Faith and Philosophy: Journal of the Society of Christian Philosophers

Volume 26 | Issue 5 Article 2

12-1-2009

Artificial Intelligence and Personal Identity

Alexander R. Pruss

ARTIFICIAL INTELLIGENCE AND PERSONAL IDENTITY

Alexander R. Pruss

Persons have objective, not socially defined, identity conditions. I shall argue that robots do not, unless they have souls. Hence, robots without souls are not persons. And by parallel reasoning, neither are we persons if we do not have souls.

Introduction

Could a computer or robot be a *person*: a being that, at least under normal circumstances, is a thinker and an agent, responsible for its thoughts and actions?

Let me clarify the question a little. Research in Artificial Intelligence is progressing. It is, admittedly, progressing more slowly than was once expected, but it is moving ahead. It seems not unlikely that one day robots will be able to function in ways that will look just like the functioning of a person. If so, we will be able to have what seem to be conversations with them, and get what seem to be the sorts of answers that a person would give. I am not asking in this paper whether *this much* is possible. I am asking, rather, whether even if we achieved all this, this would be genuine personhood. This question is, I think, tightly bound up with the question whether the robots would be thinking and acting rationally, or whether it would merely *appear* that they are.[1]

The mere *appearance* of thinking and acting rationally is not so hard to get. If I can get a good estimate of the sorts of questions someone might ask a computer, I can program the computer to give pre-programmed answers to those questions. You ask what the computer's name is, and the computer utters the sound: "I am HAL." If I am good enough at this, I can make people think that the computer is really communicating, is really telling them what it is thinking. But just pre-programming a lot of answers to questions does not give the computer an understanding of the questions and answers.

Even a somewhat more sophisticated program need not have understanding. I can right now put into Google the question "What is seven plus five?" and immediately Google comes back with "seven plus five =

[1]Cf. John Searle, "Minds, Brains and Programs," *Behavioral and Brain Sciences* 3 (1980), pp. 417–457.

FAITH AND PHILOSOPHY
Vol. 26 No. 5 Special Issue 2009

twelve." The engineers at Google programmed their servers to recognize arithmetical expressions in English, and then to compute answers. But there is no *understanding* on the part of the computer at least as yet—there is just the non-conscious processing of patterns of characters, with no responsibility or thought.

While it is an interesting *technological* question whether we can make a robot that externally seems to behave just like a person, passes psychological tests and so on, the *philosophical* question is whether we could make such a robot that would *be* a person, and not just *seem* to be a person. Or would such a robot just be non-consciously processing patterns, with no responsibility or thought?

I will approach this question through considerations of personal identity. We can ask various questions about the identity of persons, for instance across times. We can ask whether *this* person here and now is the same as *that* being there and then. There was a certain five-year-old who grew up to be me. Was that five-year-old the same person as I am? I think the answer is clearly "Yes." On the other hand, there was a once a person named Queen Victoria, and now there is a corpse at Frogmore, Windsor. Is Queen Victoria the same person as the corpse? No: for the corpse is not a person at all. (So strictly speaking it is not correct to say that Victoria is buried at Frogmore—only her corpse is.) If you have a pair of Siamese twins, is that one person or two? Surely two—each is a person, and each is a distinct person. These questions are easy to answer.

But if I erased your memory completely, and then tortured the resulting amnesiac, would *you* be the one feeling the pain, or would the pain be someone else's? Unlike the questions about the five-year-old in my past, Queen Victoria and Siamese twins, the answers to *this* question are controversial. Some maintain that your identity is guaranteed by the persistence of your body or maybe your brain, and so the amnesiac would be *you*, and hence the pain would be yours.[2] Some hold that your identity is guaranteed by a stream of memories,[3] in which case it wouldn't be *you* who is tortured after the amnesia, and so while it makes sense to *feel sorry* for the person being tortured, there is no need to have a first person *fear* of the pain, because you will no longer exist after the amnesia. And, finally, some hold that your identity is guaranteed by the presence of something over and beyond the body, a *soul*.[4] If so, then answering the question is difficult, for we would have to know whether the soul remains in the body after amnesia.

