Mind Uploading: A Philosophical Analysis

David J. Chalmers

[Published in (D. Broderick and R. Blackford, eds.) *Intelligence Unbound: The Future of Uploaded and Machine Minds* (Blackwell, 2014). This paper is an excerpt from "The Singularity: A Philosophical Analysis," published in *The Journal of Consciousness Studies* (17: 7-65, 2010).]

In the long run, if we are to match the speed and capacity of nonbiological systems, we will probably have to dispense with our biological core entirely. This might happen through a gradual process in which parts of our brain are replaced over time, or via a process of scanning our brains and loading the result into a computer, then enhancing the resulting processes. Either way, the result is likely to be an enhanced nonbiological system, most likely a computational system.

The process of migration from brain to computer is often called *mind uploading,* or simply *uploading* for short. It can take many different forms. It can involve gradual replacement of brain parts (gradual uploading), instant scanning and activation (instant uploading), or scanning followed by later activation (delayed uploading). It can involve destruction of the original brain parts (destructive uploading), preservation of the original brain (nondestructive uploading), or reconstruction of cognitive structure from records (reconstructive uploading). We can only speculate about what form uploading technology will take, but some forms have been widely discussed.[1]

For concreteness, I will describe relatively specific forms of three sorts of uploading: destructive uploading, gradual uploading, and nondestructive uploading.

Destructive uploading: It is widely held that this may be the first form of uploading to be feasible. One possible form involves serial sectioning. Here one freezes a brain, and proceeds to analyze its structure layer-by-layer. In each layer one records the distribution of neurons and other relevant components, along with the character of their interconnections. One then loads all this information into a computer model that includes an accurate simulation of neural behavior and dynamics. The result might be an emulation of the original brain.

Gradual uploading: Here the most widely-discussed method is that of nanotransfer. One or more nanotechnology devices (perhaps tiny robots) are inserted into the brain and each attaches itself to a single neuron, learning to simulate the behavior of the associated neuron and also learning about its connectivity. Once it simulates the neuron's behavior well enough, it takes the place of the original neuron, perhaps leaving receptors and effectors in place and uploading the relevant processing to a computer via radio transmitters. It then moves to other neurons and repeats the procedure, until eventually every neuron has been replaced by an emulation, and perhaps all processing has been uploaded to a computer.

Nondestructive uploading: The nanotransfer method might in principle be used in a nondestructive form. The holy grail here is some sort of noninvasive method of brain imaging, analogous to functional magnetic resonance imaging, but with fine enough grain that neural and synaptic dynamics can be recorded. No such technology is currently on the horizon, but imaging technology is an area of rapid progress.

In all of its forms, uploading raises many questions. From a self-interested point of view, the key question is: will I survive? This question itself divides into two parts, each corresponding to one of the hardest questions in philosophy: the questions of consciousness and personal identity. First, will an uploaded version of me be conscious? Second, will it be me?

Uploading and Consciousness

Ordinary human beings are conscious. That is, there is something it is like to be us. We have conscious experiences with a subjective character: there is something it is like for us to see, to hear, to feel, and to think. These conscious experiences lie at the heart of our mental lives, and are a central part of what gives our lives meaning and value. If we lost the capacity for consciousness, then in an important sense, we would no longer exist.

Before uploading, then, it is crucial to know whether the resulting upload will be conscious. If my only residue is an upload and the upload has no capacity for consciousness, then arguably I do not exist at all. And if there is a sense in which I exist, this sense at best involves a sort of zombified existence. Without consciousness, this would be a life of greatly diminished meaning and value.

Can an upload be conscious? The issue is complicated by the fact that our understanding of consciousness is so poor. No-one knows just why or how brain processes give rise to consciousness. Neuroscience is gradually discovering various neural correlates of consciousness, but this research program largely takes the existence of consciousness for granted. There is nothing even approaching an orthodox theory of why there is consciousness in the first place. Correspondingly, there is nothing even approaching an orthodox theory of what sorts of systems can be conscious and what systems cannot be.

One central problem is that consciousness seems to be a further fact about conscious systems, at least in the sense that knowledge of the physical structure of such a system does not tell one all about the conscious experiences of such a system.[2]

Complete knowledge of physical structure might tell one all about a system’s objective behavior and its objective functioning, which is enough to tell whether the system is alive, and whether it is intelligent. But this sort of knowledge alone does not seem to answer all the questions about a system’s subjective experience.

