Can I Be Transplanted into a Computer?

John N. Hooker
Carnegie Mellon University, john@hooker.tepper.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/tepper
Part of the Economic Policy Commons, and the Industrial Organization Commons

This Working Paper is brought to you for free and open access by Research Showcase @ CMU. It has been accepted for inclusion in Tepper School of Business by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
Can I Be Transplanted into a Computer? *

J. N. Hooker
Graduate School of Industrial Administration
Carnegie Mellon University, Pittsburgh, PA 15213 USA

February 1989
Revised May 1991

Abstract
I examine whether the notion of transplanting a person’s consciousness into a computer makes sense, no matter how advanced the technology. I conclude that it does, but arriving at this answer leads us to rethink our concept of selfhood in a way that can have far-reaching repercussions. For instance, it is to some extent a matter of choice whether we ascribe personhood to a computer (or a human body), and no person can survive the passage of time because there is no personal identity over time.

1 Introduction
I have heard computer enthusiasts say that we will one day be able to transfer a person’s consciousness to a machine and thereby achieve immortality. Such a transplant could also be seen as the ultimate goal of prosthetic medicine. It is not obvious, however, that this whole notion makes sense, no matter how advanced may be the technology. It may be incompatible with the very concept of personhood.

I will argue that transplantation to a computer does make sense, but arriving at this answer requires that we rethink our concept of selfhood in a way that leads to radical revisions of it. Indeed, one benefit of our encounter with computer technology is that it may encourage us to develop a more adequate way of looking at ourselves—one that could have far-reaching implications for our ideas of mortality.

My question really has two parts. (1) Can a computer be a person? (2) Can a person be transplanted into a computer? The first part asks what traits are required for personhood, and whether a computer can have them. I will conclude

that a computer can, in principle, have them all, but to draw this conclusion I adopt an analysis whereby personhood is not a property inherent in an object but is a trait with which we can, to some extent, “choose” to endow an object. The second part of the question asks whether, granted that a computer can be a person, it can be the same person that once occupied a human body. For this an analysis of personal identity, what it means to be the same person over time, is in order. I will defend an analysis that makes such a transplant conceptually possible but that differs radically from ordinary notions of personal identity.

By “computer” I mean a digital computer, not any contrivance whatever. For instance, I do not mean a flesh-and-blood facsimile of a human brain. A digital computer may, however, have any sort of sensors and output devices its designers believe are necessary to support consciousness. It may be implanted in a robot or even in a (brainless) body of flesh and blood. As a technical feat this may be forever impossible. Worse, the computation speeds needed for a digital computer to reproduce human intelligence may violate physical law. But I will grant for the sake of argument that the technological barriers are superable. My concern is with conceptual possibility.

2 Can a Computer Be a Person?

The concept of a person is confusing partly because it is really a cluster of concepts. Persons have, at some stage in their lives, a number of splendid characteristics simultaneously: intentionality, reflective consciousness, intelligence, moral agency, feelings, and perhaps a sixth, more mysterious property I will describe later.

Too often, personhood is identified with one or two of these characteristics, with the result that certain remarkable creatures (smart chimpanzees, smart computers, etc.) are classified either as fully persons or fully nonpersons, both with misgivings. But when we come to see that a creature may have some but not all of the requisites of personhood, we can make precise our feeling that it is “almost but not quite” a person. We can also decide whether it makes sense to talk to it, love it, have moral obligations to it, etc., depending on which characteristics apply.

2.1 Intentionality

The most primitive characteristic, which I somewhat loosely call intentionality, is in a way the most arcane. It is related to the “intentionality” introduced by the philosopher Franz Brentano and modified by Edmund Husserl and is only remotely related to the English “intention” [2, 11]. It is best explained by example. Consider a dog trained to hunt a prey, or a person who believes in unicorns, or even a guided missile that homes in on a target. In each case, behavior is in some sense “directed to” an object, called an intentional object,
whether it be a prey, unicorn or target. The dog, believer or missile somehow has the object “in mind.”

The mentalistic activities thinking, believing, hoping, loving, wondering, etc., are clearly intentional. But it is important not to read too much into intentionality. When a person hunts for something, he generally “desires” to find it in some sense, “reasons” about how to find it, and so on. So, intentional activities in people usually involve more advanced traits that nonetheless can be conceptually distinguished from intentionality in my sense.

One way to analyze intentionality, championed by D. M. Armstrong [1] and others, is as a disposition to behave grounded in a particular type of causal relation. A homing missile is disposed, under certain conditions, to “head toward a certain factory” – it behaves “as though” it were seeking the factory. Moreover, it does so because of a feedback mechanism whereby the missile continually ascertains its position and corrects its course accordingly. This combination of homing behavior and feedback causality qualifies the missile as having intentionality in Armstrong’s sense.

Daniel Dennett [4] has suggested that this limited sense of intentionality can also be analyzed as a way of interpreting behavior. Sometimes we can more effectively explain and predict an object’s behavior by taking an intentional stance toward it; that is, by ascribing to it goals and rationality and viewing its behavior as an attempt to achieve its goals. We might better predict how a chess-playing computer is going to move, for instance, by regarding it as an “intentional system,” to use Dennett’s phrase. But in so doing we take no position on whether the computer “really” thinks, is conscious, or has moral obligations.