We may not know what the answers to personal identity questions are. But it is, I submit, a part of the concept of a person that there *exist* answers

---

[2]The classic piece here is Bernard Williams, "The Self and the Future," *Philosophical Review* 79 (1970), pp. 161–180.

[3]The classic account is Sydney Shoemaker, "Persons and Their Past," *American Philosophical Quarterly* 7 (1970), pp. 269–285.

[4]The best defense is Richard Swinburne, *The Evolution of the Soul*, revised edition (Oxford: Oxford University Press, 1997).

to such questions, though perhaps the answers are beyond our knowledge. It is crucial that either it is true or it is false that *x*, who is a person, is the same being as *y*. For instance, the notion of *responsibility* presupposes identity: you are only responsible in the relevant first-person way for having thought or done something if you are the being who thought or did it.

All my arguments will have the form of a *reductio ad absurdum*: I assume that computers or robots can be persons, and then I argue that some considerations connected with personal identity probably lead to absurdity. Consequently, I conclude that the assumption that computers or robots can be persons is false. The basic theme is that I will ask a question related to personal identity, assuming computers or robots can be persons, and argue that the question probably cannot be answered in the case of robotic persons. Since the question is one that *would* have an answer if robots were genuine persons, it follows that robots are probably *not* genuine persons.

I will formulate some of my arguments in terms of computers rather than robots. A robot is, after all, just a computer plus a more sophisticated input/output system than the ones that typical computers have. Or we may think of a robot being related to a computer in the way that a normal person is related to a brain in a vat. Since one could be a person despite being a brain in a vat, it is very plausible that if robots can be persons, computers can, as well.

Finally, I shall stipulate that computers and robots do not have souls—that they are made of nothing but the hardware, software and data. Richard Swinburne has suggested to me that the hypothesis that a soul would come into existence when a sufficiently sophisticated robot was made does not appear that improbable to him. But I simply will not count that as a robot. I shall be talking of *mere* robots and computers.

1. Argument A: Power Switches

Suppose Robby is a robotic person, and I turn Robby off. Is Robby still in existence when turned off? In other words, is Robby the same person as the turned-off robot? I will argue that the answer, given the presupposition that Robby is a person, is both *yes* and *no*. Since the answer *cannot* be both *yes* and *no*, we must reject the presupposition in the question that it is possible to have a robotic person.

So, first let me argue that Robby is still in existence when turned off.

Argument A1a: Robby is an artifact like a vacuum cleaner or car. These artifacts certainly continue to exist when turned off, and hence so does Robby.

Note: This argument, plausible as it may seem, will not convince everyone. One might think that the essential component of Robby is the *software*, and in this way he is different from artifacts like vacuum cleaners or cars (though as vacuum cleaners and cars get more sophisticated, this objection becomes weaker). In particular, Argument A1a will not convince those who have a subtler view of the connection between Robby the robot-

ic person and the physical artifact than just saying Robby *is* the physical artifact. One might instead say that Robby is *constituted* by the artifact, and is somehow to be identified with the functioning of the software running on the artifact. If one turns off Robby, the software no longer runs, and so Robby ceases to exist.

Argument A1b: To be a person, one does not have to be *actually* thinking and acting. Otherwise, we would cease to be persons while in deep sleep. All one needs is a capacity—or at least a well-developed capacity[5]—for thinking and acting. But Robby when he's turned off surely has a well-developed capacity for thinking and acting.[6] He just can't exercise this capacity until one turns him on. Being turned off is like being asleep, rather than non-existence.

Argument A1c: Some people think it is impossible to exist, then not exist, and then to exist once again. If such temporally gappy existence is impossible, then Robby is still in existence when turned off, since it is clear that he exists before being turned off and after he is once again turned on.