A famous illustration here is Frank Jackson’s case of Mary, the neuroscientist in a black-and-white room, who knows all about the physical processes associated with color but does not know what it is like to see red. If this is right, complete physical knowledge leaves open certain questions about the conscious experience of color. More broadly, a complete physical description of a system such as a mouse does not appear to tell us what it is like to be a mouse, and indeed whether there is anything it is like to be a mouse. Furthermore, we do not have a "consciousness meter" that can settle the matter directly. So given any system, biological or artificial, there will at least be a substantial and unobvious question about whether it is conscious, and about what sort of consciousness it has.

Still, whether one thinks there are further facts about consciousness or not, one can at least raise the question of what sort of systems are conscious. Here philosophers divide into multiple camps. Biological theorists of consciousness hold that consciousness is essentially biological and that no nonbiological system can be conscious. Functionalist theorists of consciousness hold that what matters to consciousness is not biological makeup but causal structure and causal role, so that a nonbiological system can be conscious as long as it is organized correctly.

The philosophical issue between biological and functionalist theories is crucial to the practical question of whether or not we should upload. If biological theorists are correct, uploads cannot be conscious, so we cannot survive consciously in uploaded form. If functionalist theorists are correct, uploads almost certainly can be conscious, and this obstacle to uploading is removed. My own view is that functionalist theories are closer to the truth here. It is true that we have no idea how a nonbiological system, such as a silicon computational system, could be conscious.

But the fact is that we also have no idea how a biological system, such as a neural system, could be conscious. The gap is just as wide in both cases. And we do not know of any principled differences between biological and nonbiological systems that suggest that the former can be conscious and the latter cannot. In the absence of such principled differences, I think the default attitude should be that both biological and nonbiological systems can be conscious. I think that this view can be supported by further reasoning.

To examine the matter in more detail: Suppose that we can create a perfect upload of a brain inside a computer. For each neuron in the original brain, there is a computational element that duplicates its input/output behavior perfectly. The same goes for non-neural and subneural components of the brain, to the extent that these are relevant. The computational elements are connected to input and output devices (artificial eyes and ears, limbs and bodies), perhaps in an ordinary physical environment or perhaps in a virtual environment. On receiving a visual input, say, the upload goes through processing isomorphic to what goes on in the original brain. First artificial analogs of eyes and the optic nerve are activated, then computational analogs of lateral geniculate nucleus and the visual cortex, then analogs of later brain areas, ultimately resulting in a (physical or virtual) action analogous to one produced by the original brain.

In this case we can say that the upload is a functional isomorph of the original brain. Of course it is a substantive claim that functional isomorphs are possible. If some elements of cognitive processing function in a noncomputable way, for example so that a neuron's input/output behavior cannot even be computationally simulated, then an algorithmic functional isomorph will be impossible. But if the components of cognitive functioning are themselves computable, then a functional isomorph is possible. Here I will assume that functional isomorphs are possible in order to ask whether they will be conscious.

I think the best way to consider whether a functional isomorph will be conscious is to consider a gradual uploading process such as nanotransfer.

Here we upload different components of the brain one by one, over time. This might involve gradual replacement of entire brain areas with computational circuits, or it might involve uploading neurons one at a time. The components might be replaced with silicon circuits in their original location, or with processes in a computer connected by some sort of transmission to a brain. It might take place over months or years, or over hours.

If a gradual uploading process is executed correctly, each new component will perfectly emulate the component it replaces, and will interact with both biological and nonbiological components around it in just the same way that the previous component did. So the system will behave in exactly the same way that it would have without the uploading. In fact, if we assume that the system cannot see or hear the uploading, then the system need not notice that any uploading has taken place. Assuming that the original system said that it was conscious, so will the partially uploaded system. The same applies throughout a gradual uploading process, until we are left with a purely nonbiological system.

What happens to consciousness during a gradual uploading process? There are three possibilities. It might suddenly disappear, with a transition from a fully complex conscious state to no consciousness when a single component is replaced. It might gradually fade out over more than one replacements, with the complexity of the system's conscious experience reducing via intermediate steps. Or it might stay present throughout.

Sudden disappearance is the least plausible option. Given this scenario, we can move to a scenario in which we replace the key component by replacing ten or more subcomponents in turn, and then reiterate the question. Either new scenario will involve a gradual fading across a number of components, or a sudden disappearance. If the former, this option is reduced to the fading option. If the latter, we can reiterate. In the end we will either have gradual fading or sudden disappearance when a single tiny component (a neuron or a subneural element, say) is replaced. The latter seems extremely unlikely.