It seems clear that computers can have intentionality in either of these senses. Computers can and do exhibit goal-directed behavior that is controlled by a servo-mechanism. The physical basis for a computer’s behavior might be so complex that we can predict its action more reliably by regarding it as an intentional system.

2.2 Reflective Consciousness

A central trait of personhood is reflective awareness or self-consciousness, an ability that is often misunderstood. Obviously the lower animals are aware of goings on in and around their bodies, can feel pain, and so on. But it is unlikely that any but a handful have second-order awareness—that is, are aware of their own awareness. Human beings, however, have a sense of self, in that we can reflect upon our own awareness. We not only suffer pain but apprehend our own suffering of the pain. People often conceive of self-consciousness as one’s having a “little person” or “homunculus” inside his head. This homunculus is a way of picturing the faculty whereby one monitors his own awareness. Immanuel Kant argued two centuries ago that there is a necessary connection between self-consciousness and historical consciousness—the ability to experience oneself
as lasting through time [14].

It is perhaps man's reflective consciousness that most markedly distinguishes him from other creatures. It endows him with ability to narrate and therefore theorize about past events, as well as to fantasize and therefore plan for the future. His reflection on his own sensations and feelings leads to artistic expression. His projection of the future compels him to face his mortality, tempts him to question the meaning of his fleeting existence, and drives him to religion for salvation from his predicament.

Insofar as awareness consists of being informed about what is happening, it is unclear why a computer cannot have second-order awareness. If monitoring circuits can allow a computer to become aware of events outside itself, then it can become aware of its own awareness by monitoring its own circuits. It may be objected that complete self-awareness is impossible for a machine, since such would require monitoring circuits for its monitoring circuits, and so on in an infinite regress. But it is unclear that human beings have complete self-awareness in this sense. It seems enough that one be able, on demand, to add to the events in his consciousness at the moment an awareness of those events. A machine can in principle do this by turning on another monitor.

But I confess a lingering suspicion that there is more involved in second-order awareness than being informed about one's awareness. Human beings are not only able to know what they are experiencing but somehow participate directly in its sensuous quality. We savor our experiences. Up to now I have analyzed reflective awareness as a behavioral disposition (ability to report on one's internal state) and as a causal connection (the connection between the internal state and the monitor). But some irreducible something may be missing from this formula, and this is the mysterious property I mentioned earlier. I will postpone its discussion until I have covered the remaining traits of personhood.

2.3 Intelligence

A person must not only undertake intentional activities, and reflect on his awareness, but do so with intelligence. Intelligence is notoriously hard to define in detail, but it at least involves adeptness at intentional activities. These activities include solving problems, using language, and simply getting along in the world from day to day. A computer-driven robot that can find only one tool in a particular corner of his creator's laboratory is not intelligent. But one that can find any of a hundred tools in any of a billion rooms around the world may have something approaching human intelligence. We might also want to say that intelligence entails ratiocination, or some kind of reasoning process. Perhaps a robot that can find tools but cannot reason about where to find them is not yet intelligent, but only adept. Ratiocination, however, is a difficult concept that is best discussed in connection with agency, below.

Nowadays an enormous amount of effort is spent to develop artificial intelligence (AI), and aside from some chess-playing software, the results have
fallen short of early expectations. Critics such as Hubert Dreyfus have not only pointed out the failures but have advanced arguments that the whole AI enterprise is wrong-headed [6]. Unfortunately Dreyfus’ polemics regarding AI’s technical failures often divert attention from the underlying philosophical argument. I am sympathetic with this argument, which is informed by his background in the Continental school of philosophy, and I have in fact written a related critique based on Anglo-American philosophy [9]. But it is important to understand that these critiques are directed only against developing AI on an information processing model—what John Haugeland has called “good old-fashioned AI” [8]. This is the view roughly that the brain embodies a formal system, and any embodiment of it (such as a computer) would be equally intelligent. The critiques of this model do not refute the notion that a digital computer might, for instance, solve a system of differential equations that simulate the working of a human brain and thereby generate intelligent responses. For, this does not require that the brain instantiate a formal system. The simulator solves differential equations and therefore must embody a formal system (a mathematical model of electrochemistry). The brain, however, does not solve the equations and does not need to embody a formal system. It is the phenomenon the equations describe.

This leaves the question, however, as to whether a computer that can simulate the operation of the brain and thereby generate humanlike behavior is intelligent in the same sense that human beings are. It obviously can if intelligence can be analyzed as a purely dispositional trait, and much of the debate about artificial intelligence in effect centers around this issue. The view that intelligence and other mental traits are dispositional is known as philosophical behaviorism, which was popular two or three decades ago. Philosophical behaviorism holds that a statement such as, “Hubert understands Chinese,” can be analyzed as a conjunction of counterfactual conditionals of the form, “if the circumstances were A, B and C, then Hubert would perform behaviors X, Y and Z.” On this view, a computer that can, say, converse fluently in Mandarin Chinese understands what it is saying in the same sense that Mandarin Chinese understand what they are saying. This behavioral analysis seems implicit in the famous “Turing test,” which says that if, on conversing via teletype with a computer and a human being, I cannot tell which is which, then the computer understands language [20].