Note: This argument will be rejected by some on the basis of the case where your watch is disassembled and reassembled, and is the same watch again after the reassembly, since if that is possible, then temporally gappy existence should be possible for Robby as well. However, there may be a difference between the two cases. Perhaps the identity of artifacts like watches is simply a matter of social convention, which allows watches to have strange survival conditions, while the identity of persons had better not be simply a matter of social convention, since the existence of social conventions presupposes the existence of persons.[7]

Argument A1d: To conserve energy, a computer may turn itself off until a timer goes off or sensor is activated, which will then turn the computer back on at a later time when it is needed. ("Standby mode" is something like that.) Suppose Robby turns himself off for one second in this way, with a timer turning him back on a second later. It seems very plausible to say that Robby is in existence during that second. But now suppose that instead of an internal timer, there is an alarm clock attached to Robby's outside with rubber bands, in such a way that when Robby's switch is flipped to off, the alarm clock is set to go off in an hour, and when the alarm clock goes off in an hour, it flips Robby's switch to on. Surely the fact that the timer is physically on the outside rather than inside doesn't matter.

---

[5]The "well-developed" qualifier is required by Mary Anne Warren, "On the Moral and Legal Status of Abortion," *The Monist* 57 (1973), pp. 43–61.

[6]One might worry that this is not a natural capacity because robots lack natural capacities. But if robots lack natural capacities and personhood requires natural capacities, then there cannot be robotic persons anyway, and my arguments are moot.

[7]I have heard Robert Koons give arguments along these lines.

Perhaps, though it matters whether the timer is external or internal in this way. Maybe an external timer is not really a part of the robot. But it would be strange if something attached with rubber bands wasn't a part of the robot, whereas something welded to it would be. So this way out doesn't work. The externality of the timer doesn't matter, and so Robby continues to exist even while turned off by the alarm clock, just as he would even were he turned off by an internal timer.

But suppose that while Robby is turned off, I detach the alarm clock, and bring it back half an hour later. Have I really made Robby cease to exist for that half hour by detaching the alarm clock? And anyway, why does it matter *what* turns Robby back on, whether it is the alarm clock by itself, or *me* plus the alarm clock (as when I take the clock away and bring it back), or just me by myself? It seems that either in all of these cases Robby is existent when turned off, or in none of them is he existent when turned off. Since in the case where the alarm clock is securely attached, Robby is still in existence while turned off, he is in existence while turned off in all of these cases. Hence, he is in existence when I turn him off with the intention of turning him back on.

While there are objections available to some of these arguments, I think there is overall a very strong case for a *yes* answer to the question whether Robby exists while turned off.

But there is also a very strong case for a *no* answer.

Argument A2a: Robby, on our assumptions, is a person without a soul. Only a person with a soul can exist while not *alive* (and even that possibility is controversial). But Robby is not alive when he is turned off: life requires *active* functioning. If Robby, then, is like we would be if we were persons without souls, then Robby does not exist when he is turned off.

Note: Whether one accepts this argument depends on what one thinks about frozen humans.[8]

Argument A2b: Let's think a bit about what it could physically mean to "turn off" Robby. One way to do this would be to press a switch that disconnects the electrical connection between the battery and the rest of the robot. But it seems to me that the answer to the question whether Robby would continue to exist when turned off should not depend on exactly *how* turning him off works. My son has a battery-powered toy car where the switch works by physically pushing the battery away from its connector. We could imagine that Robby's off switch works by pushing the battery out of him. But the battery is, surely, a crucial part of Robby. Without such a crucial part, Robby does not exist, just as we would not exist without a heart unless we have souls. So if turning him off works by pushing out a battery, he doesn't exist when turned off.

[8]See Peter van Inwagen, *Material Beings* (Ithaca, NY: Cornell University Press, 1990), pp. 146ff.

> But it shouldn't matter exactly *how* we turn him off, and hence no matter how we turn him off, he doesn't exist when turned off.

