The Metaphysics of Artificial Intelligence

Eric T. Olson

In Mihretu Guta, ed., Consciousness and the Ontology of Properties, Routledge 2019: 67-84.

Abstract Debates about the possibility of artificial intelligence have focused on the question of whether programming a computer in the right way could produce genuine thought. But for there to be thought is for there to be thinking beings. What sort of being might be made intelligent by programming a computer? Would it be the computer itself--a physical object? Some part of the computer? The program running on the computer? Or something else? There has been almost no discussion of this question. Yet if artificial intelligence is possible, it must have an answer. A satisfying account is elusive.

1. What is artificial intelligence?

Many people believe in the possibility of artificial intelligence: they think it is possible to produce intelligence just by programming a computer in the right way. Or maybe programming alone is not enough. Maybe the computer's internal states would also need to relate in the right way to its surroundings, giving it something like perception. And maybe it would need to cause perceptible changes that would manifest its intelligence, giving it something like action. Artificial intelligence might only be possible in a sort of robot. Let us suppose that any such further requirements are satisfied.

By 'computer' I mean an electronic digital computer. 'Artificial intelligence' normally means electronic intelligence, not biological intelligence created artificially, as with Dr Frankenstein's monster. By 'intelligence' I mean mental phenomena generally: belief, desire, emotion, awareness, and so on--thought and consciousness for short. It may be that some mental phenomena can be produced by programming computers and others cannot: belief and desire, perhaps, but not emotion or conscious awareness. This would be an important fact about those phenomena. But set it aside. My interest is in the view that programming computers can produce any mental phenomena at all.

The term 'artificial intelligence' has other meanings. Most commonly it means what we might call intelligent behavior in computers: getting machines to do things that human beings can do only by exercising mental powers (recognizing junk emails, playing chess, driving cars, and so on). It can also mean programming computers to simulate or model mental phenomena. The artificial intelligence of this chapter is the production of genuine thought or consciousness in computers. The claim that this is possible is called strong artificial intelligence.

2. Artificial thought, artificial thinkers

It is usually assumed that in order to know whether artificial intelligence is possible, we need only think about the nature of mental phenomena in themselves. We need to establish whether the nature of mental phenomena requires them to have a special sort of substrate. We know that thought and consciousness can occur in biological organisms. Is there anything in their nature that would rule out their occurrence in computers? Those who believe that mental phenomena are entirely characterized by their causal roles--so-called functionalists--argue that the substrate is unimportant as long as it does the right sort of thing. Thinking is like timekeeping: anything can keep time as long as it undergoes changes at regular intervals and keeps track of how many of them have elapsed. It doesn't matter what it's made of, or whether the process involves gears and wheels, sand, water, or electronics. Others say that mental phenomena are not like timekeeping and require more than activity with the right causal role. These philosophers are often sceptical about artificial intelligence.

But showing that nothing in the nature of mental phenomena restricts it to a biological substrate would not suffice to establish the possibility of artificial intelligence. It requires something further. Not only must there be artificial thought or consciousness, but something must be the subject of that thought or consciousness. For there to be thought or consciousness is for there to be something that thinks or is conscious--just as for there to be life is for there to be living things, and for there to be movement is for something to move. For there to be artificial intelligence, then, there must be an artificially intelligent being: a thing that is intelligent because of what a computer does.

So there are two different questions concerning the possibility of artificial intelligence. One is whether anything in the nature of thought itself prevents it from occurring in computers. We might call this the question of artificial thought. The other is whether anything could be an artificial thinker. We might call this the question of artificial thinkers. The second question has to do with the sort of entity an artificial thinker would be. What properties would it have, in addition to its mental properties? Would it be a material thing? If so, what matter would make it up? If not, what sort of immaterial thing could it be? What might it be made of if not matter?

This question must have an answer. An artificially intelligent being would have to have some nature or other (just as natural thinkers do). It would have to be either composed entirely of matter or not. If it is, then some particular atoms--certain portions of the computer hardware and robotic machinery, perhaps--would have to compose it at each time. Which ones would that be? The answer need not necessarily be precise. There might be atoms that are neither definitely parts of an artificial thinker nor definitely not parts of it, just as an atom may be neither definitely a part nor definitely not a part of you or me. But there would have to be some true story of which atoms (if any) were parts of it, which were not, and which had an

indeterminate status. And if artificial thinkers were not composed of matter, there would have to be some account of their immaterial nature.

In order to assess the possibility of artificial intelligence, then, we need to know something about the nature of thought and consciousness in themselves, but also something about the nature of thinking, conscious beings. And just as there are grounds for doubt about whether anything a computer could do would count as real thought or consciousness, there may be doubt about whether there is an acceptable account of the metaphysical nature of artificial thinkers.

Nearly all discussion of the possibility of artificial intelligence has been devoted to the question of artificial thought. The question of artificial thinkers, by contrast, is almost entirely unexplored. Philosophers ask whether artificial intelligence is possible by asking whether anything computers could do would count as thought, but they say almost nothing about what sort of things these "computers" would be--or whether an artificial thinker would have to be a computer at all, as opposed to, say, a computer program.