John Searle, on the other hand, has advanced a well-known argument against the Turing test [19]. Suppose, he says, that a person ignorant of Chinese is installed in a room whose walls are lined with manuals containing exhaustive and detailed instructions for how to carry on a conversation in Chinese. A Chinese query on a slip of paper is dropped into the room through a slot, and the operator inside eventually returns a slip bearing an intelligent response. Searle observes that the operator in no sense understands Chinese. Nor does the system consisting of the room and the operator under-
understanding Chinese one whit better.

Again, Searle's argument is directed not against machine intelligence in general but against the information processing model—the view that if the Chinese room embodies the brain's formal system, it can understand as well as the brain. Thus he, like Dreyfus, militates only against "good-old-fashioned AI," which he calls "strong AI," and seems comfortable with a dispositional/causal analysis of understanding.

Nonetheless Searle's position seems to rule out the sort of digital intelligence I would recognize, because of the type of causality he believes it must involve. He says that a machine can think if it has "a nervous system, neurons with axons and dendrites, all the rest of it, sufficiently like ours... If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality and all the rest of it using chemical principles different from those human beings use" ([19], p. 422). Thus although Searle allows the possibility, in principle, of artificial intelligence, I interpret his view as being contrary to mine. Whereas I grant intelligence to any digital simulator that can calculate what a human brain would do and direct its body on that basis, he seems to require that the simulator itself be structurally and causally analogous to the human brain.

But Searle's Chinese room argument does not even refute the Turing test, much less establish a particular type of causality. If the operator is asked in English about his Chinese comprehension, he is disposed to say that he understands not a word, whereas a truly bilingual person would respond differently. The philosophical behaviorist can maintain that this difference, suitably refined and qualified, is the behavioral difference between his understanding Chinese and not. It does not involve relative Chinese fluency, but it is a behavioral difference nonetheless, and the behaviorist needs no more.

Searle might respond that he can instruct the operator of the Chinese room to lie about his ignorance of Chinese and thereby pass the behaviorist's test. But the behaviorist maintains that lying itself is to be understood behaviorally. Here, "the operator is lying" may mean in part that under a threat of death, with no matter of principle at stake, and with what he takes to be an infallible lie detector attached to his body, he would confess ignorance of Chinese, whereas a speaker of Chinese would not. The behaviorist may be wrong, but the Chinese room argument does nothing to show he is wrong. At best it begs the question, since the truth of behaviorism is precisely what is at issue.

Despite the failure of the Chinese room argument, I would grant Searle that a purely behavioral analysis of understanding is inadequate, and I agree that it must involve a particular kind of causality. In fact it is difficult to make sense of the counterfactual conditionals whereby dispositional traits are analyzed unless they are grounded in some kind of causal connection. But I can find no justification in Searle's writings for his view that understanding must involve something structurally similar to the flesh-and-blood causality of the human body.
2.4 Agency

A fourth element of personhood is agency. An agent is a being that deliberates about and chooses what to do and can be held responsible for its choice. The lower animals merely behave, whereas a human being acts. It is crucial to understand, however, that in an important sense, action is in the eye of the beholder. True, behavior cannot be reasonably described as action unless it meets certain conditions. It must be intentional in certain ways (the result of deliberation or conscious choice, directed toward accomplishing a goal, etc.), and it must show a certain amount of intelligence. Since deliberation and conscious choice require self-consciousness, the agent must be self-conscious. (Thus it is unlikely that a dog, for instance, could be an agent in this sense.) But there is another element. If I say that you are an agent, I not only vouchsafe that you satisfy these conditions, but I take the further step of choosing to regard you as an agent.

An illustration can help explain this. It is one that, as it happens, was suggested by Searle in another context and ultimately by Ludwig Wittgenstein, although Searle at least would take issue with my analysis of it [18, 21]. Imagine that an anthropologist from another culture observes a baseball game and infers the rules. To improve his understanding of this odd behavior he dons a uniform and goes through the motions of playing a few innings. His behavior is indistinguishable from that of the players, but it makes sense to say he is not really playing baseball. He is only going through the motions. When he swung the bat three times he can say he was not really out, on the ground that he was not a member of a team and was not really part of the game. By the same token, he cannot say that he scored a home run, even though he hit the ball into the bleachers.

My understanding of this situation is that if the anthropologist’s behavior satisfies certain conditions (he follows the rules of baseball), it is optional whether I regard him as really playing baseball. But this much is not optional: once I regard him as playing baseball, then I must regard him as out when he thrice swings the bat. Thus for me to say that someone is “playing baseball” is partly to say that his behavior meets the minimal conditions and partly to commit myself to making certain other baseball judgments about him when his behavior merits them.

I therefore believe that agency is like baseball. If I say that you are an agent, I say in part that you meet the minimal conditions and in part that I commit myself to making certain other judgments, many of them quite serious. In particular, I must concede that, within limits imposed by the situation and your own behavior, you are entitled to the rights of an agent, whatever I take these to be. Thus if my household robot is smart, goal-directed and self-conscious, I can choose whether to regard him as an agent. But if I do so regard him, I cannot conscienably throw him in the trash when I fancy a new model, because I believe it is wrong to do this to an agent. Some may object that this view leads
to moral nihilism, since it seems to imply that I can choose whether to regard another human being as an agent and therefore whether I am to be bound by obligations toward him or her. Indeed, people have on some tragic occasions chosen to regard certain other human beings as less than agents and thereby tried to escape their obligations to slaves or those of another race. But I do not think my view of agency leads me to nihilism. I, like the moral philosopher Alan Gewirth, would argue that when I choose to regard myself as a moral agent with certain rights, consistency requires me to regard other self-conscious human beings (and perhaps robots) likewise.