Maybe, though, Robby's memories survive while he is turned off, because they are recorded on a kind of memory that does not need electricity. Could we say that Robby, then, survives because of his memories while turned off, even though he is lacking a crucial part? No. Our memories are presumably recorded in our brains, but we're dead as soon as we stop functioning, even though, quite likely, for a few minutes—or maybe longer—our memories could still be recovered from traces in our brains, if only we had the technology for it (there are science fiction stories about this sort of thing). That a record of memories exists does not mean that life continues, and similarly that a record of memories exists does not mean that Robby continues to exist.

But suppose you're not convinced by this. Suppose you think that as long a record of memories exists, then Robby continues to exist. Well, then, imagine a different thought experiment: All of Robby's memories are printed out on a very long piece of paper in small type. Then the electronic copy of the memories is destroyed while Robby is turned off. When we turn him back on, the memories are typed in again from the piece of paper, maybe by a human typist, maybe by a trained monkey, or maybe by an electronic scanner that reads the piece of paper. Then, a record of Robby's memories does continue to exist while he is turned off and his battery is removed. In *this* case, Robby shouldn't count exist when his battery is removed, despite the printed record of his memories existing. But there should be no metaphysical difference between a *printed* record and a record on a *disk*, so neither does he exist in the case where the memories are held on a disk.

Maybe, though, you think removing the battery is not enough to make Robby cease to exist. Well, remove more parts, one by one. Eventually, Robby doesn't exist—there is just a desk full of parts. But at which point does he cease to exist? There is no sharp line once the battery is removed. The removal of the battery *is* a fairly sharp line—once the battery is gone, Robby doesn't function, doesn't in any sense live. But after the battery is removed, there are no more such sharp lines. Since the cessation of existence is a sharp line, removal of the battery is what makes Robby cease to exist.

So, we have a strong case for a *yes* and a strong case for a *no*. We could decide on a *yes*. But then the *no* arguments would be against us. Or we could decide on a *no*. But then the *yes* arguments would be against us. The right decision is to reject the supposition that the question was based on, namely the supposition that there can be a robotic person.

I suspect that in the case of a robot, the answer to the question whether the robot is still existing when turned off is given by social convention not by the objective fact. In the case of a person, the question whether the person is still existing at a given time is a matter of objective fact (though

perhaps in some cases, such as those of brain death, this fact is hard to determine). Therefore, robots are not persons.

2. How Many Electronic Persons Here?

When we're dealing with persons, the question "How many different persons are here?" should make sense for appropriate senses of "here" ("here" need not be physical—it could indicate a context instead).

There are two basic ways to try to answer this question for electronic persons. The first is to correlate persons with pieces of computing hardware. Some of the arguments in the previous section were based on that kind of a view. On a hardware-based view, if there are three intelligent computers, then we have three persons, even if one of these computers is multitasking several intelligent programs, each communicating with a different user through its own window. A second way would be to focus on software, and to correlate persons not with pieces of computing hardware, but with streams of computation. Thus, a single computer could be "inhabited" by a dozen intelligent persons, each constituted by a separately running process. I shall argue that neither approach succeeds in giving satisfactory answer to the "How many" question.

2.1. The Hardware Approach

On the hardware approach, a difficult question is how to count pieces of computing hardware. For instance, I am writing this paper on a laptop with a dual core processor. Is this one piece of computing hardware or two? A dual core processor is, basically, two processors in a single package. While one of the processors may be processing my typing, the other may be checking for viruses. Yet to the ordinary user the laptop behaves like a single computer, and Microsoft treats it as such (you only need to buy one Windows license for it).

To some degree, our two-hemisphere brain may function like a dual core processor. But surely the brain is a single piece of computation machinery—we are not actually two persons (leaving aside cases of split brain patients). So if we're counting electronic persons by counting pieces of machinery, we should count my laptop as a single piece of computing machinery.