Discussions of artificial intelligence are often phrased in ways that obscure the question of artificial thinkers. A typical statement of the question of artificial thought asks "whether intelligence can be embodied only in systems whose basic architecture is brainlike..., or whether it can be implemented in some other manner" (Boden 1990b: 1). This diverts us from the question of articial thinkers in two ways. First, it speaks of artificial "systems": a vague term that (like 'substrate' and 'medium') can refer to almost anything. To ask whether "the system" in a certain case would be intelligent is perhaps to ask whether <u>some</u> rather comprehensive object would be intelligent, but without further specification it does not identify any particular entity. It discourages us from asking more precise questions about this object. Second, it asks whether intelligence is "embodied" or "implemented" in something. This is not to ask whether that thing actually <u>is</u> intelligent, but merely whether it stands in some intimate but unspecified relation to the property of intelligence. Knowing that intelligence could be "embodied in" a computer would not tell us what the intelligent thing was. Such formulations are deliberately chosen in order to state the question of artificial thought without raising the question of artificial thinkers. There is nothing wrong with that: we can't ask all the questions at once. But it can prevent us from seeing that the question of artificial thinkers even exists.

3. Must there be a thinker?

I said that if artificial intelligence is possible, there must be something that could be the subject of this intelligence, as for there to be thought is for there to be a thinking being. Most friends of artificial intelligence appear to accept this claim[1] But should they? Could it be that while life requires living

[1] At any rate this seems to follow from the fact that they ask whether there could be artificial intelligence by asking whether computers could think (see

things and movement needs a mover, thought can take place even though nothing thinks? If so, artificial intelligence would not require artificial thinkers, and its defenders need not worry about their nature. The question of artificial thinkers would not arise.

My reason for supposing that artificial thought would have a subject is the same as my reason for supposing that human thought has a subject.[2] I believe that these thoughts--the ones I am now expressing in writing--are the thoughts of someone. They are states of someone or activities that someone is engaged in. This chapter has an author. And if any being is thinking these thoughts or writing this chapter, I am. If these thoughts had no thinker and this chapter had no author, it would follow that I do not exist. There would be no such thing as me, just as there is no such thing as the tooth fairy. And if I don't exist, you don't either. Yet I believe that I do exist and that you do too.

But if it were possible for thought to take place in a computer without there being anyone or anything doing that thinking, it would be possible for thought to take place in a human being without there being anyone or anything doing that thinking. The thought that you and I are engaged in right now would not require anyone to think it. Things could appear exactly as they do even if we did not exist. In that case they would not appear that way to anyone; but if there could be thought without a thinker, then there could be appearances without any being to whom things so appear. And such a possibility would undermine any grounds we might have for believing in our own existence.

It is not necessarily mad to doubt the existence of thinking beings, ourselves included, just as it need not be mad to doubt the existence of the physical world (see van Inwagen 1990: 115-123, Olson 2007: 180-210). But without a powerful reason, such doubt would be unwarranted. In the absence of such a reason, we ought to accept that we exist and that artificial thought, like our own, would have a thinker.

4. The computer-hardware view

Suppose, then, that it is possible to produce genuine thought by programming a computer in the right way and in the right circumstances. What sort of thing would be the subject of this thought? What would artificial intelligence make intelligent?

Given that artificial thought would consist of states or activities of a computer, the most obvious answer is that computers themselves would be

---

the next section).

[2]I labor this point only because it is common for philosophers (and students) discussing Descartes' Second Meditation to question it, even if they would never otherwise question their own existence. The reason for the disparity is not that Descartes does anything to undermine this claim, but that he argues for it. The surest way to get a philosopher to question something is to give an explicit argument for it.

intelligent. This, as we noted earlier, is what those discussing artificial intelligence most commonly say. They ask whether artificial intelligence is possible by asking whether computers could think (Turing 1950, Putnam 1964, Newell and Simon 1976, Searle 1980: 417, Haugeland 1985: 2, 76, 106; Copeland 1993: 33, Russell and Norvig 2010: 1020). They don't argue for this claim, or consider alternatives; they simply take it for granted. Nor do they try to say exactly what objects computers are. But they appear to be assuming that what becomes intelligent when the computer is appropriately programmed is a piece of computer hardware: a physical object made of metal and plastic and silicon, nowadays manufactured in East Asia.

This is analogous to an attractive account of the subjects of natural intelligence: that we are biological organisms. That's what we see in the mirror. We appear to be material things, and we don't appear to be any larger or smaller than organisms. We should expect artificial thinkers to stand to computers as we ourselves stand to human animals. If we <u>are</u> those animals, then artificial thinkers must be computers.

Call the view that artificial thinkers are computers, or parts of them, the <u>computer-hardware view</u>. I think this is the best answer to the question of artificial thinkers. But it is not an entirely easy view to accept. The most obvious objection is that it conflicts with widely held views about identity over time--that is, about the persistence of thinking beings.

If artificial thinkers are pieces of computer hardware, then programming a computer for intelligence makes a previously unintelligent being intelligent. And when the program stops running, the intelligent being loses its intelligence and becomes once again an ordinary piece of computer hardware. Running an artificial-intelligence program on my desktop computer just once for an hour would make that machine (or some part of it) intelligent for an hour. For the rest of its career it would be no more intelligent than my desk is. That can easily sound wrong. It is tempting to suppose that programming the computer for intelligence does not merely give a previously unintelligent being the property of intelligence, but brings an intelligent being into existence. And shutting down the program and erasing the data destroys the intelligent being, rather than merely depriving it of intelligence. Yet running and then stopping the program does not create or destroy any piece of computer hardware. It follows from the tempting thought, then, that artificial thinkers would have different histories from the computers they run on: the computer would exist before and after the thinker did. More generally, artificial thinkers would differ from computers in their persistence conditions: computers can survive being switched off and having their data erased, but artificial thinkers cannot. In that case artificial thinkers could not be computers.