My analysis of agency is reminiscent of Dennett’s notion of an intentional stance, discussed earlier. On his view, we should not ask whether a thing is “really” an intentional system in his sense, but whether it is useful so to regard it. But whereas Dennett separates the question of whether a thing is an intentional system from whether it has moral responsibilities, I propose that anything granted agency in the sense explicated above is ipso facto granted the moral rights and responsibilities that go with agency.

I think that a computer can satisfy the minimal conditions for agency. If I have been right up to now, it can meet the intentionality, intelligence and self-consciousness requirements. Deliberation, however, requires more explanation. Deliberation consists essentially of the application of a “moral syllogism,” which is an argument that has the form,

\[
\begin{align*}
X, \ Y \text{ and } Z \text{ are my ends.} \\
A \text{ is the only means to } X, \ Y \text{ and } Z. \\
\text{Therefore, I will do } A.
\end{align*}
\]

It may seem unreasonable to require that I choose the only means to my ends. But typically one of my ends is to use the most convenient means that comes to mind to further my other ends. Barring ties, only one means satisfies this requirement. The point is that, at least in the paradigmatic case, an agent has a reason for his action. If I choose arbitrarily among several means, I do not have a reason for my choice.

To resume the thread, I can regard a being as an agent only if I can reasonably interpret its behavior in various circumstances—its response to questions, and so on—as indicating that it sometimes reasons from ends to means in a moral syllogism. This is again a behavioral/causal requirement. The agent should be able to adduce the reasons for an action (the ends to which they are means) when challenged, and it should tend to behave similarly in other situations in which the same reasoning would apply. This is the behavioral requirement. Also there should be a causal connection between the deliberative process and the behavior it is supposed to result in—the causal requirement. Therefore, a computer can meet the deliberation requirement.

Deliberation is clearly similar to ratiocination in general. If it is reasonable to interpret some of a computer’s behavior as indicating that it moves from premises to conclusion, and if this movement from premises to conclusions is
the cause of the intelligent behavior that indicates its existence, then I have the option of regarding the computer as undergoing a reasoning process. So, if understanding necessarily involves reasoning (and I think it does), then intelligence is in the eye of the beholder in the way that agency is.

I conclude that a household robot can meet the minimal conditions for agency, and if so, that I have at least the option to regard it as an agent. Whether I can consistently regard myself but not a robot as an agent is a further issue I will not consider here.

2.5 Feelings

Perhaps the most common objection to investing a computer with personhood is that computers cannot have “feelings.” But this notion of feelings is a complex one. It is related to but different from that of emotions. Emotions involve an excited physiological state, dispositions to behave in certain ways, and certain “feelings” – that is, ways that one “feels inside.”

Probably a machine made of nuts and bolts cannot have emotions in the full sense of the word, since emotions seem definitionally connected to the state of a human body. But this is no obstacle for a digital computer implanted in such a body. Also an embodied computer with intentional states and reflective consciousness can exhibit the behavioral dispositions that go with feelings. The notion of “feeling a certain way inside” seems to be at the core of misgivings about computers having feelings. It involves reflective consciousness, but again reflective consciousness does not seem to exhaust the matter. Human beings not only know that their minds are in a certain state, but they participate directly in its sensuous quality. It has been remarked that whereas thought once seemed the exclusive province of human beings (and gods), now we find it much easier to accept that a computer can think than that it can have “raw feels” – that it can “feel” what hot and cold are like, or know what it “feels like” to be in an angry state. So I think that at the heart of the “feelings” objection is again the idea that reflective consciousness involves something more than a consciousness monitoring of its own state, something more mysterious.

2.6 A Mysterious Property

Richard Rorty and some other philosophers influenced by Wittgenstein have insisted that the notion that something yet unanalyzed lies at the root of self-consciousness is based on confusion [17]. It is based on the curious Western notion that one’s self is a homunculus inside one’s head (or wherever it might be; see [5] for an amusing essay on the location of consciousness). Descartes, who concocted mind-matter dualism to solve the homunculus problem, is allegedly the historical root of this confusion, and prior to him mankind was blissfully free of it. Plato and Aristotle were never troubled by “raw feels,” called “qualia” in much of the recent literature (e.g., [13]). The problem of the mysterious
something underneath self-consciousness is a pseudo-problem symptomatic of the egocentric modern worldview.

The difficulty is, one can let the Rortian deconstructionists analyze away this pseudo-problem as often as they like and still walk away unsatisfied. It helps somewhat to regard the mysterious property as in the eye of the beholder in much the way that agency and ratiocination are. When another human being displays the five traits of personhood I have already discussed, I think it is fair to say that we decide to regard him as having the same sort of feeling self or core of consciousness that we have.