But how does one, then, distinguish between a single laptop with two processors and two laptops with one processor each? The physical condition that to have two laptops they would be in separate plastic cases is clearly not right. If I took the insides of the two laptops and placed them in a single box, they would still be *two* laptops—even if I glued them together (Siamese twins are two persons). Nor can I say that I have one laptop whenever there is only one keyboard and only one screen. After all, it's easy to hook up a second screen and a second keyboard.

Such physical criteria are, surely, beside the point. If anything makes the laptop a single computer, it is that it's functioning *as a whole*. There may be two processors, but they are working together, in a well-coordinated way.

But this coordination is a matter of software, not hardware. I could, after all, connect two computers wirelessly to the Internet, and, running the right software on both, operate them as a *cluster* that functions as a single, larger computer. So it is not some kind of purely physical contiguity that makes my laptop be a single piece of computing machinery.

If we employ hardware-based criteria for counting electronic persons, we will at best get things wrong. But in fact, not only will we get things wrong, we will get no answers in general. The question "How many computers are here?" is not answerable in general. Do we count two laptops glued together as a single, more complex computer, or do we count them as two individual computers? The same problem comes up for other kinds of artifacts. Suppose I take three chairs and tie them together, side-by-side. Do I have a single, new piece of furniture—a bench with twelve legs—or do I still have three pieces of furniture?

In fact, I think it objectively does not matter what we say. The answer is surely just a matter of social convention. There is no objective fact to be discovered by metaphysical investigation of the chairs or laptops. It is simply up to us to decide whether we count a set of three chairs tied together as one piece of furniture or three (or maybe even four). Of course, the answer may matter, say, for legal purposes. If I have an insurance policy that covers only one computer, and then the two laptops glued together are destroyed, then there will be a *legal* question whether I can claim the total loss or only half of it. But the answer is one to be determined by the courts, or by linguistic convention, rather than by mind-independent facts.

But this is not so for *persons*. The question of how many persons there are, whether here there are two persons or one person, has an *objective* answer, one the courts can be wrong about, though sometimes we may not be able to find that objective answer. And what I said about computers applies just as much to robots. Therefore, if the hardware approach to counting electronic persons is right, then robots can't be persons. For if they were persons, there would be *objective* answers to the question of how many of them there are. But there are no objective answers available for such questions about electronic persons, at least on the hardware approach.

2.2. The Software Approach

The software approach is more promising. On this view, if I run one intelligent program on one processor core and another on another core, I have two electronic persons, but likewise if I run them both on *one* core, I still have two electronic persons. If, on the other hand, I run a single intelligent program in parallel fashion on several processors, indeed on several computers, each computer doing its part of the total computation task, then I have only one person, spanning multiple computers. This software approach is promising as it embodies the conviction of those who believe in the possibility of strong Artificial Intelligence that it is not the physical substrate that matters for personhood, but what matters is the *computation* that is going on.

However, the software approach to counting persons also runs into difficulties. One of these is that if it is applied *to us*, it may mean that a human being with multiple personalities is literally more than one person—there is more than one stream of computation going on there. Maybe, though, you do not think that is absurd, or maybe you think the approach cannot be applied to us, but only to electronic persons.

A second difficulty is that it is sometimes difficult to count streams of computation. Suppose that I want to compute the positions of the planets in 10,000 years. But I want to be *really* sure of the result. So I take eleven computers with the same hardware specifications. One of these computers, then, sends to each of the other ten the task to compute the positions of the planets in 10,000 years using the laws of physics and the present positions, giving each of the ten the same program to run. The coordinating computer then continually checks to make sure that the memory state of each of the ten computers is, at each time, exactly the same. As soon as the memory state of one of the ten computers deviates from the others, the coordinating computer modifies the deviant to match the others. (If more than one deviates at the same time, the coordinating computer goes crazy and explodes everything, maybe, or maybe conforms the minority to the majority if possible.)