A second tempting thought is that an artificial thinker could move from one piece of computer hardware to another. This would not require the guts of one computer to be physically removed and wired into another. An

electronic data transfer would suffice. Otherwise transferring all the data from one computer to another would result in the first computer's losing all its mental properties--memories, beliefs, preferences, cognitive capacities, and so on--and the second computer's acquiring them. It would give the second computer the false belief that it was the first computer, and that it did all the things the first computer did before the transfer. It is tempting to suppose that the being resulting from the transfer would be the thinker it thought it was and remembered being, and that it was transferred from the first computer to the second along with the data.

But this too is incompatible with the hardware view. You cannot move a piece of hardware--a physical object made of metal and plastic--by an electronic data transfer. According to the tempting thought, artificial thinkers would have a property that no piece of hardware has, namely being moveable by electronic data transfer. The computer-hardware view implies that no sort of electronic transfer could move an artificial thinker from one computer to another.

Those familiar with debates about human thinkers will know that the view that we are biological organisms--"animalism"--is criticized for analogous reasons (see Olson 2007: 39-44). Suppose your brain were transplanted into my head (or that the psychological information in your brain were transferred to mine), so that the resulting person was psychologically continuous with you and not with me. He would have your pre-operative memories, beliefs, preferences, and other mental properties rather than mine. Most people say that he would be you: the operation would not give me a new brain, but would give you a new body. But the operation would not give an organism a new body. It would not pare down an organism to the size of a brain, move it to a new location, and then supply it with a new complement of extracerebral parts. It would simply move an organ from one organism to another, just as a liver transplant would. It follows that human thinkers have a property that no organism has, namely being being moveable from one organism to another by brain transplant (or by "brain-state transfer"--see Shoemaker 1984: 108-111). If so, we human thinkers cannot be organisms. And for analogous reasons, artificial thinkers cannot be computers.

The general point is that both animalism and the hardware view conflict with the claim that the persistence of a thinking being consists in some sort of psychological continuity: if a thinking being $\underline{x}$ exists now, then something $\underline{y}$ existing at another time is $\underline{x}$ just if $\underline{y}$ is in some way psychologically continuous, then, with $\underline{x}$ as it is now. (Set aside complications to do with "branching", where on thinking being is psychologically continuous with two.) In other words, some condition involving psychological continuity is both necessary and sufficient for a thinking being to continue existing. But no condition involving psychological continuity is either necessary or sufficient for either a human organism or a piece of computer hardware to continue existing.

Friends of artificial intelligence could of course give up the psychological-continuity view. They could accept that programming a computer for intelligence cannot create an intelligent being, but can at most make a previously unintelligent being intelligent, and that erasing the relevant data would not destroy any intelligent being but merely render it unintelligent. (So an intelligent being could survive complete psychological discontinuity.) They could deny that an artifical thinker could ever move from one piece of hardware to another, no matter how much psychological continuity there may be. And they could say that the result of putting your brain into my head would give me a new brain rather than giving you a new body. They could deny that any sort of psychological continuity is either necessary or sufficient for an intelligent being to continue existing. I am not saying that this would be a mistake. I myself reject the psychological-continuity view (because I belielve that we are organisms). But it would be important news if the possibility of artificial intelligence ruled out the dominant view about personal identity over time.

I will return to the hardware view in §9. First I will consider some alternatives. (Firm friends of the hardware view may want to skip ahead.)

5. The constitution view

It may be possible to accommodate the tempting claims about the persistence of artificial thinkers by saying that such thinkers would not be computers themselves, but rather material things constituted by computers (Pollock 1989: 31-46, Baker 2000: 109). Each artificial thinker is made of precisely the same matter as the computer whose programming makes it intelligent. The two objects are physically identical. But they have different histories and modal properties.

The idea is that the artificial thinker stands to the computer as a clay statue stands to the lump of clay from which it is fashioned. It is often said that the lump and the statue do not differ physically while the statue exists, but the lump exists before the statue does. Kneading the lump into the shape of Descartes (say) does not make the lump into a statue. It does not merely change the lump by giving it a new shape. Rather, it creates a statue that did not exist before. And squashing the statue would not change its shape and deprive it of the property of being a statue, but would destroy it. The statue would cease altogether to exist. Yet the lump would endure. And if we replaced an arm with a new one made of different clay, the statue would come to share its matter with a different lump. Even if the statue and the lump did not differ historically, and coincided throughout their existence, they would differ modally--in what could happen to them.

In much the same way, the idea goes, when a human foetus or infant acquires the properties of consciousness and intelligence, it does not merely change psychologically, but comes to share its matter with a person who did not exist before. When it loses those properties, the person ceases to exist, though the organism may survive. And a person whose brain is transplanted

would share its matter with different organisms at different times.

Just so, when a computer is programmed for artificial intelligence, it does not itself become intelligent; rather, the process creates a new thing, made of the same matter, that thinks. Shutting down the program and erasing the relevant data would destroy the thinker, though the computer sharing its matter would endure. And the thinker could be share its matter with different computers at different times by an electronic data transfer. Call this the constitution view.

But few friends of artificial intelligence will be happy with the constitution view. This is because it implies that a thing's physical properties, surroundings, and history do not suffice to fix its intrinsic mental properties. It violates this "weak supervenience" principle:

> Necessarily, if things have the same physical properties, spatial surroundings, and historical properties (including historical surroundings) at a certain time, then they have the same mental properties at that time.

