In a chapter entitled “Of Identity and Diversity” (Essay II.xxvii) written for the second edition of *An Essay Concerning Human Understanding*, John Locke presents his account of personal identity. It is surely the most famous treatment of that topic in the history of Western philosophy. But while it often draws high marks for originality, it has also been criticized as circular and inconsistent. A common charge of circularity, datable at least to Joseph Butler (1736), is that Locke’s attempt to analyze personal identity in terms of memory fails because memory itself presupposes personal identity. One common charge of inconsistency, datable at least to George Berkeley (1732) and Thomas Reid (1785), is that Locke’s theory of personal identity leads, through the transitivity of identity, to contradictory consequences.

A second common charge of inconsistency, however, concerns Section 13 of the chapter, in which Locke considers the possibility of a “transfer of consciousness” in which “one intellectual Substance” would “have represented to it, as done by it self, what it never did, and was perhaps done by some other Agent.” Locke describes such a transfer of consciousness as involving “a fatal Error” that would “draw Reward and Punishment with it,” and he claims that God’s goodness could be expected to prevent the occurrence of such transfers if they were otherwise possible in the nature of things. It has seemed to many commentators that this section—which I will call “the Fatal Error Passage”—is utterly

---

incompatible with Locke’s own theory of personal identity. Hence, it seems that Locke, in this section, not only discusses but commits a “fatal error.”

I will argue, on the contrary, that the Fatal Error Passage is entirely consistent with Locke’s theory of personal identity, properly understood. I will begin by explaining why commentators have frequently judged the Fatal Error Passage to be inconsistent with that theory, and I will set out four questions that any complete interpretation of the passage must answer. In preparation for answering them, I will describe some of Locke’s key doctrines concerning four related topics: memory, identity, consciousness, and rewards and punishments. I will then explain briefly how these doctrines provide Locke with the resources to avoid both the common circularity objection concerning memory and the common inconsistency objection concerning the transitivity of identity. Next, I will examine two well-known interpretations that aim to acquit Locke of the charge of inconsistency in the Fatal Error Passage—one offered jointly by William P. Alston and Jonathan Bennett, and the other offered by Paul Helm—and I will argue that neither is successful. Finally, I will present what I believe is the correct interpretation of the Fatal Error Passage, answering the four questions about it that any complete interpretation must answer and giving a reading of the entire passage in the light of those answers—a reading that renders it consistent with Locke’s treatment of personal identity. Although my central focus is thus on the interpretation of the Fatal Error Passage, the examination of Locke’s theory of personal identity that is required in order to understand the passage will also serve to demonstrate the overall coherence and promise of that theory.

1. Personal Identity and the Apparent Contradiction of “Fatal Errors”

*Personal Identity.* Locke begins the chapter “Of Identity and Diversity” with a general characterization of identity as a relation that holds between any thing considered “as existing at any determin’d time and place” and itself, even when existing “at another time” (Essay
II.xxvii.1). The leading idea of his approach to identity, however, is that “that which has made the Difficulty about this Relation has been the little care and attention used in having precise Notions of the things to which it is attributed” (Essay II.xxvii.1), and so he quickly turns to the application of that relation to particular kinds of things, as these are determined by particular ideas. After discussing the application of identity to God, finite spirits, particles of matter, compound masses of matter, living creatures, and machines, he draws a crucial distinction between the idea of man and the idea of person. A man is a member of the human species (i.e., a living human animal) that may, but in principle need not, be rational; a person, in contrast, is a rational being that may, but in principle need not, be human. (Because of its centrality to his chapter, I will simply follow Locke’s technical usage of these terms, even though it has the counterintuitive consequence that women are properly describable as “female men.”) Drawing on his previous discussions of finite spirits and bodies, he also distinguishes, from both men and persons, the particular substances—whether immaterial or material—that actually perform the thinking in men, persons, and other thinking beings.

These distinctions allow Locke to extend specifically to persons his general methodological principle that “to conceive, or judge of [identity] aright, we must consider what Idea the Word it is applied to stands for; … for such as is the Idea belonging to that Name, such must be the Identity” [Essay II.xxvii.7]. In pursuance of this approach, he analyzes what it is to be a person, asserting that 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 itself as it self, the same thinking thing in different times and places; which it does only by that consciousness, which is inseparable from

---

2 He also writes, “In this consists Identity, when the Ideas it is attributed to vary not at all from what they were that moment, wherein we consider their former existence” (Essay II.xxvii.1). By this he apparently means that the idea of identity is formed by comparing an idea of a thing as it is at one time with an idea of a thing as it is at the same or another time, but where each state is conceived as a temporal stage of the same enduring thing. This is in accordance with Locke’s general account of the origins of ideas of relations: all such ideas involve comparing two or more ideas without combining them (Essay II.xii.7 and II.xxv).
thinking, and as it seems to me essential to it: It being impossible for anyone to perceive, without perceiving, that he does perceive. When we see, hear, smell, taste, feel, meditate, or will any thing, 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 diverse Substances. (Essay II.xxvii.9)

From this account of personhood as constituted by reason, reflection, and conscious consideration of self at different times and places, he concludes that the temporal extent of a person is determined by consciousness of the past:

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: And as far as this consciousness can be extended backwards to any past Action or Thought, so far reaches the Identity of that Person; it is the same self now it was then; and ’tis by the same self with this present one that now reflects on it, that that Action was done. (Essay II.xxvii.9)

The Apparent Contradiction of “Fatal Errors.” After giving this general account of personal identity, Locke returns more specifically to the relation between the identity or diversity of thinking substances and the identity or diversity of the persons in which they think. In Section 12, he poses a two-part question: “whether if the same Substance, which thinks, be changed, it can be the same Person, or remaining the same, it can be different Persons.” Section 13 constitutes Locke’s answer to the first part of the question, but it also includes a lengthy and puzzling digression concerning “transfer” of consciousness and “fatal errors.” The entire section reads as follows:

3 By Locke’s usual standards for the use of the term ‘substance’, men (and presumably persons) also qualify as substances (Essay II.xii.6). However, in his discussion of personal identity, he consistently uses the term ‘substance’ (and related terms such as ‘same substance’) more restrictedly, to refer more specifically to the substances that compose or help to compose men or persons. While allowing that God has the power to make material substances think, Locke grants that it is “the more probable opinion” (Essay II.xxvii.25) that human thought is performed by an immaterial substance (i.e., a “finite spirit” or “soul”). He conceives of such immaterial substances, on the model of material corpuscles, as spatially located entities persisting through time; unlike material substances, however, they are not extended. See Alston and Bennett 1988 for further discussion of Locke’s use of the term ‘substance’ in this chapter.
13. But next, as to the first part of the Question, Whether if the same thinking Substance (supposing immaterial Substances only to think) be changed, it can be the same person. I answer, that cannot be resolv’d, but by those, who know what kind of Substances they are, that do think; and whether the consciousness of past Actions can be transferr’d from one thinking Substance to another. I grant, were the same Consciousness the same individual Action, it could not: But it being but a present representation of a past Action, why it may not be possible, that that may be represented to the Mind to have been, which really never was, will remain to be shewn. And therefore how far the consciousness of past Actions is annexed to any individual Agent, so that another cannot possibly have it, will be hard for us to determine, till we know what kind of Action it is that cannot be done without a reflex Act of Perception accompanying it, and how perform’d by thinking Substances, who cannot think without being conscious of it. 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 it self, what it never did, and was perhaps 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 representations in Dreams are, which yet, whilst dreaming, we take for true, will be difficult to conclude from the Nature of things. And that 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. How far this may be an Argument against those who would place Thinking in a System of fleeting animal spirits, I leave to be considered. But yet to return to the Question before us, it must be allowed, That if the same consciousness (which, as has been shewn, is quite a different thing from the same numerical Figure or Motion in Body) can be transferr’d from one thinking Substance to another, it will be possible, that two thinking Substances may make but one Person. For the same consciousness being preserv’d, whether in the same or different Substances, the personal Identity is preserv’d. [Essay II.xvii.13]
Antony Flew, in his well-known 1951 article on Locke’s theory of personal identity, characterizes this passage as “a confession of sorts” by Locke that his official theory of personal identity is inadequate. Flew writes:

In his desperation Locke falls on his knees …. But the assistance for which Locke supplicates is beyond the resources even of Omnipotence. For on Locke’s view there could be no sense in his own fear that people might lose or escape their deserts because they remembered doing what they had not in fact done: if anyone can remember doing something then necessarily—according to Locke’s account—he is in fact the same person as did that deed. (164 [1968 Martin and Armstrong version]; boldface added for emphasis)

In a similar vein, J. L. Mackie (1975) remarks that Locke’s appeal to God will not do … for it presupposes that there is something else which really constitutes personal identity, which is the true bearer of responsibility, and which therefore needs to be protected from the unjust effects of a transfer of consciousness. But 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. (184-85; boldface added for emphasis)

Thus it appears that, in addressing the question of whether a person can survive a change in thinking substance, Locke recognizes at the outset of the Fatal Error Passage that his own theory of personal identity entails that a person would be identical with an earlier person having a different thinking substance just as long as the later person had “consciousness” of the actions of that earlier person. However, Locke next appears to worry

---

4 Flew continues:

By making this desperate appeal, Locke both tacitly confesses the inadequacy of his own account of personal identity and provides one more example of a phenomenon already all too familiar to the student of religious apologetic—the hope that the sheer physical power of a postulated God can make contradictions consistent or by itself make utterances to which no sense has been given sensible.

5 Mackie continues, “The actions of which someone thus becomes 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.” More recently, Kenneth Winkler, while showing more sympathy for Locke than does either Flew or Mackie, has allowed that “Flew’s objection [to the Fatal Error Passage] shows that the dominant themes in chapter xxvii—the subjective constitution of the self, and the possibility of objective criticism and adjustment—cannot easily be combined” (Winker 1991: 170).
that a later person having such consciousness but lacking the same thinking substance would be committing a “fatal error,” evidently on the grounds that that later person would not, after all, be the same person as the earlier agent. Yet after invoking God’s goodness as a reason to expect that the kind of fatal error in question would not in fact occur, Locke then returns at the end of the passage, with seemingly remarkable good cheer, to answer his original question in the affirmative—re-affirming the seemingly-just-abandoned claim that persons would survive a change of thinking substance if the same consciousness were preserved—and to re-endorse the original theory of personal identity that entailed it. Even for a philosopher sometimes thought to be inattentive to consistency, such a performance of the general form “p; although q, because not-p; but to return to the question, p” within a single paragraph would be truly breathtaking.

Is this indeed what Locke is doing? In order to answer this question, we must determine the answers to the following four questions, each of which must be addressed by any complete interpretation of the Fatal Error Passage:

1. What is a “transfer of consciousness?”
2. What is the “fatal error” that a transfer of consciousness does or might involve?
3. How would such a fatal error “draw reward and punishment with it?”
4. Why would God’s goodness lead Him to prevent the occurrence of such errors?

2. Some Key Lockean Doctrines

In order to see how Locke would answer these questions, it is essential to examine some of his key doctrines on the topics of memory, identity, consciousness, and rewards and punishments.
Memory. Locke presents his theory of memory in the chapter of the Essay entitled “Of Retention” (Essay II.x). He defines Retention as the “keeping of those simple Ideas, which from Sensation or Reflection [the mind] hath received.” One species of retention is contemplation, by which the mind keeps “the Idea, which is brought into it, for some time actually in view”; the other species of retention is memory, defined as “the Power to revive again in our Minds those Ideas, which from Sensation and Reflection it hath received” (Essay II.x.1). Thus, Locke imposes, by definition, the

Revival in Self Requirement on Memory: Memory requires the revival of a perception that one has oneself actually had in the past.6

He repeats and emphasizes this requirement several times in the course of the chapter (see especially Essay II.x.2,7). He also emphasizes repeatedly (Essay II.x.2-8) that this requirement is compatible with his doctrine that ideas cannot exist unperceived; for although the ideas themselves do not exist when unperceived, the power to revive them may nevertheless do so. This power may be explained through the operation of “organs” related to memory (see especially Essay II.xxvii.29) that allow the “laying up” and “rouzing” of material or immaterial traces (Essay II.5-7).

Immediately after the definition just cited, Locke characterizes memory more fully as “a Power … [of the mind] to revive Perceptions, which it has once had, with this additional Perception annexed to them, that it has had them before” (Essay II.x.2; boldface added for emphasis). Thus, he also holds the

Self-Representation Requirement on Memory: In memory, one not only has a perception, but also represents that this perception was had by oneself in the past.

---

6 Because the publication of the chapter “Of Retention” in the first edition of the Essay predates by several years Locke’s treatment of the self in the later chapter “Of Identity and Diversity,” it is not surprising that he writes of “the mind’s” revival of a perception that “it” has had before. However, given his repeated allowance in the later chapter that persons remember their earlier perceptions despite possible changes of thinking of substance, there can be little doubt that by ‘the mind’ and ‘it’ in this context he means what he later characterizes as the ‘self’—i.e., the conscious thinking person, rather than any particular thinking substance.
Identity. As previously noted, Locke characterizes identity as a relation holding between any thing considered “as existing at any determin’d time and place” and itself, even when existing “at another time”; and as this implies, he holds that all of the things to which identity applies exist in both time and place (Essay II.xxvii.1-2). Spatio-temporal location thus pertains not only to simple and compound “bodies” of various kinds, but also to God (who is everywhere at all times), to “finite intelligences” (which exist in a definite place even if they are immaterial and unextended—see also Essay II.xxiii.20), and even to the modes and relations of all these kinds of substances (for modes and relations are located in the same places as the substances whose modes or relations they are—see Essay II.xxvii.2,28). The universality of spatio-temporal location is important to Locke because it plays a crucial role in individuation:

For we never finding, nor conceiving it possible, that two things of the same kind should exist in the same place at the same time, we rightly conclude, that whatever exists any where at any time, excludes all of the same kind, and is there itself alone. When therefore we demand, whether any thing be the same or no; it refers always to some thing that existed such a time in such a place, which it was certain at that instant was the same with itself, and no other. 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. (Essay II.xxvii.1; boldface added for emphasis)

As the final sentence indicates, he accepts the

Uniqueness of Location: No place can be occupied by more than one thing of any given kind at a time, and no thing can have two different places at a single time.

The Uniqueness of Location shows, according to Locke, how “Existence itself” is the “principium individuationis” that serves to distinguish items of the same kind from one another by “determin[ing] a Being of any sort to a particular time and place incommunicable to two Beings of the same kind” (Essay II.xxvii.3). Things of different kinds, however—for
example, the omnipresent deity and a body, or (presumably) a man and a person—may occupy the same place at the same time without violating this principle.

Just as location plays a central role in individuation at a time, continuity plays a central role in identity through time. Locke first cites the example of

an Atom … existing in a determined time and place: … it is the same, and so must continue, as long as its Existence is continued: for so long it will be the same, and no other. (Essay II.xxvii.3; boldface added for emphasis)

Again, he writes that a plant’s identity lies in

that particular concrete existing constantly from that moment both forwards and backwards in the same continuity of insensibly succeeding parts. (Essay II.xxvii.4; boldface added for emphasis; see also Essay II.xxvii.6 for the application of this point to men and II.xxvii.25 for a reference to the “self” or “same consciousness” as “continued for the future”).