Should I see this situation as consisting of *ten* streams of computation of the positions of the planets? Or maybe *eleven* (ten individual ones plus the whole consisting of the coordinator plus the ten subsidiaries)? But the ten streams of computation are highly interdependent. The coordinating computer ensures that as soon as any deviation occurs, the coordinating computer cancels out the deviation. If it is interdependence that defines a single stream of computation, then I think the right thing to say in this case is that there is only one stream of computation.

Now let us imagine a version of this hypothesis for *intelligent* programs (it doesn't require intelligence to compute the positions of the planets). We have ten intelligent computers, that is ten computers running an intelligent program. And we have an eleventh computer—this one isn't intelligent since its task is simple—which gives them all the same input, and monitors their functioning. As soon as any of the ten were to diverge from the others, the eleventh would push the divergent computer back to the same state as the others. But let us suppose that *in fact* there is no divergence. Here, I think, we can make a good case for the hypothesis that we have ten intelligent programs, each running on one computer, as well as for the hypothesis that we have *one* intelligent program, running on a system consisting of eleven computers.

First let me argue that we have ten intelligent programs. Let us suppose that in fact none of the ten computers diverges from the others—no malfunctions occur.[9] The fact that they are always thinking the same thing

[9]And if there are any indeterministic events going on in the computation, they happen to go the same way for all of them.

does not make them be one person. After all, it is quite possible, though unlikely, for two people to be always thinking the same thing—imagine you and an identical twin on a planet just like ours, where everything is just like here. And why should the existence of a coordinating computer make them all be one person, if the coordinating computer does not actually *do* anything to them?—it just *watches* for deviations, but if there are no deviations, as on our present hypothesis there are not, it does nothing.

Suppose you are one of three identical triplets who always think the same thoughts. If Big Brother watches you and your two triplets, and will force the thoughts of each to conform to the thoughts of the majority whenever there is divergence (and blow up all of you if there is no majority), that does not make the triplets into a single individual, at least if they all always happen to agree with one another without having to be forced to (this is, of course, like Frankfurt cases in discussions of free will).[10]

On the other hand, it seems clear that *overall* we have a single computational system, made up of eleven sub-parts. This single computational system is running an intelligent program with the additional feature that the intelligent program is more resistant to hardware failure (since deviations get canceled out).

Perhaps, then, the right way to look at this is to say that we have *eleven* persons. One is the *system* as a whole (running on an aggregate of eleven computers), and then there are the ten *component* persons (the coordinating computer does not count as a person, because it isn't intelligent—it just automatically cancels out deviations in functioning). So we have one person who has ten *more* persons as parts. That isn't of itself absurd, perhaps. (Or is it? Maybe it would imply that each of us is a person who has two persons—one per brain hemisphere—as parts?)

But what does seem absurd is that just by inserting a coordinator who in fact does nothing (because nothing is to be done), one has created a new person. Suppose that you and your two identical triplets are going along thinking the same thoughts. And then a non-intelligent computer is put into place, whose job is to ensure that your and your triplets' thoughts never diverge. As soon as the thoughts diverge, the computer will make them converge again. But in fact, your thoughts do *not* ever diverge. So *in fact* the computer doesn't affect anything. Yet, a fourth person is thereby immediately created if we accept the view above that there are eleven persons in the computer case. This seems absurd.

Furthermore, if inserting the unintelligent coordinator creates a new person, then destroying the coordinator kills the person, and this is, of course, morally wrong without strong reason. But there seems to be nothing wrong with destroying the coordinator.

Moreover, I perhaps don't even need a coordinating computer to get the problematic result. Let's take ten computers, and run different copies

[10]Harry G. Frankfurt, "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66 (1969), pp. 829–839.

of the same intelligent program on each, and give each the same input. We can *think of* the ten computers as together running a single program in a more reliable way simply by treating them as a unit—once the output comes, we simply accept the majority output (typically the outputs will be the same). It is up to us, after all, how to interpret the outputs. But surely how *we think of* a bunch of persons does not affect how many persons there are.