Let the God of the philosophers create a computer programmed for intelligence ex nihilo, and annihilate it after the program has run for a year. Suppose it has all the right relations to its surroundings. On the constitution view, the computer would share its matter with a thinking being at every moment during its existence--a being that would cease to exist if the relevant data were erased. But the computer itself would have no mental properties. (Otherwise programming it for intelligence would result in two thinkers, the computer itself and also the thing constituted by it. No one wants to say that.) Yet the computer and the thinker it constituted would have the same physical properties, surroundings, and history. If there could be artificial consciousness, the computer would be a "zombie" in the philosophical sense: an unconscious being physically identical to a conscious being, with the same behavior.

In fact the constitution view appears to imply that there actually are zombies (Olson 2018). If artificial thinkers are material things constituted by computers, then you and I are material things constituted by organisms (or by lumps of flesh). But almost no one who believes this takes those organisms (or lumps) to be mentally just like ourselves. Otherwise there would be two thinking beings for every human being: the organism (or lump) and the person it constitutes. How could you ever know which one you were? The usual view is that the organisms (or lumps) constituting human people have no mental properties at all (see e.g. Shoemaker 2008, Johnston 2007: 55). They are physically identical to us, with the same surroundings, yet lack any conscious awareness. There are as many zombies as there are human beings.

Friends of artificial intelligence who deny that there are zombies should reject the constitution view.

6. The program view

Maybe artificial thinkers would not be material things at all--that is, things made of matter, with a size, shape, mass, and chemical composition. Some of those who speak of intelligent computers speak just as readily of intelligent computer programs.[3] Could it be the program running on the computer that would think?

A program is normally defined as a sequence of instructions that a computer can follow. This description has a type-token ambiguity. I am using a program called Mariner Write 3.9.5 to compose this chapter. What sort of thing is Write 3.9.5? On the one hand there is an entity that was created (or at any rate given physical realization) by certain programmers at a certain time, is subject to copyright, and is running on many other computers. This is how the word 'program' is most commonly understood. On the other hand there is (perhaps) a particular copy of that program now running on my computer and nowhere else. The first is a type or universal; the second is a token or particular, a concrete instance of the type.

Consider first the view that artificially intelligent beings are program types. Call this the program view. Imagine that the Mariner Software corporation develops a program for artificial intelligence, Think 1.3, which you can download from their website. Then if you run that program on your computer--the same one that other people run on theirs--it's the program that thinks.

The program view is impossible to take seriously. For one thing, a computer program--a type--exists as soon as a certain sequence of instructions exists. It is created when those instructions are first thought of or written down. It does not come into being when it first runs on a computer. It can exist without ever running. If artificially intelligent beings are sequences of instructions, then they too can exist before running on any computer. In fact they could exist even if there were never any computers. It would follow that artificial intelligence of any level of sophistication--and artificial thinkers--could exist even if there had never been any computers. Few friends of artificial intelligence will accept this.

You might reply that although a program can exist before it runs on a computer, it cannot think or be conscious until it does so. Before it is executed, the sequence of instructions has the mental powers of a stone; then it acquires mental powers like yours and mine. But this only leads to another problem. Suppose running Think 1.3 on your computer makes it intelligent. And suppose it runs simultaneously on mine. Then the thinker in or on my computer could be happy while the thinker in or on your computer is not. Yet on the program view they would be the same thinker: Think 1.3. It would both have and lack the property of being happy.

[3] They often see the views as interchangeable (e.g. Russell and Norvig 2010: 2, 4). But a computer and a computer program (type) are metaphysically as different as two things could ever be.

And universals don't change. Write 3.9.5--the type--doesn't change when I start it up on my machine, or type a sentence and save it, any more than the colour white changes when I spill coffee on a piece of paper. Or at least these things don't change in their intrinsic properties, but only in the way that the number 22 changes by ceasing to be the number of players on the field when someone is sent off. It would be the same for Think 1.3. At most a particular concrete instance of the program can change. But conscious, thinking beings must be able to change intrinsically: in their beliefs, preferences, and perceptual states.[4]

7. The bundle view

Computer-program types--sequences of instructions--don't literally do anything. What does the work--performing the calculations, fetching the web pages, or, in the case of artificial intelligence, thinking--is the thing that follows those instructions. That thing is not the sequence itself. It would seem to be the computer hardware. But we've already considered that view.

It may be that when a program runs on a machine at a particular time, there is a particular instance or token of the program located there and then: an electronic event or process, or a "collection" or "bundle" of such events. (It might consist of all the movements or causal interactions of all the electrons involved in the electronic circuit that executes the program.) And someone might suppose that artificially intelligent beings would be such entities. Call this the bundle view. It would avoid the problems facing the program view: electronic events are not created when the program-type is created, and they can change.

If artificial thinkers would be bundles of electronic states and events, we should expect natural thinkers to be bundles of states and events too, though of course not electronic ones. At any rate it would be surprising to discover that bundles of electronic states and events could think or be conscious but bundles of brain states and events could not. I cannot think of any explanation for this dramatic difference. And if there is a thinking, conscious bundle of events within me, then that is what I am. (I clearly think, and I could hardly be a second thinker in addition to the thinking bundle.)

Only a handful of philosophers have seriously held such a view. Here is one reason why. A bundle of brain events is not a material thing. It may consist of the activities of material things--molecules and brain cells. But those activities themselves, though physical, are not material things. They are not made of matter. They have no mass or surface or electrical conductivity. That was the point of the bundle view: if it said that thinkers were material things, it would imply that artificial thinkers were computers or parts of them, making it a version of the hardware view rather than an alternative to it.

[4]For further objections to the program view, see Olson 2007: 145-49.

