hard, to make an Embryo, one of Years, mad, and sober, the same Man by any Supposition, that will not make it possible for Seth, Ismael, Socrates, Pilate, St. Austin, and Cesar Borgia to be the same Man. (§6, 331-2.)

And whatever is talked of other definitions, ingenious observation puts it past doubt, that the Idea in our Minds, of which the Sound Man in our Mouths is the Sign, is nothing else but of an Animal of such a certain Form: Since I think I may be confident, that whoever should see a Creature of his own Shape and Make, though it had no more reason all its Life, than a Cat or a Parrot, would call him still a Man or whoever should hear a Cat or a Parrot discourse, reason, and philosophize, would call or think it nothing but a Cat or a Parrot; and say, the one was a dull irrational Man, and the other a very intelligent rational Parrot. (§8, 333.)

Whereas elsewhere (as in §8) when he writes,

‘[T]is not the Idea of a thinking rational Being alone, that makes the Idea of a Man in most Peoples Sense; but of a Body so and so shaped joined to it; and if that be the Idea of a Man, the same successive Body not shifted all at once, must as well as the same immaterial Spirit go to the making of the same Man. (§8, 335.)

he endorses the view that “man” means a body and soul joined together. Such a quick about-face certainly invites puzzlement; but the turnaround is not so puzzling on the causal hypothesis. The “fitly organized body” (§6) which makes a man is also an “[a]nimal of such a certain Form” (§8); and for the Aristotelian, both the “form” and the organizing principle of a man is the rational soul. The human form and rationality go together; are, indeed, inseparable. So Locke really means by “man” a body and a soul together. My point is not so much that
Locke intended us to think of causality here. Rather, having it in the back of his mind; he did not notice that his reference to “form” was ambiguous and would not carry to others all the connotations of rationality it did to him.

00. Conclusion

From the above considerations, I conclude that although it contains some counter-intuitive implications, Locke’s theory of personal identity is nevertheless a practical criterion for the reidentification of persons over time. Admittedly, it is not (on the relative identity and diachronic concept of a person interpretations) a formally adequate theory since it is viciously circular. However, this inadequacy can be significantly ameliorated (if not eliminated) by replacing the view that Locke is a relative identity theorist with the view that he is an Aristotelian four-causal identity theorist. Although this move has the drawback that Locke nowhere articulates such a theory, I believe that it is implicit in the text; and that further research will reveal many more places where it is a useful interpretive device.

Thus ends our discourse.

ef.
Bibliography


Appendix A: Varieties of Privileged Access

In "Varieties of Privileged Access" (1970), William Alston delineates six levels or degrees by which we might claim that our knowledge of our own mental states might have an "epistemic privilege"—we have an advantage in our knowing them (if indeed we do) over the way we know other kinds of propositions having to do with external events and facts. Bearing in mind that for Alston knowledge is "justified true belief", the six types of privilege are:

Infallibility. Each person is so related to propositions ascribing current mental states to himself that it is logically impossible for him to believe that such a proposition is true without knowing it to be true; while no one else is so related to such propositions. (230.)

Omniscience. Each person is so related to propositions ascribing current mental states to himself that it is logically impossible for such a proposition to be true without his knowing that it is true; while no one else is so related to such propositions. (230.)

Indubitability. Each person is so related to propositions ascribing current mental states to himself that it is logically impossible both for him to believe that such a proposition is true and for anyone to have any grounds for doubting that he knows that proposition to be true; while no one else is so related to such propositions. (230.)

Incorrigibility. Each person is so related to propositions ascribing current mental states to himself that it is logically impossible both for him to believe that such a proposition is true and for someone else to show that proposition is false; while no one else is so related to such propositions. (230.)

Truth-sufficiency. Each person is so related to propositions ascribing current mental states to himself that it is logically impossible both for such a proposition to be true and for him not being justified in believing it to be true; while no one else is so related to such propositions. (234.)

Self-warrant. Each person is so related to propositions ascribing current mental states to himself that it is logically impossible both for him to believe that such a proposition is true and not be justified in
holding this belief; while no one else is so related to such propositions. (235.)