Thus, on the software view, it is also true that the question of how many programs are running is answered by our subjective decision as to how we want to consider a situation: do we want to think of it as a system running one program, or several systems running several copies of a program, or maybe in some other way? Again, there does not appear to be an objective answer to the question of how many intelligent running programs there are. And so running programs are not persons, since there is an objective answer to "how many" questions in regard to persons.

Still, the software view seems to work better than the hardware view. So for the next identity question, I will only consider the software view.

3. Identity Over Time

I can take a piece of software running on one computer, record all the computer's memory to disk, erase the memory, put the disk in another computer, restore the data from disk to the memory of that computer, and continue running the software there. On a software view of the nature of electronic persons, if the software constitutes a person, the person should survive such transfer—after all, the stream of computation continues.

One question to ask is whether the alleged electronic person exists when the software is not running and all we have is a memory record on the disk. We asked this question when we considered whether Robby existed while turned off. This question leads to problems for the proponent of the possibility of electronic persons. For it is absurd to suppose that an inert disk, or even the information on it, could be a person. On the other hand, how does the case where the data is temporarily on a disk actually differ that significantly from what happens during the ordinary execution of a modern computer program? After all, when a program is running, sometimes the computer just keeps it stored inertly in memory while doing something else—one of the ways modern computers multitask is by switching tasks. So if the disk or the data on it is not a person, neither would electronic persons exist except precisely when the computer is doing something with the programs that constitute them. But if so, then by parallel, it seems that should my mental activity pause for a moment, I would not exist then, which appears false.

But I want to focus on a different, and very standard in the theory of personal identity, set of questions. Call the electronic person I had before the memory recording "Robby." Suppose that I actually make two copies of the memory record, and then simultaneously put them into two computers. Where does Robby go? Does Robby inhabit *both* computers

now? But that seems absurd. After all, the two computers might now be given different inputs, and thus might be doing different things at the same time. Does the program as running on one of the computers need to be afraid if the program as running on the other computer is facing pain (after all, if computers could be intelligent, then probably they could suffer pain)? Surely not.

Or does Robby inhabit only one of the computers? But which one? After all, the data was put into both simultaneously. Again, there does not seem to be an answer to this question.

Or, perhaps, Robby inhabits *no* computer after his data has been sent to two computers. But then it follows that if Robby's data is restored on only one computer, Robby continues to exist, but if it is restored on two, Robby ceases to exist. This, too, seems very strange. Let's suppose that when Robby's data is put on a disk, a copy of the disk is carried in a spaceship going to a far away star. Why should it affect the question whether Robby exists on earth what is done with the copy of the disk on the spaceship? Yet, if on the spaceship the data is restored at the same time as the one on earth, then Robby ceases to exist on this hypothesis, but if only the disk that is on earth is restored, then Robby continues to exist on earth. And if both disks are restored, but one before the other, then, plausibly, Robby resumes existence in the place where the earlier restoration happened (after all, if the later restoration doesn't happen, then Robby will exist where the earlier one happened, and where Robby is shouldn't depend on what disks are restored in the future). But then the answer to the question whether Robby comes to be on the spaceship or on earth may well depend on the reference frame—for on relativistic grounds it may well depend on the reference frame which restoration happened first.[11]

So it seems that there is no good answer to the question whether and where Robby continues to exist in this thought experiment. But in any real situation, when dealing with persons, there has to be an objective fact whether the person continues to exist or not.

4. What About Human Persons?

But there is a serious weakness in all of the above arguments. In these arguments, I suggested that certain questions about personal identity have no answers in the case of electronic persons. But there are exactly parallel questions that we can ask about human persons.[12]

Parallel to the question whether Robby continues to exist when he is turned off, we can ask whether people continue to exist while in a coma.

---

[11]One might think that Robby only survives if the two restorations are time-like separated (i.e., one is in the light-cone of the other), and if they are space-like separated (neither is in the light-cone of the other), Robby perishes. But then whether Robby survives on earth depends on what happens at a space-like separated point of space. This suggests that there is something very much like faster-than-light causation. For it seems that by restoring the disk, the technicians on the space-ship manage to prevent Robby from coming back into existence on earth!