So the bundle view implies that thinking beings are not material things, but rather bundles of events that material things are engaged in. This has troubling implications. One is that material things can never think. If any material thing could think, it would be a healthy, mature human organism or brain. In that case you and I should be organisms or brains, and not bundles of events. (No one would suppose that each of us is a second thinker in addition to the organism or brain.) And if natural thinkers were organisms or brains, artificial thinkers would be pieces of computer hardware, contrary to the bundle view. Friends of the bundle view will have to deny that pieces of computer hardware, brains, organisms, or any other material things could ever have mental properties. They will need to explain why this is. This task is made more difficult by the fact that thoughts are states or activities of material things: organisms, brains, or perhaps computers. For thinking to be going on in me is for this organism or brain to engage in mental activity. For there to be pain in me is for this organism or brain to be in a state of pain. How, then, could something be engaged in thinking, or in a state of pain, without thinking or being in pain? What's the difference? It would seem to be human organisms or brains that think and are conscious. But then it would be absurd to suppose that bundles of the thoughts that those organisms or brains are engaged in <u>also</u> think and are conscious.

A second implication is that what <u>does</u> think are states and events or bundles of them. Most of those who hold this view say[5] that thinkers are bundles of <u>mental</u> states and events: we are composed not of atoms, but rather of beliefs and memories and sensations and fears. But saying that what thinks are thoughts themselves, or collections of thoughts, is like saying that what moves are movements or collections of movements. Dances dance and games play. It sounds like the most elementary sort of metaphysical confusion.

Friends of the bundle view may try to avoid these problems by rejecting the distinction between material things and their activities, or between substances and events. <u>Every</u> concrete thing, they might say, is a process or event. Stones are slow process and wildfires are rapid ones, but there is no metaphysical difference between them. The dancer <u>is</u> the dance. Or at most the dancer may have a longer temporal extent, existing both before and after the dance. But during the dance they don't differ: both weigh 140 pounds, are composed of 60% water, and are fond of newts. It would follow that the activities of material things are themselves material things, and that thoughts are not states or activities of nonthinking things. They are not strictly states or activities <u>of</u> anything. It would not be an elementary mistake, but rather a profound truth that movements move and dances dance.

"Process" ontologies are poorly understood, and a systematic treatment of them would be a large project. But I will briefly discuss what might be a version of it: the ontology of temporal parts.

[5] As Hume did--see 1978: 252.

8. The temporal-parts view

We began with the view that artificial thinkers would be pieces of computer hardware, and showed that it conflicts with the dominant view about the persistence of thinking beings. We then considered some alternatives: that artificial thinkers are not computers themselves but rather things "constituted by" computers, that they are computer-program types, and that they are bundles of electronic events. These proposals looked significantly worse than the computer-hardware view. But there is a variant of the hardware view that would avoid some of the objections to do with persistence: artificial thinkers might be temporal parts of computers.

Suppose that all persisting things are composed of arbitrary temporal parts. A temporal part of something is a part of it that takes up "all of that thing" at every time when the part exists. Barry Manilow's nose is a part of him, but not a temporal part, as it doesn't take up all of him while it exists. His adolescence, though, if there is such a thing, would be a temporal part of him. His temporal parts are exactly like him at all times when they exist. They differ from him only by having a shorter temporal extent. If he is a material thing, his temporal parts are too; if he is immaterial, so are his temporal parts. To say that persisting things are composed of arbitrary temporal parts is to say that for any period when a thing exists, there is a temporal part of it existing only then.

Suppose also that composition is unrestricted: for any entities whatever, no matter what their nature or arrangement, there is a larger thing composed of them. (Some things, the xs, compose something y $=_{df}$ each of the xs is a part of y, no two of the xs share a part, and every part of y shares a part with one or more of the xs.) So if there are such things as Plato's nose, Barry Manilow's adolescence, and the Soviet Union, then there is also an object scattered across space and time made up of those three things.

Both these claims--that all persisting things are composed of temporal parts and that composition is unrestricted--are highly controversial. But grant them for the sake of argument. Together they imply that every matter-filled region of spacetime is exactly occupied by a material thing. Quine took this to abolish the distinction between substances and events:

> Physical objects, conceived thus four-dimensionally in space-time, are not to be distinguished from events or, in the concrete sense of the term, processes. (1960: 171)

That's why I called the current view a version of the process ontology (though not all friends of temporal parts agree with Quine on this point, and not all process ontologists believe in temporal parts).

I said in §4 that friends of artificial intelligence are likely to believe that when a computer is programmed for intelligence, an artificially intelligent

being comes into existence, and that it perishes when the program stops running and the relevant data are erased. And they may believe that an artificial thinker could move from one piece of computer hardware to another by an electronic data transfer. Both beliefs imply that the artificial thinkers in question are not pieces of computer hardware. But they are compatible with the temporal-parts view.

It follows from the assumption that persisting things are composed of arbitrary temporal parts that any computer programmed for intelligence has a temporal part that exists just when the program is running. If running the program creates intelligence, the subject of that intelligence might be that temporal part of the computer. And if a data transfer brings it about that the program runs on a second computer and no longer on the first, the temporal-parts ontology entails that there is an object composed of the temporal part of the first computer that exists just when it is programmed for intelligence and the temporal part of the second computer that exists when that program runs on it--an object that "moves" discontinuously from one computer to the other. In that case the artificially intelligent being might be a material thing composed of pre-transfer temporal parts of one computer and post-transfer parts of another.

But the temporal-parts view has other implications about the persistence of artificial thinkers that appear less attractive. Suppose once again that a certain computer runs an intelligence-generating program for a day. On the temporal-parts view, the day-long temporal part of the computer is intelligent: it begins to exist when the program starts running, and ends when the program stops and the relevant data are erased. But it also implies that the computer itself is intelligent for that day. That's because for a thing to have a property at a time, on the temporal-parts ontology, is for the temporal part of it located at that time to have that property without temporal qualification (Quine 1960: 172f., Olson 2006: 745-48). So the computer's having an intelligent temporal part implies that the computer is intelligent.