In fact it is an old epistemological question how we can know that other bodies house minds like our own, rather than simply behaving as though they do (v. [16]). We may now have an answer: perhaps there is nothing to know here at all, once we establish that the behavior is sufficiently complex. We simply choose to assume that a consciousness like ourselves haunts another body, in the same way we may choose to regard someone who swings a bat as playing baseball.

This view allows us to explain how self-consciousness can genuinely be something above and beyond a physical mechanism, without postulating a mysterious property. For on this view, no physical property alone is tantamount to self-consciousness; I can ascribe an object any physical property I like and, without contradiction, deny it self-consciousness, provided I do so in a consistent manner from object to object. This helps, but I cannot be sure it exhausts the matter. One knows that when his brain dies, it will not (immediately) cease to exist; there will only be a modification the chemical reactions taking place in it. Yet he has the unshakable intuition that something will be totally annihilated (or will resume its existence elsewhere). This intuition gave rise to Cartesian dualism but predates it by centuries, since it is clearly evident in the Pythagoreans’ notion of an immortal psyche or soul. Pythagoras identified the soul with the rational faculty, whereas we tend to identify it with a “feeling” faculty, but the dualistic intuition is present in either case.

My response to this impasse is to acknowledge that there may be a property of consciousness I do not understand but to argue that it is no harder to attribute it to a digital computer than to a lump of grey matter. For argument I revert to a thought experiment that is often used to defend the possible personhood of computers. Suppose I undergo brain surgery, and the surgeon discovers, not a brain, but a digital computer inside my skull. I would not on that account infer that I lack “raw feels” or any of the other advantages of personhood. It must therefore be possible for a digital computer to have or generate raw feels.

This argument is not nearly as potent as it seems. Otherwise I could have adduced it in the beginning to show that a computer can be a person and dispensed with the rest of my arguments. A thought experiment that would accomplish this involves much more than imagining that a surgeon finds what appear to be wires and circuit boards inside my skull. He must also find that these wires and circuit boards comprise a digital computer that is responsible for
my behavior. To imagine that he does this, I must concoct, in detail, a specific
digital mechanism that is adequate to generate the traits of personhood, and
imagine that the surgeon finds this mechanism. Until I do this the thought
experiment is incomplete, and no one has ever been able to do it.

However, once I have established by other means (the arguments above)
that a computer can support the less mysterious traits of personhood, perhaps
such a thought experiment can show that the computer can have "raw feels."
For I can now claim to imagine that in my skull is a digital computer that is
responsible for everything about me but (perhaps) raw feels, but that no one
clearly understands how it works (since the precise operation of the computer
no longer need be part of my thought experiment). Now I need to imagine that
this imperfectly understood digital computer is also responsible for my raw feels.

How to do this? Granted, it is difficult to say how in principle one can show that
a computer, or any mechanism, causes raw feels. But this does not stop us from
making this connection with the brain. When neurologists probe and stimulate
certain parts of an imperfectly understood brain, and thereby stimulate raw feels
in the patient, we do not hesitate to say that the brain is somehow responsible for
the raw feels. I suggest that we imagine an electronic engineer who generates raw
feels in me by probing certain parts of my ill-understood cranial circuit boards.
This should count equally well as showing that the computer is responsible for
my raw feels.

It may be helpful to remark here that an old tradition in the philosophy of
mind sees mind and matter as the same thing viewed from different aspects.
Hinduism reflects this view in the notions of brahman and atman, and in the
West Baruch Spinoza is a well-known exponent of such a "dual aspect" theory.
On this view raw feels are not something caused by certain neural firings, but
they are the very same occurrences as the neural firings, seen "from the inside."
Thus there is no need to account for raw feels in addition to their physical
correlates. Dual aspect theories have never made this idea clear, but they could
eventually dispel the mystery surrounding raw feels.

3 Can a Computer Be Me?

Granting that a computer can be a person, can it be the same person as now
occupies a human body? This sort of question is hardly new. In the eighteenth
century John Locke recalled the story of the cobbler who swaps bodies with
a prince, in order to raise just this issue of how to identify persons over time
[15]. Ordinarily we say that a person I meet today is the same person as one I
met yesterday if and only if the body, including the brain, is the same. If this
bodily criterion is correct, then swapping bodies is obviously impossible, not to
mention such notions as my transplantation into a computer, transmigration of
souls, and disembodied life after death. But to appeal to this criterion without
argument is to beg the question, since many people believe quite strongly that
a person is separable from a particular body, and the prince and cobbler story, while fanciful, does seem at least to be conceivable.

### 3.1 The Memorial Criterion of Identity

If a person is separable from a body, however, it is difficult to say what personal identity is. David Hume in fact argued with merciless logic that no criterion of personal identity makes sense, and he drew the rash conclusion that no person can survive the passage of time [10]. Then as well as now the main alternative to the bodily criterion is one based on memory, and defending it was Locke's purpose in citing the prince and cobbler story. On his view, the prince today and the pauper yesterday are the same person because the prince remembers being the pauper, and the pauper remembers being the prince. By extension, a computer is me if it can remember occupying my body before the transplant. We may not want to say that memory is a necessary condition of identity, since I am identical with the person who did any number of things I cannot and will never recall, but it is at least a sufficient condition.