[12]E.g., Derek Parfit, *Reasons and Persons* (Oxford: Clarendon, 1987).

Parallel to the question of how many electronic persons there are, we can ask how many humans there are—think of Siamese twins, for instance, as a way of making this question problematic. And parallel to the question whether Robby's data is restored on two computers, we can imagine thought experiments where my brain is split in half, and the two halves are put in different bodies—where, if anywhere, would I be then?[13]

I suggested that in the case of electronic persons these questions cannot be answered. But then how can they be answered in the case of humans? Here I need to note an assumption in my previous arguments. I was assuming that there was nothing to electronic persons but the hardware and the software (including data), that there was no further metaphysical reality beyond these. Thus if we were going to come up with an answer whether Robby continues to exist while asleep, this answer would have to depend only on the hardware and the software—there is nothing else there. And the hardware and the software fail to give an answer to the question.

But exactly the same point can be made about humans. If all there is to us is a bunch of molecules and a bunch of data encoded in these molecules, then questions of personal identity do not always have objective answers. If these questions are to have objective answers, there must be *more* to us than just molecules and data. What could be this "more" that makes answers possible? I think it has the traditional name "soul."[14]

But even supposing a soul, we do not know what the answer *is* for some of these questions. If my brain is split in half, where do I go? Well, if I have a soul, then I can say that the question is ill-defined. "What if" questions only make sense if sufficient information is specified. Suppose you are in a crowded room, and someone asks you: "What if you moved one third of the human beings from this room to another room, which room would I be in?" In fact, this has no answer because sufficient information is not specified. One needs to know whether the questioner would be among the one-third moved or the two-thirds remaining to answer the question. Likewise, the question: "If my brain is split in half, where do I go?" has no answer unless one further specifies which half of the brain my soul goes with or maybe that my soul goes with none. I will go where my soul

[13]It may be worth noting that the thought experiments are fanciful in the human case and not at all fanciful in the computer case—it's easy to duplicate computer memory. I do not know what exactly to make of this disanalogy. One might want to say that the concept of a "person" does not apply in fanciful situations (see Richard M. Gale, "On Some Pernicious Thought Experiments," in *Thought Experiments in Science and Philosophy* ed. T. Horowitz and J. Massey [Savage, MD: Rowman and Littlefield, 1991], pp. 297–304). If one says this, then one will be able to resist the analogy between the human and electronic case. In the case of humans, we do not expect answers for duplication cases, since these are fanciful cases. In the case of electronic persons, we would expect answers for duplication cases, since for robots and computers the cases are not fanciful. But I want to resist this move of relativizing the concept of "person" to non-outlandish cases.

[14]One might also opt for an unanalyzable non-supervenient fact theory of identity, but if the robot is just hardware and software/data, it's not clear where that additional fact would come from. Divine fiat seems the only explanation. It's also worth noting that I intend what I say about souls here to be neutral between substance-dualist and hylomorphic views.

will go. This is basically Richard Swinburne's personal identity argument for the existence of a soul: only if we have souls can there be answers to certain questions.[15]

However, in the case of an electronic person, when we describe what happens to the hardware and the software, we are in an appropriate sense describing *everything* relevant, and so we should be able to get answers. Assuming electronic persons don't have anything beyond the hardware and the software, the question *is* sufficiently specified once we've given the facts about what happens to the hardware and the software, such as in my example where Robby's data is recorded in a disk and restored on two computers. And yet, even though the question is sufficiently specified, as there is nothing further to specify, there is still no answer.

If this is right, then computers and robots cannot constitute persons unless, somehow, there is more to them than hardware and software, namely unless computers and robots will have souls. And by parallel, we cannot be persons unless we have souls.[16]

Baylor University

[15]See Swinburne, *The Evolution of the Soul*.

[16]I would like to extend my gratitude to the participants and organizers of the *Science and Human Nature* conference.