At the conclusion of the chapter, he makes the point completely universal:

[W]hatever 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. (Essay II.xxvii.29; boldface added for emphasis)

This is in keeping with his primary strategy, noted previously, of deriving the specific conditions for the identity of things of particular kinds directly from the ideas of those kinds. Thus, Locke endorses the

Sufficiency of Continued Existence for Identity: Given the existence of a thing of a particular kind, the continued existence of something satisfying the conditions determined by the idea of that kind of thing is sufficient for identity between the original thing and the continuing thing.7

7 This formulation of the principle leaves open, as Locke’s own remarks do, the question of exactly what kind of continuity is required for continued existence. For example, it leaves open the question of whether the continued existence must be entirely continuous temporally (i.e., involving the occupation of every moment, from the first beginning of the continued existence to its final ending, without temporal “gaps”). A fortiori, it leaves open the question of whether the continued existence must be spatio-temporally continuous (i.e., involving the occupation, from first beginning to final ending, of a spatial location at each moment that is either the same as, overlapping, or immediately adjacent to the spatial location occupied at the immediately
Consciousness. Locke defines consciousness in the first chapter of Book II, entitled “Of Ideas in General,” as “the perception of what passes in a Man’s own mind” (Essay II.i.19). However, most of his remarks about this kind of perception occur in the chapter “Of Identity and Diversity,” and especially in Sections 9-14. There he characterizes consciousness more explicitly as “perceiving [that] one does perceive” and asserts firmly that consciousness so defined is “inseparable from thinking, and … essential to it: It being impossible for any one to perceive, without perceiving that he does perceive” (Essay II.xxvii.9).8 Consciousness is able to play an essential role in personal identity for Locke precisely because it is an act of perception whose content is not merely that some particular thing is perceived but also that that particular thing is perceived by oneself—i.e., because Locke accepts the

Self-Representation Requirement on Consciousness: Consciousness always represents a perception as being perceived by oneself.

By its nature, consciousness provides both the origin of, and the basis for applying, the very idea of self; and this is what Locke means when he writes of consciousness that “by this every one is to himself, that which he calls self …. [C]onsciousness always accompanies thinking, and ’tis that, that makes every one to be, what he calls self” (Essay II.xxvii.9, boldface added for emphasis).

Crucially, then, memory (in virtue of the Self-Representation Requirement on Memory) and consciousness (in virtue of the Self-Representation Requirement on Consciousness)
each include a representation of the self as part of their content: the former representing the self as having perceived in the past, the latter representing the self as perceiving in the present. But since all perception is conscious, every act of memory is inseparable from an act of consciously remembering; and an act of consciously remembering will contain more than one representation of self (or at least, more than one use of the same representation of self). Because memory always represents oneself as having previously had a certain perception in the past, the consciousness that accompanies memory will always represent oneself as currently perceiving both that an earlier perception occurred and that that perception was perceived by oneself. When a complex representation of this kind occurs, one ipso facto represents oneself (i.e., one’s present self) as being the same self as the self that perceived an earlier perception or (because voluntary actions are themselves perceived) performed an earlier action. The present self thereby appropriates the self who was conscious of the earlier perception or action as its self—i.e., as the same self that has the present consciousness. In doing so, the present self appropriates the perceptions and actions of the earlier self as well (Essay II.xxvii.25-26), considering itself—as Locke puts it in his definition of personhood—“as it self, the same thinking thing in different times and places.”

At least when it occurs in genuine memory, Locke holds, such a past-appropriating consciousness allows a rational, reflective being to “repeat” an earlier act of perception with what may properly be called “the same consciousness”; and this sameness of consciousness is sufficient for the earlier and the later person actually to be the same person:

---

9 It may be noted that if the memory of a past perception content also happens to include memory of the consciousness that accompanied that content in the past (i.e., if the memory now represents not only that the content was perceived by the self in the past, but also represents that the content was consciously perceived by the self in the past), then the present act of memory will also contain a third representation of self (or at least will make a further use of an existing representation of self).

10 As I am using the term ‘appropriation’, it requires only a representation that a perception or action was, is, or will be had or performed by oneself—i.e., by the self that one now is. Such a representation need not be veridical, for it may be (as we shall later have occasion to see) a misrepresentation, on Locke’s view. A representation of a perception or action as had or performed by oneself will be veridical, however, whenever it constitutes a genuine memory and hence occurs within the same consciousness. At Essay II.xxvii.10, Locke alludes to a possibility that there may be “thought” in sleep without “that consciousness, which remarks our waking Thoughts.” This should not be understood as a claim that thought
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 the same consciousness can extend to Actions past or to come … the same consciousness uniting those distant Actions into the same Person, whatever Substances contributed to their production. (Essay II.xxvii.10; boldface added for emphasis)

Thus, Locke accepts the

**Sufficiency of Sameness of Consciousness for Personal Identity:** The revival of a past idea with sameness of consciousness, in a being with reason and reflection, is a sufficient condition for identity with the person whose idea is revived.

The sense of appropriation provided by sameness of consciousness is, in Locke’s view, psychologically very powerful:

Had I the same consciousness, that I saw the Ark and Noah’s Flood, as that I saw an overflowing of the Thames last Winter, or as that I write now, I could no more doubt that I, that write this now, that saw the Thames overflow’d last Winter, and that view’d the Flood at the general Deluge, was the same self … than that I that write this am the same my self now whilst I write … that I was Yesterday. For as to this point of being the same self … I [am] as much concern’d … 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. (Essay II.xxvii.16; boldface added for emphasis)

Indeed,

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. 

(Essay Iixxvii.18; boldface added for emphasis)

Thus, he accepts the

Irresistibility of Appropriating Consciousness in Memory: The appropriation of an earlier perception or action by the same consciousness in memory is psychologically sufficient for an irresistible conviction that one is the same person who perceived or performed it.

As Locke’s final observation about the “finger” example indicates, a representation through consciousness of oneself as the same self that performed an earlier action is also necessary for one to appropriate the action as one’s own. He makes the same point again more generally at (Essay II.xxvii.26):

Whatever past Actions [the self] cannot appropriate to that present self by consciousness, it can be no more concerned in, than if they had never been done.

Rewards and Punishments. Locke’s primary discussion of rewards and punishments occurs in the chapter of the Essay entitled “Of other Relations” (Essay II.xxviii). There he defines reward and punishment as “Pleasure or pain, attending our observance, or breach of the Law, by the Decree of the Law-maker” (Essay II.xxviii.6). As this definition indicates, rewards and punishments must be actual pleasures or pains that result from (at least supposed) obedience to or violation of a law. Locke distinguishes three kinds of laws, which differ from one another both through their particular lawmakers and through their characteristic sanctions of pleasure and pain. The divine law, establishing “sins and duties,” is made by God and results in pains or pleasures in the afterlife. The civil law, establishing “criminal” and “innocent,” is made by the commonwealth, and results in the forfeiture or retention of life, liberty, or goods in the present life. Finally, the law of opinion or reputation, establishing “vice” and “virtue,” is made by societies of human beings, and results in the pain of “disgrace” or the pleasure of “credit.”
In the chapter “Of other Relations,” Locke writes of laws, rewards, and punishments as applying simply to “Men”; this is not surprising, for the chapter antedates by several years the explicit development of the distinction between the concepts of *man* and *person* involved in Locke’s writing of the chapter “Of Identity and Diversity” for the second edition of the *Essay*. In “Of Identity and Diversity, however, Locke expressly clarifies the basis for the appropriate application of law, reward, and punishment, writing that “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, and not mattering what becomes of any substance not joined to, or affected with that consciousness.” (*Essay* II.xxvii.18). Personal identity can provide a proper basis for rewards and punishments, in Locke’s view, because the past-directed relation of appropriation is paralleled by a similar but future-directed relation: one feels *self-interested concern* (i.e., “there is some thing that is himself that he is concerned for and would have happy”) when one represents, with present consciousness of oneself doing so, perceptions and actions as being perceived and performed in the future by *oneself*—a future self that can, in turn, appropriate one’s present self (*Essay* II.xxvii.25). Hence, one can represent future happiness or unhappiness as occurring to *oneself* for deeds that will have been performed by *oneself*, and one can deliberate in the knowledge that *oneself* will, in the future, be gratified or regretful for the rewarded or punished actions of *oneself* in the present. For this reason, only ‘person’ is, for Locke, “*a forensick term appropriating actions and their merit*” (*Essay* II.xxvii.26).\(^\text{11}\) Thus, he endorses the

*Forensic Uniqueness of Personality:* Only identity with the *person* who performed an earlier action, and no other identity, is a necessary condition for just rewards and punishments.