The memorial criterion of identity is very plausible on first hearing but quickly withers under examination. Joseph Butler, a contemporary of Hume's, pointed out that remembering something always entails the possibility of being wrong [3]. For instance, if the prince remembers mending a shoe in a certain humble shop yesterday, then it must make sense for me to say he was really dancing at a royal ball. In ordinary circumstances, this means if we traced the whereabouts of the prince's body back through time, we would find it at the royal ball yesterday. But since this bodily criterion is inappropriate in the fairy story, we must fall back on the memorial criterion. Thus when I identify the prince at the ball with the prince who remembers mending shoes, I can only mean that the prince who remembers mending shoes remembers being at the ball (at the same time), which is nonsense.

### 3.2 A Causal Criterion of Identity

Butler infers from this argument that the memorial criterion presupposes a bodily criterion and therefore cannot explain what we mean by identity when a person leaves his body. For the only way he sees to check the accuracy of a person's memory is to check whether his body did the remembered deed. While it seems clear that the memorial criterion does presuppose another criterion and so fails to help explain identity, it is not obvious that the other criterion need be a straightforward bodily one. A bodily criterion rejects body swapping outright, while our intuition that it somehow makes sense has yet to be explained away. On the other hand, if we try to understand identity totally without reference to one's body or bodies, then we become lost in a metaphysical netherworld.

One compromise solution is, rather than to trace a body through time, to trace a causal chain through one or more bodies. By a causal chain I mean a
possibly complex series of events in which each event influences one or more later ones in the series (perhaps as well as other events outside the chain). Certainly all the events in the universe can be viewed as a single causal chain, and it is somewhat arbitrary how one divides these events into subchains. But one can often identify a series of events in which the causal interconnections are dense enough to make it useful, for purposes of explanation, to isolate the series as a causal chain. One such chain is that which resides in a single human body throughout its life span. Certainly the body influences and is influenced by events outside it, but medical and other sciences find it quite useful to identify a single causal nexus that evolves inside the body. In particular, the causal connection between an experience and its recollection often lies wholly within the body.

A causal chain that conveys personal identity would be one that, above all others, is useful in explaining how a person came to be the way he is. The causal chain that resides in one’s body is obviously such a chain and therefore accounts for traditional cases of personal identity. But similar chains can account for a wide array of pathological fairy-tale and science-fiction cases. One scenario is that in which the state of every molecule in a person’s body is recorded on computer tape, while he is asleep, whereupon his body is annihilated. Minutes later a remarkable synthesizer reconstructs a body exactly like the old one from raw materials, in which every molecule has the state dictated by the computer tape. The person soon awakens and reports an experience exactly like that of an ordinary night’s slumber. We have a compelling intuition that a single person survives this process, and the necessary causal chain is indeed present, although a single body is not.

The sort of causal chain I have described would be necessary for identity but not sufficient, since one may want to impose at least two additional necessary conditions. The first is a similarity condition. There may be limits to how drastic a change a person can undergo and remain the same person, even when there is a tight causal link. This is clearly true of inanimate objects, such as furniture. If a grand piano gradually transmogrifies into an easy chair, we are nonetheless inclined to say that the piano is not the same piece of furniture as the easy chair. Second, one may want to require that a transformed person have at least some memories from before the transformation, or, as in the case of an amnesiac sufferer, at least be able to recover some memories after appropriate therapy. The memorial criterion may therefore reappear in a weaker form, but it no longer helps explain identity, but merely limits the sorts of causal connections establishing identity to those that preserve some memories.

The case of the prince and the cobbler is complex for two reasons. To begin with, it may run afoul of the similarity condition, since it is unclear whether one can really conceive waking up in a body totally different (except for the cerebral cortex?) from that in which he fell asleep. But supposing this is conceivable, there is no obvious causal mechanism linking the prince with his cobbler past. I think, however, that while there are no computer tapes involved, we tend to
assume that some supernatural agency is responsible for the transformation. Perhaps a magic fairy brings it about and thereby supplies the causal link that explains it, albeit in a supernatural mode.

On the account of identity I have proposed, I can clearly be identical with the computer that replaces my body. The technicians need only supply a mechanism that informs the computer of my bodily state before the transplant. This mechanism constitutes the necessary causal link. Perhaps, as in the case of the prince and cobbler, it is difficult to conceive myself occupying a very different body, particularly one made of nuts and bolts, but we can presume that I am transplanted into a new body just like the old one but for a digital brain. We can also presume that my memory traces are installed in the circuitry.

3.3 Dispensing with Personal Identity

The causal theory of personal identity I have suggested explains a number of our intuitions and allows for my transplantation into a computer, but it has one serious drawback. It is not a theory of identity. Another science-fiction example will make this clear. Recall the fellow whose body, during sleep, is annihilated and reconstructed. Let us suppose that two copies of his body are reconstructed rather than one. If there are grounds to identify the sleeper with the occupant of one new body, then those same grounds would seem to require his identification with that of a second new body. But the occupants of the two new bodies are clearly distinct persons; one could die while the other survives. Thus our theory identifies a single person with two distinct persons. This is an absurdity, since identity is transitive; if A and B are identical with C, they are identical with each other. It follows that our theory is not a theory of identity.