Or suppose that electronic data transfer moved an intelligent being from one computer to another. Again, the temporal-parts view implies that the first computer is also intelligent, and stays where it is and loses its intelligence when the data are transferred, and that the second computer acquires intelligence when the transfer is complete. If an artificial thinker were to wonder what would happen to him when the relevant data were transferred, he may be uncertain, since there would be two beings asking the question, one of which was going to move to another computer and one of which was going to stay put and lose its mental properties.[6]

So the temporal-parts view implies that only some artificially intelligent beings are created and destroyed when computers are and then cease to be appropriately programmed. Others persist through these changes. And it implies that only some artificially intelligent beings move from one computer

[6]Noonan 2010 proposes a solution to this epistemic problem.

to another when the relevant data are electronically transferred, while others stay put. The difference between the temporal-parts view and the computer-hardware view is less than it may appear.

9. The parts of thinkers

Here is a difficulty facing the temporal-parts view that I have been ignoring up to now. It arises equally on the computer-hardware and constitution views. It is the question of what would determine the spatial boundaries, or the spatial parts, of an artificial thinker. (On the temporal-parts view, the question amounts to what would count as a momentary temporal part or "stage" of an artificial thinker.) Imagine, once again, that the computer on my desk is intelligent. No one would say that all the parts of what we ordinarily call the computer, including the keyboard, mouse, display screen, and power cable would be parts of the artificial thinker. What things would be parts of it, then?

If there could be artificially intelligent beings and they would be material things, this question must have an answer. Some atoms must be parts of a given thinker and others not. Maybe some could be neither definitely parts of it nor definitely not parts. But a material thing must have some boundaries, sharp or fuzzy.

There is a precisely analogous question about naturally intelligent beings. Where do our boundaries lie? Which atoms are now parts of me and which not? (And which, if my boundaries are vague, are neither definitely parts of me nor definitely not parts?) This question too must have an answer--assuming, anyway, that human thinkers are composed of atoms.

We should expect there to be a principle that determines the answer to these questions: an account of why the boundaries of a thinking being lie where they do, or of what makes something a part of a thinker. If my hands are parts of me and my gloves are not, there must be an account of what makes this the case. It could hardly be a "brute fact" about which nothing further can be said. There must be a principle of the form

> Necessarily, if x is a thinking being at t, then y is a part of x at t if and only if...x...y...t....

If natural thinkers are biological organisms, the question of what determines their boundaries is the question of what determines the boundary of an organism. Here is a proposed answer to that question:

> Necessarily, if x is a natural thinking being at t, y is a part of x at t if and only if y is, at t, caught up in x's life,

where a life is a biological event or process that takes in particles from its surroundings, imposes a complex biochemical form on them, and later expels them in a degraded form (van Inwagen 1990: 82-97). An organism's life is

roughly the sum of its physiological, immune, and metabolic activities. My hands are parts of me because they are caught up in my life: they and all their parts are nourished by my bloodstream and participate in my metabolic processes. My gloves are not parts of me because they are not caught up in my life. There are many hard questions about what counts as a life, but this is at least a start.

Obviously nothing like this could apply to artificial thinkers. What would be the corresponding principle for them? If my computer's central processing unit could be a part of an artificial thinker but its keyboard could not, why should this be? The best proposal I can think of is something like this:

> Necessarily, if x is an artificial thinking being at t, then y is a part of x at t if and only if y is directly involved in x's thinking at t.

The keyboard does not seem to be involved in the computer's thought at all. And although its power supply is involved--the computer could not produce thought without it--its involvement seems only indirect, compared to certain parts of the computer's digital circuitry. This suggests that an artificial thinker would be composed entirely of electronic components and the wires connecting them. It would be a thin, spidery thing made of metal and silicon weighing only an ounce or two.

If the parts of an artificial thinker are just those directly involved in its thinking, we should expect the same to hold for natural thinkers. The principle appears to derive from the nature of thinking beings as such--with what it is to be a thinker. It has nothing to do with artificial thinkers in particular.

This would be incompatible with the earlier suggestion that our parts are determined by the extent of our biological lives. But that is not surprising, seeing as that proposal assumed that we are organisms, whereas the current one appears to rule it out. It appears to imply that my hands are not parts of me. Although they may contribute to my mental activities by helping to nourish me and providing me with tactile information, their involvement would seem at best indirect. It would be metaphysically impossible for a thinking being to have hands as parts. The same would presumably go for all my vital organs apart from my brain. Some philosophers have said something much like this, and inferred from it that we are brains (Hudson 2007). But the proposal seems to imply that many parts of my brain are not parts of me: its blood vessels, for instance, seem only indirectly involved in my mental activity. Presumably I should be composed entirely of nerve cells. (And probably not even every part of an active nerve cell would be directly involved in thinking, but only those that transmit signals. The involvement of the nucleus and mitochondria, for instance, would appear to be only indirect.) We too should be thin, spidery things weighing only a few ounces.

I have called the claim that a thinking being must be composed of all and only the things directly involved in its mental activity thinking-subject minimalism (Olson 2007: 87-90). It is troubling in a number of ways. One worry is that it is hard to say what it is for something to be directly involved in a being's thinking, or in any other activity. Which atoms are directly involved in someone's walking--as opposed to only indirectly involved or not involved at all? I doubt whether this question has any principled answer.