But while personal identity is a necessary condition for just rewards and punishments, it is not a sufficient condition. One further necessary condition is appropriate *liberty* to

\(^{11}\) See also *Essay* II.xxvii.22, which states that “punishment [is] attached to personality.”
perform the act (Essay II.xxi.56), which includes the power to “suspend” determination in order to deliberate. Presumably because of the psychological power of act appropriation, as exemplified in the Irresistibility of Appropriating Consciousness in Memory, Locke evidently holds as a third requirement for fully just reward and punishment not merely that the person rewarded or punished be the same person that performed the action rewarded or punished, but also that, at the very time of reward or punishment, the person actually represent himself or herself as being the person who performed the action.  

This further requirement is implied in Locke’s remark that

in the Great Day, wherein the Secrets of all Hearts shall be laid open, it may be reasonable to think, no one shall be made to answer for what he knows nothing of; but shall receive his Doom, his Conscience accusing or excusing him. (Essay II.xxvii.22)

A few pages later, he returns to the Great Day of Judgment again, this time citing the apostle Paul and explicitly invoking justification:

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. 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. (Essay II.xxvii.26; boldface added for emphasis)

3. Circularity Concerning Memory and Inconsistency Concerning Transitivity

These key doctrines provide Locke with the basic resources needed to avoid two well-known objections to his theory of personal identity: the objection that it is circular in its use of memory and the objection that it is inconsistent with the transitivity of identity.

Locke does allow that, because of the difficulty of proving consciousness of particular deeds, human courts may sometimes have to punish “with a Justice suitable to their way of Knowledge” even in the absence of consciousness of the deed punished (Essay II.xxvii.22).
Circularity Concerning Memory. Because it provides at least the primary source of sameness of consciousness between past and present, memory plays an essential role in Locke’s account of personal identity. It is sometimes objected, however, that this role involves him in a dilemma. If memory is understood in a weak sense, as consisting only in seeming to oneself to have perceived or done something that was in fact perceived by someone, then it seems insufficient to guarantee identity; for it seems that people can falsely imagine having perceived or done many things that may, coincidentally, have been perceived or done only by someone else. If, on the other hand, memory is understood in a more stringent sense that requires seeming to oneself to have perceived or done something that was in fact actually perceived or done by oneself, then it seems that this further requirement, at least, can be analyzed only in terms of one’s personal identity with the earlier agent or perceiver. But because such an account of memory presupposes personal identity, it cannot be used to explain it. Thus, memory in the first sense (“phenomenal memory” or “seeming memory”) is by itself too weak to play the central role in an account personal identity, while memory in the second sense (“genuine memory” or “veridical memory”) is sufficiently strong to yield the right results but only at the cost of circularity, allowing us to analyze personal identity only in terms of a concept of memory that must itself be analyzed in terms of a prior concept of personal identity.13

In fact, however, Locke’s account of memory provides him with an avenue to avoid this dilemma. On the one hand, Lockean memory is not too weak to guarantee personal identity. Because of the Revival in Self Requirement on Memory, memory in Locke’s sense goes beyond mere phenomenal memory—contrary to the interpretations of Flew, Mackie, and others—in two potentially distinguishable ways. First, it must be the revival of an earlier perception; and regardless of similarity or felt familiarity, a later perception is not, prima

---

13 For discussion of the circularity objection, see Perry 1975b and Perry 1975c; see also Flew 1951, Wiggins 1976, and the portion of Butler 1736 reprinted as Section 5 in Perry 1975a.
facie, a revival of an earlier perception unless, at a minimum, its occurrence results from the activation of traces causally derived from that earlier perception. (Analogously, a merely coincidental similarity to earlier performances, no matter how striking, is not sufficient for the revival of a Broadway play; a revival of a play must exploit a continuing power stored in traces—such as scripts and scores—to produce performances similar to earlier performances of the play of which it is a revival.) Second, memory in Locke’s sense requires the revival of a perception that one has oneself had in the past. This latter requirement cannot reduce to mere phenomenal memory for Locke, for the following reason. When two different persons in two different places perceive the same content at the same time, any later phenomenal memory of that content must be equally a phenomenal memory of each. If phenomenal memory were sufficient for identity, therefore, a person having the phenomenal memory would be identical with each of the two earlier persons and hence would have existed at two different places at the same time—contrary to the Uniqueness of Location. In addition, the two earlier persons, although not identical with each other, would each be identical with the later person and hence would both exist at the same place at the same (later) time—again contrary to the Uniqueness of Location.

Of course, Locke’s requirement that memory be of a perception that one has oneself had in the past threatens circularity—the other horn of the dilemma—unless it can be spelled out in terms that do not presuppose personal identity. However, the concept of an idea’s revival through causal traces in a way that involves a dual representation of self (first as having perceived an earlier perception or action and second as presently perceiving that it so perceived) offers at least a basis for such an account. Locke could refine the account by further specifying the causal requirements for revival, by stipulating additional requirements for memory in addition to revival, or both. To be sure, he does not actually develop such an account; and the difficulties of doing so are many. Furthermore, since he regards memory as sufficient for sameness of consciousness, and sameness of consciousness as sufficient for
personal identity, he must take care that his account of memory not allow for memory
“fission” cases (in which the perceptions or actions of a single person are remembered by
two distinct persons) or memory “fusion” cases (in which the perceptions or actions of two
distinct persons are both remembered by a single person), on pain of violating the
Uniqueness of Location. Nevertheless, the strategy of refining or supplementing revival as a
requirement for memory offers the prospect of a middle course between a conception of
memory that is too weak and a conception of memory that could not be invoked without
circularity.

Inconsistency Concerning Transitivity. It is often objected that Locke’s account of
personal identity is inconsistent with the transitivity of identity—i.e., with the principle that
if X is identical with Y and Y is identical with Z, then X is identical with Z. The best-known
version of this objection is Thomas Reid’s discussion of the example of a “Brave Officer.”14
In the example, a young boy is flogged for robbing an orchard; the boy then grows into a
brave officer, who takes the enemy’s standard in battle while still remembering the flogging;
and the officer then becomes, in advanced age, a general who remembers taking the enemy’s
standard but can no longer remember the flogging. Reid interprets Locke’s claims about
conscious appropriation as entailing that a person P2, existing at time t2, is identical with a
person P1 existing at an earlier time t1, if and only if P2 can remember at t2 the action or
perception of P1 at t1. On this interpretation, the general is identical with the brave officer,
and the officer is identical with the boy, but the general is not identical with the boy. Yet by
the transitivity of identity, if the general is identical with the officer and the officer is
identical with the boy, then the general is identical with the boy. It follows, then, that the
general is identical with the boy; and Locke is thereby caught in a
contradiction.