There are at least three possible responses to this quandary. One is simply to reject the causal theory of identity and search for another. But I am unaware of another plausible theory, and if I were, it might founder no less on such body-duplication examples. A second tack, popular with some philosophers of mind, is to refuse to take these pathological examples seriously. Precisely because we have never faced such dilemmas, they do not help us to understand our core notion of personal identity, for which the paradigm cases are persons who occupy a single body throughout their lives. But the whole point of my exercise here is to decide what constitutes personal identity in just such cases. If this requires revision of our core notion of identity, then our core notion needs revision. In fact, as I said at the outset, computer technology can do us a favor by nudging us into a more adequate view of ourselves.

A third response is more promising. It is to say that Hume was right after all—that no person, to the extent that he is regarded as an entity in his own right, can survive the passage of time. Since persons do in some sense endure through time, this is another way of saying that persons are not entities in their own right at all. When we speak of persons we are really speaking of streams of causes and effects, which are not truly entities because, like rivers on the
landscape, there are no clear conditions for identifying and distinguishing them. As Willard Quine says, "no entity without identity." If two rivers merge, which of the tributaries is to be identified as the "same river" as their confluence, if either? There is no point in debating the matter; it is enough to say that both are tributaries. Similarly, if a person splits into "two," there is no point in debating which, if either, of the offspring is the same person as the original. It is enough to say that both flow out of the original.

It may appear that I have ignored a fourth option, namely that personal identity is identity but not transitive like ordinary identity. But I cannot distinguish this from the position I have just taken. When I say that personal identity is not truly identity, I simply mean that it is not transitive, from which it follows that persons are not entities in the sense that they can be identified and distinguished.

This simple view has profound implications. It implies that when a question of personal identity arises, there is no answer, because identity is not a concept that applies to persons. There is a large and complex causal link between me and the person sitting at this keyboard a minute ago, whereas there is a relatively tenuous link between me and the announcer heard on my car radio some hours ago. I cannot go on to say that I am the same person as the guy at the keyboard but distinct from the announcer, unless I mean by this only that the causal coupling is much tighter in the former case than the latter. The difference in the two cases is purely one of degree. Furthermore there are no borderline cases. Rather than agonize over whether two persons are identical, we merely describe the causal connection between them, and that is the end of the matter; there is no further decision to be made.

If identity does not apply to persons then in particular it does not apply to agents, and this may seem quite upsetting. Our whole system of justice seems to rely on our ability to hold a person responsible for past deeds and thus on our ability to identify him with the doer of the deeds. To begin with, if a theory of justice is based on a misunderstanding of personal identity, then so much the worse for it. But this aside, it is unclear that our system of justice, or something very much like it, really needs the concept of personal identity. I obviously cannot go into detail here, but note that we are already less inclined to hold persons responsible for deeds done in their distant past than for deeds done yesterday (except perhaps for the most heinous crimes). This indicates a recognition of the more remote causal link between oneself and his distant past; personal identity cannot explain this intuition, since identity does not come in degrees. Also a major reason for holding persons responsible for past deeds is to deter future crimes, and deterrence depends on the strength of the causal link, not on identity. Under ordinary circumstances, the strongest causal links are of course those provided by a common body, and it is appropriate to hold a person accountable just for actions committed by his body.

Another radical consequence of rejecting personal identity is vaguely comforting. We often speak of a dead person "living on" in his children or students
or others he influenced. On the theory I propose this is literally true in a quite
unsentimental sense. For it says that the connection that binds one with what
he calls his later self is not different in kind, but only in degree, from the con-
nection that binds one with his offspring or protegés. One can quite literally
achieve a degree of immortality by influencing his survivors, since one’s influence
on his own body’s future, were it to keep living, would differ only in degree.

Similar reasoning causes immortality of the body to lose its luster. It is more
than a bit specious to refer to the present occupant of an immortal body as “the
same person” as its occupant a thousand years hence, which we would be obliged
to do if identity applied to persons. It is more plausible to speak of a tight causal
link between me today and “me” tomorrow and a very tenuous one between me
today and “me” a thousand years hence. I am more closely “identified” with
“other” people I influence today than with “myself” a thousand years in the
future. The death of a friend today is accordingly a much more grievous loss
to me than my “own” death a thousand years down the road. It is true that
even if I were literally identical to my future self I might “discount” my future
value when comparing it with the present value of others, due simply to the
time lapse. But the theory I propose would justify “discounting” even if the
time lapse made no difference.

If my view of personal identity (or nonidentity) is correct, then how can I
explain the fact that people universally subscribe to a totally different view?
Clearly one reason we tend to treat persons as discrete entities is that, in the
biological world we know, the associated causal streams are tied to relatively
independent bodies and are therefore relatively distinguishable. If our bodies
were linked by a complex network of neural fibers, the causal streams would
intertwine, and the illusion of distinguishable persons would disappear.

My view may also seem to contradict our deeply-felt intuition that when a person dies, there is total annihilation of some entity, and not just
the dissipation of some causal chain. But this sense of finality has several
components that should be analyzed separately. One obvious component is our
often overwhelming fear of death. But as Hume might have pointed out, this
is a biological trait of our bodies that need have nothing to do with personal
identity. It is conceivable that a person be biologically disposed to dread the
death of his children as much as his own, and some persons may in fact do so.