More seriously, the proposal is hard to generalize. It could hardly be the case that any being engaged in any activity must be composed of just the things directly involved in that activity. Suppose a being could see only if it were composed of just the atoms directly involved in its seeing, and could remember only if composed of just the atoms directly involved in its remembering. Neurologists tell us that seeing and remembering use different parts of the brain. This suggests that the atoms directly involved in someone's seeing are not the same as those directly involved in her remembering. It would follow that no one could both see someone's face and also remember her name. The thing that sees will be either too big to remember names, by having parts not directly involved in that remembering, or too small, by not including such parts--or both, of course. The seer and the rememberer, composed of different atoms, would have to be distinct. And the same is likely to hold for artificial thinkers: any artificial thinker that sees would be distinct from any that remembers names, as different bits of circuitry will be directly involved in these activities.

In all likelihood, no being would be able to engage in more than one mental activity, because no two such activities will have precisely the same atoms directly involved in them. What appeared to be a single being performing many mental activities involving different parts of the brain would in fact be many different beings--overlapping, perhaps, but distinct--each performing only one such activity. A "general" thinker capable of both seeing and remembering would have to be mereologically simple (lacking parts) and presumably immaterial.

It may be that thinking-subject minimalism can be generalized in a more plausible way, or that the parts of artificial thinkers are determined by a principle fundamentally different from minimalism. Friends of artificial intelligence who believe that artificial thinkers would be material things will need a solution to this problem.

10. The relaxed attitude

Material things are composed of atoms. So for each material thing there must be certain atoms that are parts of it at a given time and certain atoms that are not parts of it then.[7] And if artificial thinkers would be material

[7] On the temporal-parts view, material things are typically composed not of atoms but of temporal parts of atoms. Still, an atom is a part of an object at a time in the sense that the temporal part of the atom located then is a part, without temporal qualification, of the temporal part of the object

things, there must be an answer to the question of which atoms would be parts of them. I said that this must come under some principle, such as thinking-subject minimalism: that the parts of an artificial thinker are just those things that are directly involved in its thinking. But that principle looks unacceptable, and it's hard to think of a better one. (The problem does not arise for natural thinkers, at least if they are organisms: what things are parts of them has nothing to do with direct involvement in thinking.) This looks like a serious worry for the possibility of artificial intelligence.

I can imagine someone taking a relaxed attitude towards this problem. Suppose my desktop computer is programmed for intelligence and there are various thoughts located within it. What would be the subject of those thoughts? What sort of artificial thinker would we have created, or made intelligent? Maybe there is no interesting answer to this question. There are many candidates, so to speak, for being the thinker. There is the computer itself, including mouse, keyboard, screen, and power cable. There are various parts of the computer. Maybe not just any part of the computer is a candidate for being the thinker: perhaps we can rule out the space bar on the keyboard. It might have to include the parts of the computer directly involved in thought and consciousness, if we can make any sense of that notion (the lright-to-left conjunct of thinking-subject minimalism). But there are no further restrictions. The thing composed of the computer together with the desk is a candidate, and so is the thing composed of the temporal part of the computer located this week and the the temporal part of the desk located last week. (The relaxed attitude involves a commitment to temporal parts and unrestricted composition.) And so on. But there is no saying which of these is the thinker of the thoughts going on in the computer. They're all thinkers. There is no psychological difference between them.

The same, according to the relaxed attitude, goes for natural thinkers. What is the thinker of <u>my</u> thoughts? Which thing am I? That question has no interesting answer either. There are many candidates: my brain; certain parts of my brain; this organism; various "arbitrary, undetached parts" of the organism, such as its upper half; the thing composed of my brain and my desk; and so on. But there is no saying which of these is the thinker of these thoughts. They all are.

At best there are certain linguistic conventions governing the use of personal pronouns and associated names and predicates. We use such words as 'I' and 'Olson' to refer to organisms and not to brains, or upper halves of organisms, or things composed of a brain and a desk. (For the most part we don't speak of such arbitrary and gerrymandered entities at all. For obvious practical reasons we ignore them.) That's what makes it correct to say that I have hands and that I weigh 150 pounds, and wrong to say that I weigh three pounds and am located within my cranium where no one has ever seen me. Similar rules govern the application of such general

---

located then.

terms as 'person', 'philosopher', or 'Hindu'. But again, there is no difference between the psychological properties of the beings to which we conventionally apply such terms and those of the beings to which we don't apply them (provided, at least, that such beings all include the things directly involved in the relevant mental activity).

So we needn't accept thinking-subject minimalism (or more precisely, its problematic left-to-right conjunct). More generally, we needn't worry about the question of artificial thinkers. There are no interesting facts about the metaphysical nature of thinking beings, natural or artificial--or at least none beyond the fact that they are material things.

But whatever merits the relaxed attitude may have, it fits badly with strong artificial intelligence. The attitude presupposes that there is never any one thing that thinks. If a thing has certain mental properties, then many other things, including many of those that have the first thing as parts, have the same mental properties. If my brain is conscious, so is my upper half, all of me but my left hand, this entire organism, and the thing composed of the organism and my desk. I am unsure how to formulate this principle in its full generality, but we might sum it up by saying that mental properties are "size-invariant".

A question now immediately arises: What other properties are size-invariant? Mass, for instance, is not: my upper half, this organism, and my brain cannot all have the same mass. Shape, color, location, temperature, hardness--and, obviously, size--are not either. It's hard to think of any other familiar property that is size-invariant. It would seem to be peculiar to mental properties. Why should this be so? What could it be about mental properties that makes them size-invariant?