14 See Reid 1785 III.4.6, reprinted as Sections 6 and 7 in Perry 1975a; see also Grice 1941 and Quinton 1962.
Stewart 1997 traces the “brave officer” example to an earlier use by George Campbell. For what appears to be
the earliest version of the objection itself, see Berkeley 1732 VII.8.
It is sometimes proposed that Locke could avoid this contradiction simply by rejecting the transitivity of identity, at least in its application to persons. But while this would prevent the derivation of the contradiction in question, it would not constitute a remedy acceptable to Locke. For if the general and the boy are each identical with the brave officer, then it follows that both occupy whatever time and place the brave officer occupies; and if the general and the boy are nevertheless not identical with each other, then they will constitute two different things of the same kind (namely, persons) occupying the same place at the same time—contrary to the Uniqueness of Location.

Locke can avoid the objection, not by rejecting the transitivity of identity, but rather by rejecting the view that a person \( P_2 \), existing at time \( t_2 \), is identical with a person \( P_1 \) existing at an earlier time \( t_1 \) only if \( P_2 \) can remember at \( t_2 \) the action or perception of \( P_1 \) at \( t_1 \). It is true, for Locke, that one will find oneself appropriating an earlier action or perception if and only if one represents it as done or perceived by oneself—typically, and most forcibly, by remembering it. In addition, he evidently holds that one is justly rewarded or punished for an act only if one is actually appropriating it at the time of reward or punishment. But although Locke implies that personal identity reaches only as far as it can “be extended” or can “reach” by consciousness (e.g., Essay II.xxvii.9-10, 14, 23), he does not say that the extension by consciousness of a present person into the past is always limited to what the present person now remembers or even can now actually remember. On the contrary, following an extended discussion of the “reach” of consciousness, he concludes that to be conscious of “any of the Actions of Nestor” is to “find” oneself to be “the same Person with Nestor” (Essay II.xxvii.14; boldface added for emphasis)—which is much more than finding merely one single and discrete act of Nestor’s also to be an act of one’s own. This conclusion is just what one should expect, for to become conscious of an action of Nestor’s is, on Locke’s account, to become convinced that one is the same self as Nestor and ipso facto to become prepared to represent any of the actions of Nestor’s self—properly
incorporated into Nestor’s self at whatever time and by whatever means—as actions performed by one’s own self; and this must include even actions that Nestor once remembered but that one cannot now specifically remember. If one cannot at present remember an action (i.e., cannot revive the perception of its performance with the representation that it was performed by self) but is at present convinced by external evidence that one is nevertheless the same self that performed it, then one will necessarily represent it as performed by oneself, but with somewhat less immediate force and conviction than is present in memory. Such a difference would be comparable to that between the forcible conviction of seeing a tree and merely having an opinion, based on other evidence, that a tree exists. Like the opinion that an unseen tree exists, the opinion that a now-forgotten perception or action genuinely belongs to oneself (because genuinely belonging to an earlier person who is oneself) may well be true.

That Locke actually holds this general doctrine—i.e., the doctrine that conscious memory of an earlier perception or action also extends the history of the person to whatever other perceptions or actions are implicated in sameness of consciousness with that earlier perception or action, regardless of present ability to remember them—is strongly suggested by his pointed limitation of the two cases he discusses in which personal identity fails despite identity of thinking substance or man. The first such case is that of total permanent erasure of traces:

As to the … question, “whether the same immaterial substance remaining, there may be two distinct persons?” which question seems to me to be built on this, whether the same immaterial being, being conscious of the action of its past duration, may be wholly stripped of all the consciousness of its past existence, and lose it beyond the power of ever retrieving it again; and so as it were beginning a new account from a new period, have a consciousness that cannot reach beyond this new state. (Essay II.xxvii.14; boldface added for emphasis)

The second is that of alternating but entirely distinct and incommunicable consciousnesses:
Could we suppose two distinct incommunicable consciousnesses acting the same body, the one constantly by day, the other by night … I ask … whether the day and the night man would not be two as distinct persons, as Socrates and Plato? … [I]t is evident that an 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 night, and you have two persons with the same immaterial spirit …. (Essay II.xxvii.23)

In both of these cases, Locke suggests, there are two different persons precisely because no perception or action of the one person can properly be appropriated by any conscious memory of the other person. But if remembering any one perception or action of a person is ipso facto to incorporate into one’s history all of the other perceptions and actions of that same person (whether available to present consciousness or not) then Reid’s general, in remembering the deed of the brave officer, also incorporates into his personal history the action of the boy, which was remembered by the brave officer. Thus, the general simply is the same person as the boy, and there need be no conflict with the transitivity of identity.¹⁵

4. Two Defenses of Locke’s Consistency Concerning “Fatal Errors”

We are now in a position to assess two previous attempts to reconcile the Fatal Error Passage with Locke’s theory of personal identity.

Punishing Innocent Substances. The criticisms offered by Flew and Mackie presuppose that the “fatal error” that God is called upon to prevent consists in a person’s having “consciousness”—which they identify with mere phenomenal memory—of certain actions

¹⁵ This amounts to saying that the ancestors of memory-based relations—often proposed (e.g., Grice 1941, Quinton 1962, and Perry 1975a Section 9) as providing a way to amend Locke’s theory of personal identity so as to avoid objections derived from the transitivity of identity—are in fact available to Locke without amendment of his theory.
that were not in fact performed by that person. Since they interpret Locke as also holding
that consciousness of earlier actions—again in the sense of mere phenomenal memory—is

**sufficient** for personal identity, they regard Locke’s admission of the possibility of such
“fatal errors” as inconsistent with his own theory. William Alston and Jonathan Bennett
(1988), however, propose an entirely different interpretation of the “fatal error.” On their
interpretation, Locke **consistently** holds that mere phenomenal memory of an earlier action is
sufficient for being identical with the person who performed that action. Rewarding or
punishing the later *person* who possesses such consciousness is therefore not itself unjust,
and the person punished or rewarded commits no error, fatal or otherwise, in taking himself
or herself to be the person who committed the act. Alston and Bennett cleverly observe,
however, that one cannot punish a *person* without at the same time inflicting punishment
that will be felt by the *thinking substance* that is then *constituting* or helping to constitute
(by providing the “seat of consciousness” of) the person. Hence, in the case of a “transfer of
consciousness” from one thinking substance to another, a later thinking substance may be
punished for what *that thinking substance* never did, even though that substance has
*consciousness* (i.e., phenomenal memory) of it. This, according to Alston and Bennett—and
not the punishment of a person for what that *person* never did—is the injustice that God
must prevent, since its possibility is not ruled out by Locke’s theory of personal identity.

Alston and Bennett thus provide definitive answers on Locke’s behalf to the four
questions posed earlier concerning the Fatal Error Passage:

1. A “transfer of consciousness” will consist of any case in which one thinking
   substance seems to remember a perception or action actually perceived or
   performed by another thinking substance.

2. The “fatal error” involved will lie in one thinking substance believing that it
   has performed an action that was in fact performed not by that thinking
   substance, but by another thinking substance.
3. This error will “draw reward and punishment with it” because reward and punishment do properly belong to the person who performed an earlier action.

4. A good God could be expected to prevent such fatal errors from occurring because the just reward or punishment of the person—reward or punishment that it might even be unjust to omit—could only be achieved by the simultaneous unjust reward or punishment of a thinking substance for what that substance never did.