Another component is the sense that death destroys something precious and
irreplaceable. This squares with my causal theory, since the causal chain that
constitutes a person is no ordinary causal chain. Although it differs only in
degree from interpersonal causal connections, the difference of degree is great,
due to the astounding complexity and organization of the causal nexus one
carries around in his body. Therefore its dissipation is indeed the destruction
of something marvelous and irreplaceable.

A third component may be a brute phenomenological datum, the seemingly
unmistakable sense that one and the same entity, me, endures through my expe-
riences. Does not one sense his identity through time directly and incorrigibly,
as one senses a pain? But this “datum” may not be so unmistakable as it seems. Although I seem to have a clear sense that the person typing at this keyboard a moment ago is identical with me, on reflection I have no such clear sense of the schoolboy who bore my name thirty years ago. I seem almost to need to make a mental effort to identify myself with him, and the identification seems almost an abstract concept that society finds convenient.

Perhaps my clear perception of identity over a moment but not over years is really a result of how I sense the passage of time. The phenomenologist Edmund Husserl, for instance, argued that one experiences the passage of a few moments, as when one hears a short tune, quite differently than the passage of hours [12]. The former seems to involves a special faculty (“retention”) that unconsciously assembles the events into a simultaneous whole so that the shape of the melody can be sensed directly, whereas the latter involves ordinary recollection, in which events must be consciously assembled before their shape becomes evident. A perception of identity over time by way of recollection cannot be direct and incorrigible, since memories can be mistaken. Careful introspection seems to reveal that any direct perception of identity is really a consequence of short-term “retention” and does not extend to longer periods. Perhaps the fear of death has led us to suppose that we have a clear introspective sense of long-term identity without actually examining our own experience. There is therefore no good reason to suppose that what I sense over the short term is personal identity, since if it were (rather than simply the phenomenon of retention) it would be unclear why I should not sense identity over longer periods.

A final component of our sense of finality may have to do with the possible existence of the mysterious feeler of raw feels I discussed earlier. Even if the physical manifestations of persons can be treated like rivers and need no criteria of identity, the identification of the underlying metaphysical selves may still pose a problem. Let me grant that my physicalist account omits some metaphysical element beyond my understanding. But now that the other components of death’s finality have been explained without recourse to personal identity, what evidence remains that the metaphysical substrate must take the form of discrete entities corresponding to persons? Our souls could be parts of a single world-soul that intermingle perhaps in much the way that causal chains intermingle in the material realm.

4 Conclusions

I have argued that personhood is actually a cluster of several distinguishable traits that need not all occur together – intentionality, reflective consciousness, intelligence, agency, and feelings. I have found no reason to deny a digital computer the conceivable possession of any of these traits. Relatively simple machines can have intentionality, and reflective consciousness in a computer is simply its ability to add to its activity, on demand, a monitor of that activity.
Humanlike intelligence in a digital computer eludes us but can in principle be achieved.

Agency (and intelligence, to the extent that it involves ratiocination) are more complex in that we can choose whether or not to regard a creature as having them, provided it meets certain minimum requirements. A sufficiently complex robot does seem to meet the requirements. We could reasonably regard it as deliberating about its behavior and therefore as an agent, provided its behavior is at least sometimes caused by a process interpretable as formulating a moral syllogism. (We may in fact be inconsistent to regard ourselves but not robots as agents.) Feelings, as well as a mysterious property that may enable a being to have “raw feels,” are likewise traits that we to some extent choose whether to ascribe to a creature, and it seems no less reasonable to attribute them to a computer than to grey matter.

As computers become more complex they are likely to exhibit some of the traits of personhood convincingly and others less so. Rather than agonize over whether to call them persons simpliciter, we should simply acknowledge that they have some of the earmarks of personhood but not all. Our relations with machines should be governed by which traits they possess. We can converse with a machine that has only intentionality, intelligence and reflective consciousness, ask it about itself, and discuss its past and future, but we owe it no obligations and cannot relate to it emotionally as we would another human being, except in some pathological way. If we choose to regard the machine as an agent, we should praise it for virtue, blame it for vice, and recognize an obligation for its well-being. It makes no sense, however, to fall in love with it. If in addition it has emotions and, from all appearances, feelings, there is no reason to shrink from regarding it as one of ourselves.

The question remains whether it makes sense conceptually, apart from technical barriers, to achieve immortality by transplanting oneself from one computer to another. I have argued that if one can accept a quasi-Humean view of personal identity, radically different from the ordinary view, then immortality of this sort does make sense. On this view persons are interpreted to be causal chains in the material world. Since the identity and distinctness of causal chains over time is not well defined, they and therefore persons are not entities in their own right. Hume was right to say that no person survives the passage of time. There is no difference in kind between my connectedness with my future “self” and with “other” persons I influence; the only difference is one of degree—the tightness of the causal link. Once this is granted, it makes sense to say that I can be connected with the future occupant of a digital computer in the same sense in which I am connected with the future occupant of my body, and in this case there need be no difference of degree.

This view of personal “identity” is, again, at odds with our ordinary notions, but for me at least it has the advantage of being the only plausible theory I can concoct at the moment. It is curious, however, that a purely materialist theory would be one to admit the possibility of immortality.
References