The most obvious suggestion is that mental properties lack the objective reality of mass, size, shape, and other familiar properties. We find it useful for certain purposes--for explaining and predicting behaviour, for instance--to attribute mental properties to certain entities--to "take up the intentional stance" towards them, as Dennett famously put it (1981). And it may be useful to attribute mental properties to different objects for different purposes: to brains rather than organisms, say, or vice versa. But such attributions cannot be right or wrong in the way that attributions of mass or size or shape can be right or wrong. There are no hard facts about the mental properties of things. At best there are facts about how expedient it is for certain purposes to attribute them to something.

Few philosophers accept this instrumentalist or anti-realist view of the mental. More to the point for present purposes, it would deprive strong artificial intelligence of its interest. There would be no question of whether computers could ever have (or produce) genuine thought or consciousness. There would only be the question of how useful it could be for certain purposes to say that they did. And we all know the answer to that question. Meteorologists clearly find it useful to say that their computers "know" more about weather systems than they do themselves. We all find it useful for

certain purposes to attribute mental properties to photocopiers ("it thinks it's out of paper, even though I've just refilled the tray") and thermostats ("it's noticed that the room has warmed up"). The psychological difference between the most sophisticated computers of science fiction and the thermostats of the 19th century would be only a matter of degree--as would the difference between thermostats and human beings.

The instrumentalist view implies that there is no difference between simulating intelligence and actually producing it. Computers would be no more interesting to the philosophy of mind than thermostats. If there is any point in inquiring about the possibility of artificial intelligence, there must be real facts about which things can think or be conscious. And that looks inconsistent with the relaxed attitude.

11. Concluding remarks

For there to be genuine artificial intelligence, programming a computer in the right way and in the right circumstances must produce real thought or consciousness. But there must also be a subject of this thought or consciousness. And it is difficult to say what sort of being might be made intelligent by programming a computer. There is no account of what the subject of artificial intelligence would be that is anything like as obvious and natural as saying that the subject of natural intelligence is an organism.

It seems that artificially intelligent beings must be material things of some sort. If nothing else, any putative artificial thought or consciousness would consist of states or activities of a material thing, and it would be a mystery if being in mental states or engaging in mental activities fell short of thinking or being conscious. This may be why most discussions of artificial intelligence appear to take it for granted that artificial intelligence would make computers think.

But that view goes against what many people want to say about the persistence conditions of artificial thinkers: that they come into being when the relevant program starts running and are destroyed when it shuts down and the relevant data are erased, for instance. And no one supposes that all the parts of what we call a computer, including the keyboard, would be parts of the thinking being that would result from programming it for intelligence. At best an artificial thinker might be some part of a computer. But it's hard to say what part it would be, and the question must have an answer.

We cannot establish the possibility of artificial intelligence without having a satisfactory account of the sort of beings that would thereby be made intelligent, or at least showing that such an account exists. This has yet to be done. I am not saying that it can't be. But it's a safe bet that any such account will have surprising consequences for the nature of both artificial and natural thinkers.[8]

---

[8]For valuable advice on ancestors of this chapter I thank Luca Barlassina, Tom Cochrane, Mihretu Guta, and Karsten Witt.

References

Baker, L. R. 2000. Persons and Bodies. Cambridge University Press.

Boden, M., ed. 1990a. The Philosophy of Artificial Intelligence, Oxford University Press: 1-21.

---. 1990b. Introduction. In Boden 1990a: 1-21.

Copeland, J. 1993. Artificial Intelligence: A Philosophical Introduction. Blackwell.

Dennett, D. 1981. True believers: The intentional strategy and why it works. In A. F. Heath, ed., Scientific Explanation.

Haugeland, J. 1985. Artificial Intelligence: The Very Idea. MIT Press.

Hudson, H. 2007. I am not an animal! In P. van Inwagen and D. Zimmerman, eds., Persons: Human and Divine, Oxford University Press: 216-236.

Hume, D. 1978. A Treatise of Human Nature. Oxford University Press. (Original work 1739.)

Johnston, M. 2007. "Human beings" revisited: My body is not an animal. In D. Zimmerman, ed., Oxford Studies in Metaphysics 3: 33-74. Oxford University Press.

Newell, A. and H. Simon. 1976. Computer science as empirical enquiry: Symbols and search. Communications of the Association for Computing Machinery 19. Reprinted in Boden 1990a.

Noonan, H. 2010. The thinking animal problem and personal pronoun revisionism. Analysis 70: 93-98.

Olson, E. 2006. Temporal parts and timeless parthood. Noûs 40: 738-752.

---. 2007. What Are We? A Study in Personal Ontology. Oxford University Press.

---. 2018. The zombies among us. Noûs 52: 216–226.

Pollock, J. 1989. How to Build a Person. MIT Press.

Putnam, H. 1964. Robots: Machines or artificially created life? Journal of Philosophy 61: 668-91.

Quine, W. V. O. 1960. Word and Object. MIT Press.

Russell, S. and P. Norvig. 2010. Artificial Intelligence: A Modern Approach, 3e. Pearson.

Searle, J. 1980. Minds, brains and programs. Behavioral and Brain Sciences 3: 417-24. Reprinted in Boden 1990a.

Shoemaker, S. 1984. Personal identity: A materialist's account. In S. Shoemaker and R. Swinburne, Personal Identity. Blackwell.

---. 2008. Persons, animals, and identity. Synthese 163: 313-324.

Turing, A. 1950. Computing machinery and intelligence. Mind 59: 433-60. Reprinted in Boden 1990a.

van Inwagen, P. 1990. Material Beings. Cornell University Press.