This interpretation of Locke has at least two promising aspects. First, it provides an account of transfer of consciousness that is compatible with the possibility of persons surviving a change of thinking substances; hence, it is compatible with his positive answer to the question with which the Fatal Error Passage begins and ends. Second, it provides an explanation for his choice of the rather generic term ‘sensible Creatures’ in his central assertion that God, “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.” For thinking substances themselves, as well as the persons who think with them, might well qualify as “sensible Creatures”; and hence, his remark can be understood as concerned with the justice of rewarding and punishing such substances.\textsuperscript{16}

Nevertheless, there are two serious objections to this interpretation. The first is that it conflicts with Locke’s doctrine of the Forensic Uniqueness of Personality, according to which personal identity is the only kind of identity necessary for the just application of rewards and punishments. As we have already observed, Locke states explicitly that “all the right and justice of reward and punishment” is “founded on personal identity” (Essay

\textsuperscript{16} One might also see an entirely different kind of injustice, one to the person, in punishment that occurs after a change of thinking substance—not punishment of the wrong person, but punishment of the person for faults of character that he or she might no longer possess. For a change of thinking substance might well involve acquiring a thinking substance more disposed to virtue, or to the habits of mind (such as calm deliberation prior to acting on one’s present desires) that Locke thinks are conducive to virtue. However, Locke shows no evidence of being concerned about issues of this kind. Indeed, his previously-noted emphasis on the motivating power of self-interested concern and reciprocal conscious acts of action-appropriation as the basis for just reward and punishment makes little obvious room for such considerations.
II.xxvii.18; boldface added for emphasis); the question of the justice of punishing a thinking substance is, for Locke, simply the question of whether it now constitutes or helps to constitute a person who could be justly rewarded or punished. Consider, by way of analogy, the closely related case of *men*. Men are, *prima facie*, as much “sensible creatures” for Locke as thinking substances would be. Yet men as Locke describes them (i.e., as human animals) cannot survive a change of body. Hence, his account of the resurrection—in which persons are justly rewarded or punished by God for their previous actions, even though they have received new bodies—entails that there is *no* injustice in punishing a man for what that *man* never did, so long as the new man constitutes the *same person* as the person who performed the deeds in question. Indeed, the point can be made more directly from Locke’s text itself, without appeal to analogy with the just punishment of men; his own description of the punishment to be incurred on the Great Day of Judgment, cited previously, specifies quite explicitly that identity or diversity of *substance* is irrelevant to its justification, for 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. (*Essay* II.xxvii.26; boldface added for emphasis)

The second objection is that the Alston-Bennett interpretation cannot make sense of Locke’s description of the *process* that produces the “fatal error,” given his acceptance of the *Self-Representation Requirement on Consciousness*. Consider again the following portion of the Fatal Error Passage:

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 it self*, what it never did, and was perhaps 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 representations in Dreams are, which yet, whilst dreaming, we take for true, will be difficult to conclude from the Nature of things. And *that 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. (Essay II.xxvii.13; boldface added for emphasis)

The Alston-Bennett interpretation requires that the reference of ‘it self’ be the same thinking substance and that the referent of ‘some other Agent’ be some other thinking substance. But as we have previously observed, Locke is emphatic that consciousness represents perceptions and actions only as perceived and performed by a self, not as perceived or performed by any particular substance. Moreover, he has already indicated in his description of consciousness (Essay II.xxvii.9) what a self is: the self is the person, as he emphasizes again at Essay II.xxvii.26 (“Person, as I take it, is the name for this self”). For a thinking substance to have something represented to it as done by it self is simply to have it represented as done by its self, the same self—i.e., person—whose activity its thinking presently helps to constitute; for a “thinking intelligent Being … can consider it self as it self … only by that consciousness … [in which] consists personal Identity” (Essay II.xxvii.9).

Again, he writes of persons (and also, in chapters of earlier composition, of men) as “Agents,” but he nowhere writes of thinking substances themselves as “Agents.” In suggesting that a thinking substance might represent something as being done by “it self” that was in fact done by “some other Agent,” Locke can only be raising the prospect that it represents as done by its own present self or person what was in fact done by some other self or person. This—and not an attribution to the wrong thinking substance—must be the “fatal error” in question. I conclude that the Alston-Bennett interpretation, while ingenious and initially appealing, is ultimately unsuccessful.

Persons as Continued “Thinking Substances.” Paul Helm (1979) defends Locke’s consistency in the Fatal Error Passage as follows:
What Locke is here supposing is that the goodness of God will prevent one person (what he here calls a thinking (or intellectual) substance) from being conscious of (i.e., remembering) what another intellectual substance did. (176)

Like Flew and Mackie, and unlike Alston and Bennett, Helm maintains that the “fatal errors” that God is to prevent would consist in a person’s having “consciousness” of certain actions that were not in fact performed by that person. According to Helm, however, Locke is right to suppose that his own theory does not rule out the possibility of such errors, because the identity of a Lockean person through time is not merely a matter of phenomenal memory (contrary to the interpretations of Flew, Mackie, and Alston and Bennett), but rather requires the spatio-temporal continuity of an individual “consciousness.” Hence, it might well be possible for a person to seem to remember (“have consciousness of”) something that that person did not actually do—because it was not done by any consciousness that is spatio-temporally continuous with the person, but was instead done by some other person.

Helm, too, provides answers on Locke’s behalf to each of our four questions about the Fatal Error Passage:

1. A “transfer of consciousness” will consist in any case in which one person/thinking substance seems to remember having performed an action that was in fact performed by some other person/thinking substance.

2. The “fatal error” involved will lie in one person/thinking substance believing that it has performed an action that was in fact performed not by it, but by another person/thinking substance.

3. This error will “draw reward and punishment with it” because “human courts sometimes punish,” erroneously, on the basis of “false memories.”

4. God’s goodness would prevent such fatal errors from occurring because it would be unjust for one person/thinking substance to be punished (or rewarded) “on the grounds provided by a misremembering” (177) for what another person/thinking substance did.
This interpretation has at least two promising aspects as well. First, unlike the interpretations of Flew, Mackie, and Alston and Bennett, it allows Locke to make a distinction between phenomenal memory and veridical memory in a way that is in keeping both with his *Revival in Self Requirement on Memory* and with the pointed analogy he draws in the Fatal Error Passage between fatal errors and “representations in dreams.” Second, it is compatible with his doctrines of the *Sufficiency of Continued Existence for Identity* and the *Sufficiency of Sameness of Consciousness for Personal Identity*—indeed, it offers a necessary condition of at least some *prima facie* plausibility for “sameness of consciousness” in terms of (spatio-temporally) continued existence.

Nevertheless, Helm’s interpretation is also subject to two objections. The first is that its answer to the third question, concerning the drawing of rewards and punishments, seems insufficient. Because a just God would surely not reward or punish a person for the deeds of some other person, Helms allows that a “fatal error” of the kind he describes would not draw any *divine* “reward and punishment” with it. In an attempt to find rewards and punishments that *would* follow such a “fatal error,” he points to the fact that criminal courts (i.e., courts of what Locke calls “civil law,” establishing “criminal and innocent”) are sometimes led to punish unjustly—for example, as a result of a false confession—through failure to recognize the falsity of a defendant’s phenomenal memory. Yet the prevention of such criminal punishments as these seems no more central to God’s concerns than does the prevention of unjust punishments based on the dishonest testimony of others—something that Locke does not describe God as taking any special care to prevent. Preventing unjust *rewards* in criminal courts resulting from false phenomenal memory seems an even more unlikely aim of special divine intervention. Moreover, such rewards would rarely be so great as to constitute matters in which “the Happiness or Misery of any of [God’s] sensible creatures is concerned”; for Locke defines *happiness* not merely as pleasure but rather as “the utmost Pleasure we are capable of,” and *Misery* not merely as pain but as “the utmost Pain” (*Essay* 28).
II.xxi.42). They are, he specifies, “two extremes, the utmost bounds whereof we know not; ’tis what Eye hath not seen, Ear hath not heard, nor hath it entered into the Heart of Man to conceive” (Essay II.xxi.41, quoting I Corinthians 2:9).