Alston's work is actually intended as substantive philosophy; in his paper he argues that we have no privilege higher than self-warrant for our beliefs about our mental states. But many philosophers (including Locke) have made claims of epistemic privilege for beliefs of this kind, such as of infallibility, omniscience, etc. For example, when Locke writes that it is "impossible for one to perceive, without perceiving that he does perceive" (from §9 of Book II, Chapter xxvii) he is claiming omniscience for our perceptions—we always have the justified true belief that we are perceiving whenever we perceive. Leibniz, on the other hand, would reject omniscience for our perceptions, since he believed we could perceive without being aware of it; and so on. So Alston's taxonomy has value independent of his own philosophic commitments for analyzing and comparing historical claims.
Appendix B: Circular Definitions

Consideration of the few words which are defined in terms of exact synonyms leads to the problem of circularity. Circularity of definition, as we all know, is a bad thing. Quine (in “Two Dogmas of Empiricism” (1971)) rejected the entire notion of analyticity because of the interdefinability of its key terms. He wrote,

Analyticity at first seemed most naturally definable by appeal to a realm of meanings. On refinement, the appeal to meanings gave way to an appeal to synonymy or definition. But definition turned out to be a will-o’-the-wisp, and synonymy turned out to be best understood only by dint of a prior appeal to analyticity itself. So we are back at the problem of analyticity.⁴⁴ (71.)

What illumination can we get out of defining one word in terms of another which is in turn defined in terms of it?

It is true that a circular definition can be unilluminating. But it need not be, if certain other factors are present. Moreover, whether or not circularity is to be avoided, it is always present in definitions as a possibility and sometimes is present in fact. Why not study the phenomenon, and try accurately to describe it? As I showed in my “Digraphs and Definitions” (Fried 2000), such a study is highly illuminating. Philosophers are very concerned with the meanings of words, and often go on to great lengths to specify such meanings. It would be well to have useful tools with which we can see how far they succeed in making

⁴⁴ Quine, unaccountably, did not feel it necessary to examine the interdefinability of the key concepts of reference.
their definitions non-circular, what words they take to be primitive (i.e., non-definable), what kinds of problems their choices of definitions either avoid or create; and many other tasks.

Moreover, the factors which make the study of definition difficult can be largely eased, if not completely removed, with the help of modeling procedures borrowed from the mathematical theory of directed graphs ("digraphs"). Application of the tools of digraph theory can reveal both the strengths and weaknesses of definitions. Most importantly, they provide us with a powerful tool for examining a proposed definition's degree of circularity. But if circularity is, in fact, a matter of degree, then some circles are better than others. If we find that the modeling process and the tools suggested in my paper provide us enough important information to make their usage to analyze definitions a worthwhile activity, perhaps we may be induced to reconsider the received view that all circular definition is bad.

Indeed, insisting on non-circularity in definitions may even be an unnecessary hindrance to the creation of good ones. If the presence of such elements is in fact unavoidable, it will be especially important that we have tools which help us understand the phenomenon and how to ameliorate it.

In my paper, I described four ways in which circular definitions may be distinguished.

1. Distance from the apex. The "apex" is the word which is the original definiendum; in our current case "person." Surely we should prefer circles which involve words in the definiens of person (or at even further remove) than those which bring us back to "person" itself.

2. Empirical content. The empirical content is the degree to which we must refer to some actual activity to understand the circle. The
"Krebs Cycle" is an empirical circle; "bachelors are unmarried men and unmarried men are bachelors" is a purely formal circle.

3. **Length of path.** It should perhaps be considered that circles with a greater number of levels of definition before the reappearance of the "circled" word are to be preferred to ones with fewer levels.

4. **Complexity of definition.** Definitions are hardly ever in the form of single words, usually there are other aspects of the definition which provide some context despite the circle. Obviously, the more complex the definition, the smaller the role the offending circle plays.

It seems we can use these four criteria to grade our circles for objectionableness numerically. Add a point for each generation in the length of path; add a point for each generation of distance from the apex; add the average complexity of the definition; and give two points for empirical content, one for some, and zero for none. Obviously larger numbers will be preferable to smaller ones.

The person/same person circle gets two points for length of path and has two points for average complexity for a total of four. The new person/efficiently caused/remembers being/person circle gets three points for length of path, two for average complexity and one for some empirical content, for a total of six. This makes it significantly less vicious than the original.

Formal details are in the original paper, which is available on request.