The second objection is that Helm’s interpretation requires that Locke use the terms ‘thinking substance’ and ‘intellectual substance’ interchangeably with ‘person’ in the Fatal Error Passage. But this is contrary to Locke’s clearly established usage throughout the chapter, where ‘thinking substance’ is always used to designate the particular immaterial or corporeal substances that perform the act of thinking. Locke is consistent and emphatic in distinguishing these from persons, and the stated point of the very section in question is to establish that the continuity or identity of a person does not entail the continuity or identity of a thinking substance. Thus, Helm’s interpretation cannot explain why Locke both begins and ends the Fatal Error Passage by giving a positive answer to the original question: “Whether if the same thinking Substance … be changed, it can be the same person.”

5. Transfer of Consciousness and “Fatal Errors”

Four Questions Answered. How, then, would Locke answer the four primary questions about the Fatal Error Passage? The first question concerns the nature of a “transfer of consciousness.” As we have seen, Locke defines consciousness itself simply as an act of perceiving that one perceives, an act in which (as the Self-Representation Requirement on Consciousness dictates) one represents that something is perceived by oneself. Hence, a “transfer of consciousness” occurs whenever one thinking substance represents its self (i.e., the self that it now helps to constitute) as having had a perception that was, in point of fact, previously represented by another thinking substance as being perceived by its self (i.e., the

17 Although the Fatal Error Passage contains Locke’s only use of the term ‘intellectual substance’ (“why one intellectual Substance may not have represented to it, as done by it self, what it never did, and was perhaps done by some other Agent … will be difficult to conclude from the Nature of things”), the phrase seems to be only a verbal variation of ‘thinking substance’. And this is just what one would expect, since Locke consistently uses the adjective ‘intellectual’ to mean having or pertaining to the faculties of the understanding.
self that that other substance then helped to constitute). A “transfer of consciousness” from one thinking substance to another, as just defined, may or may not involve the sameness of consciousness that (by the Sufficiency of Sameness of Consciousness for Personal identity) conserves personal identity. Of course, a “transfer of consciousness” will conserve personal identity if the “transfer” is an instance of genuine memory. For in that case, the later consciousness will represent (in accordance with the Self-Representation Requirement on Consciousness and the Self-Representation Requirement on Memory) that it is itself “perceiving something that was previously perceived by itself,” and (by the Revival in Self Requirement on Memory) this perception will occur as the result of a process of reviving an idea that was in fact previously perceived by that same self or person. Of course, whether or not there can actually be genuine memory of perceptions that were originally perceived by another thinking substance depends—as Locke himself insists—on the unknown mechanisms by which thinking things might acquire, retain, or transfer their powers to revive perceptions. Here, however, is the outline of one obvious explanation of how such personal-identity-preserving transfers could occur: Memory traces may be stored in a brain that is in causal interaction with whichever immaterial thinking substance is currently doing the thinking in a given man. The replacement of one immaterial thinking substance by another might then allow the new immaterial thinking substance to take on the role of the earlier substance in the same continued individual consciousness—including in this role its ability to access some or all of these stored memory traces. The new thinking substance would thereby acquire the power once possessed by the previous thinking substance to “revive” or “renew” certain ideas as memories.

Turning to the second question, what then is the “fatal error” that might infect some transfers of consciousness? In the course of addressing the question of whether persons can survive a change of thinking substance, Locke is naturally brought to address the more general question of how “transfers of consciousness” might occur and to recognize that,
while some such transfers may preserve personal identity, others may not. A “fatal error” is an error because it involves a transfer of consciousness of the kind that does not preserve personal identity—that is, it always involves a false and merely phenomenal or seeming “memory” of an action that was in actuality performed only by another person or self. Like the conscious perception of a genuine memory, the perception involved in such an error is an act having the content: “I myself am (currently) perceiving that an earlier perception or action was perceived and that that thought or action was perceived by myself.” Unlike the case of consciousness of a genuine memory, however, the earlier perception or action in the case of a “fatal error” was in fact perceived by some other person—and so not by any previous self sharing sameness of consciousness with the present self having the seeming memory. Such errors are possible because one can mistakenly represent oneself as having perceived what one has not really perceived—as would surely occur, for example, if one represented as having been perceived by oneself a perception that someone else was genuinely remembering and which had no causal relation at all to any substance currently constituting oneself. In order to see why Locke calls such errors “fatal,” however, we must turn to the third question.

The third question is this: “How would a “fatal error … draw Reward and Punishment with it?” As we have seen, Helm proposes one answer—such an error may lead to erroneous judgment and hence to unjust punishment by a criminal court. If the charge were a capital one, such an error could even be “fatal” in one literal sense of that term—death-causing—at least in a case of punishment, if not of reward. It is far more likely, however, that Locke has in mind a different seventeenth century sense of the term ‘fatal’—namely, that in which it means having to do with one’s destiny or fate. In order to see how erroneous transfers of consciousness could be “fatal” in this sense, we must recall not the civil law but Locke’s law of opinion or reputation concerning vice and virtue, with its reward and punishment of credit and disgrace; and we must observe the relation of that law to the Irresistibility of
Appropriating Consciousness in Memory. As we have observed, Locke holds that the experience of action-appropriation involved in remembering an earlier action as performed by oneself is irresistible or indubitable. But since Locke holds that genuine memory is simply a kind of representation of the past, the representations in an erroneous transfer of consciousness through false phenomenal memory could be expected have the very same irresistible phenomenological character. A good God would of course not wish to impose external punishment or reward in response to an erroneous transfer of consciousness; but Locke emphasizes, as we have already observed, that on the Great Day of Judgment each person’s own “Conscience” will necessarily “accuse” or “excuse” for each action seemingly remembered. An erroneous transfer of consciousness thus involves, by its very nature, an irresistible conscious experience of the appropriation of actions—actions for which one will therefore punish oneself with internal disgrace or reward oneself with internal credit. For although Locke’s original discussion of the law of reputation establishing vice and virtue emphasizes the opinion of other members of one’s society, there is no reason why one’s own opinion of oneself should be of lesser punitive or rewarding force—indeed, if anything, it should be much greater, on the grounds that acute awareness of it is unavoidable. Such errors, if not prevented or corrected, would indeed be “fatal”—for the internal reward or punishment resulting from them on and after the Day of Judgment would be part of one’s eternal fate.

Finally, we can see why Locke should suppose that divine goodness would require that God forestall such fatal errors, were they otherwise to be possible in the nature of thinking substances. For although a just God would not wish to bestow eternal rewards and punishments on persons for actions they had not performed, the psychology of action appropriation itself guarantees that these errors will produce unjustly rewarding or punitive

---

18 In this respect, the indubitable force of memory might be compared with the force of the sensation of searing heat that Locke describes as rendering indubitable the sensory judgment that one’s hand is in a real glass furnace (Essay IV.xi.8).
states of mind in a person’s consciousness. Furthermore, should these states of mind occur on or after the Great Day of Judgment, they would be of just the kind—namely, pleasures and pains in the afterlife—specified for obedience to and violation of the divine law, even though they would not be received for real obedience or violation of those laws. Moreover, because such rewarding or punitive states of mind at the Day of Judgment are likely to be central to one’s thought for all eternity, they bear directly on “Happiness and Misery,” in Locke’s technical senses of those terms. God would therefore be especially concerned to prevent “fatal” errors of this kind, for they would threaten to make the most damaging mockery possible of divine law.

_A Consistent Reading of the Fatal Error Passage._ We are now in a position to provide a consistent step-by-step reading of the Fatal Error Passage. Locke begins the passage by considering the question of whether a person can survive a change of thinking substances. He restricts this question, for the moment, to thinking _immaterial_ substances:

13. But next, as to the first part of the Question, Whether if the same thinking Substance (supposing immaterial Substances only to think) be changed, it can be the same person.

Locke indicates that the answer to this question depends on how and whether consciousness can be transferred:

I answer, that cannot be resolv’d, but by those, who know what kind of Substances they are, that do think; and whether the consciousness of past Actions can be transferr’d from one thinking Substance to another.

He next observes that if “sameness of consciousness” were a matter of the literal persistence of a single “Action,” it could not survive a change of substances; this is because actions are a kind of _mode_, and modes are individuated on the basis of the substances of which they are modes. But the conscious memory that underlies the sameness of consciousness sufficient for personal identity is, for Locke, a matter of present modes of thought representing earlier modes of thought; and for this reason, personal identity may be
something that can survive a change of substance. However, this general consideration about
the sameness of consciousness as constituted by memory—namely, that it involves the mere
*representation* of a distinct earlier mode of thought—raises an unanticipated and previously
unaddressed question, about the possibility of *false* representation. Thus, he writes:

> I grant, were the same Consciousness the same individual Action, it could not:
> But it being but a present representation of a past Action, why it may not be
> possible, that that may be represented to the Mind to have been, which really
> never was, will remain to be shewn.

This general possibility of false representation, in turn, raises yet a further and more specific
question—namely, whether a thinking thing might falsely represent itself (i.e., its *self*) as
having perceived, not just something that it did not actually perceive, but something that
another *self* did happen actually to perceive. Because consciousness is a perception of
oneself as perceiving, it is a “reflex” act of perception, and one that he has earlier claimed
(*Essay II.xxvii.9*) is inseparable from perceiving; hence, the possibility of one person or
agent representing itself as having perceived or done something that was actually perceived
or done only by another self likewise depends on the nature of the processes underlying this
particular kind of reflexive perception. In fact, in the absence of specific knowledge about
these processes, it must be allowed as possible that a thinking substance *may* represent to
itself as done by *itself* (i.e., by the *self* that it represents itself to be) an action that was not
performed by the same individual *self* which its thought helps to constitute, and was instead
performed by another person. Such an occurrence would, in itself, be no more surprising
than the false representations of the external world that occur in dreams (which also
sometimes involve representing oneself as doing what happens actually to be perceived or
done only be someone else). So he continues:

> And therefore how far the consciousness of past Actions is annexed to any
> individual Agent, so that another cannot possibly have it, will be hard for us to
determine, till we know what kind of Action it is that cannot be done without a
reflex Act of Perception accompanying it, and how perform’d by thinking
Substances, who cannot think without being conscious of it. 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 it self, what it never did, and was perhaps 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 representations in Dreams are, which yet, whilst dreaming, we take for true, will be difficult to conclude from the Nature of things.

A misrepresentation of this kind would involve a person in misappropriating (i.e., erroneously treating as one’s own) an earlier act that he or she did not in fact perform. Hence, if the act in question were contrary to the *law of vice and virtue*, the person would incur at least the unjust punishment of self-disgrace. Such an unjust state of affairs, at least as it would affect one’s ultimate fate in the afterlife, is something that God could be expected to prevent, a reflection that leads Locke to write:

> And that 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.\(^{19}\)

Up to this point in the Fatal Error Passage, Locke has written in accordance with his initial assumption that perception and thinking in humans are performed by immaterial

---

\(^{19}\) Locke’s use of the phrase ‘that it never is so’ is rather surprising, for erroneous transfers of consciousness do occur occasionally in earthly life, and even with punitive or rewarding psychological consequences. Nevertheless, Locke’s failure to acknowledge such cases parallels his consistent failure to acknowledge, in his discussions of knowledge by sensation (*Essay IV.xi*), the existence of actual sensory illusions and hallucinations (outside of the special case of dreams). Indeed, he connects his optimism about the actual reliability of memory with his optimism about the actual reliability of the senses in *Essay IV.xi.11*:

> As when our senses are actually employed about any object, we do know that it does exist, so by our memory we may be assured that previously things that affected our senses have existed. And thus we have knowledge of the past existence of several things, of which our senses having informed us, our memories still retain the *ideas*; and of this we are past all doubt, so long as we remember well.

Worries about the actual occurrence of false phenomenal sense experiences and the actual occurrence of false phenomenal memory experiences are both entirely absent from this passage. In any case, however, the later phrase ‘as far as the Happiness or Misery of any of his sensible Creatures is concerned in it’ has the effect of limiting the scope of Locke’s claim—or at least his interest—in the Fatal Error Passage to the cases of “utmost Pleasure” and “utmost Pain” involved in eternal rewards and punishments.
substances. He now reflects that erroneous transfers of consciousness might pose an even greater problem if thinking were instead performed by *material* substances:

How far this may be an Argument against those who would place Thinking in a System of fleeting animal spirits, I leave to be considered.

It is indeed more difficult to explain how an alleged resurrection of the person with a new body would avoid fatal errors on this materialist alternative. Locke remarks in *Essay II.xxvii.15* that “we may be able without any difficulty to conceive, the same Person at the Resurrection, though in a Body not exactly in make or parts the same which he had here, the same consciousness going along with the Soul that inhabits it.” But in the absence of an immaterial thinking soul-substance to endure from death to resurrection, the traces underlying genuine memory must presumably be somehow lodged in an arrangement of material substances—despite the disintegration of the person’s human body.

Finally, Locke returns to the answer to his original question:

But yet to return to the Question before us, it must be allowed, That if the same consciousness (which, as has been shewn, is quite a different thing from the same numerical Figure or Motion in Body) can be transferr’d from one thinking Substance to another, it will be possible, that two thinking Substances may make but one Person. For the same consciousness being preserv’d, whether in the same or different Substances, the personal Identity is preserv’d. [*Essay II.xvii.13*]

This is exactly what Locke should say; for unlike those erroneous transfers of consciousness involved in fatal errors, a transfer of the same consciousness preserves personal identity by preserving the same person through time. Thus, Locke’s positive answer to his original question is entirely consistent with his discussion of the possibility that some potential transfers of consciousness—namely, those that would not transfer the same consciousness—would constitute “fatal errors” drawing rewards and punishments with them in a way that a good God would be likely to prevent.
Conclusion. If the interpretation I have offered is correct, Locke does not contradict himself in the Fatal Error Passage, and the passage itself is no “fatal error.” On the contrary, in coming to understand more fully the theory of personal identity in which Locke’s claims in that passage are embedded, we discover the depth of its resources for meeting other well-known objections as well. Locke’s theory draws on debatable claims about the nature and content of both memory and consciousness as containing representations of self. It also requires further specification than I have provided here, particularly in connection with the necessary conditions for revival, memory, continuing existence, and sameness of consciousness, before it can be fully assessed. I hope to have shown, however, that, in addition to its originality, it also has a good deal more coherence and promise than is usually supposed.20

20 I have benefited greatly from discussion of earlier versions of this paper with Geoffrey Sayre-McCord, Margaret Atherton, David Owen, William Uzgalis, Edwin McCann, Kenneth Winkler, Gary Hatfield, Peter Unger, Cian Dorr, Hartry Field, John Richardson, and Christopher Hill, as well as with audiences at a symposium of the 2000 American Philosophical Association Pacific Division Meetings, at the University of Pennsylvania, and at New York University.
Bibliography