Gradual fading also seems implausible. In this case there will be intermediate steps in which the system is conscious but its consciousness is partly faded, in that it is less complex than the original conscious state. Perhaps some element of consciousness will be gone (visual but not auditory experience, for example) or perhaps some distinctions in experience will be gone (colors reduced from a three-dimensional color space to black and white, for example). By hypothesis the system will be functioning and behaving the same way as ever, though, and will not show any signs of noticing the change. It is plausible that the system will not believe that anything has changed, despite a massive difference in its conscious state. This requires a conscious system that is deeply out of touch with its own conscious experience.

We can imagine that at a certain point partial uploads become common, and that many people have had their brains partly replaced by silicon computational circuits. On the sudden disappearance view, there will be states of partial uploading such that any further change will cause consciousness to disappear, with no difference in behavior or organization. People in these states may have consciousness constantly flickering in and out, or at least might undergo total zombification with a tiny change. On the fading view, these people will be wandering around with a highly degraded consciousness, although they will be functioning as always and swearing that nothing has changed. In practice, both hypotheses will be difficult to take seriously. So I think that by far the most plausible hypothesis is that full consciousness will stay present throughout. On this view, all partial uploads will still be fully conscious, as long as the new elements are functional duplicates of the elements they replace. By gradually moving through fuller uploads, we can infer that even a full upload will be conscious.

At the very least, it seems very likely that partial uploading will convince most people that uploading preserves consciousness. Once people are confronted with friends and family who have undergone limited partial uploading and are behaving normally, few people will seriously think that they lack consciousness. And gradual extensions to full uploading will convince most people that these systems are conscious at well. Of course it remains at least a logical possibility that this process will gradually or suddenly turn everyone into zombies. But once we are confronted with partial uploads, that hypothesis will seem akin to the hypothesis that people of different ethnicities or genders are zombies.

If we accept that consciousness is present in functional isomorphs, should we also accept that isomorphs have qualitatively identical states of consciousness? This conclusion does not follow immediately. But I think that an extension of this reasoning (the "dancing qualia" argument in Chalmers 1996) strongly suggests such a conclusion.

If this is right, we can say that consciousness is an organizational invariant: that is, systems with the same patterns of causal organization have the same states of consciousness, no matter whether that organization is implemented in neurons, in silicon, or in some other substrate. We know that some properties are not organizational invariants (being wet, say) while other properties are (being a computer, say). In general, if a property is not an organizational invariant, we should not expect it to be preserved in a computer simulation (a simulated rainstorm is not wet). But if a property is an organizational invariant, we should expect it to be preserved in a computer simulation (a simulated computer is a computer). So given that consciousness is an organizational invariant, we should expect a good enough computer simulation of a conscious system to be conscious, and to have the same sorts of conscious states as the original system.

This is good news for those who are contemplating uploading. But there remains a further question.

Uploading and Personal Identity

Suppose that I can upload my brain into a computer? Will the result be me?[3]

On the *optimistic* view of uploading, the upload will be the same person as the original. On the *pessimistic* view of uploading, the upload will not be the same person as the original. Of course if one thinks that uploads are not conscious, one may well hold the pessimistic view on the grounds that the upload is not a person at all. But even if one thinks that uploads are conscious and are persons, one might still question whether the upload is the same person as the original. Faced with the prospect of destructive uploading (in which the original brain is destroyed), the issue between the optimistic and pessimistic view is literally a life-or-death question. On the optimistic view, destructive uploading is a form of survival. On the pessimistic view, destructive uploading is a form of death. It is as if one has destroyed the original person, and created a simulacrum in their place.

An appeal to organizational invariance does not help here. We can suppose that I have a perfect identical twin whose brain and body are molecule-for-molecule duplicates of mine. The twin will then be a functional isomorph of me and will have the same conscious states as me. This twin is *qualitatively* identical to me: it has exactly the same qualities as me. But it is not *numerically* identical to me: it is not me. If you kill the twin, I will survive. If you kill me (that is, if you destroy *this* system) and preserve the twin, I will die. The survival of the twin might be some consolation to me, but from a self-interested point of view this outcome seems much worse than the alternative.

Once we grant that my twin and I have the same organization but are not the same person, it follows that personal identity is not an organizational invariant. So we cannot count on the fact that uploading preserves organization to guarantee that uploading preserves identity. On the pessimistic view, destructive uploading is at best akin to creating a sort of digital twin while destroying me.

These questions about uploading are closely related to parallel questions about physical duplication. Let us suppose that a teletransporter creates a molecule-for-molecule duplicate of a person out of new matter while destroying or dissipating the matter in the original system. Then on the optimistic view of teletransportation, it is a form of survival, while on the pessimistic view, it is a form of death. Teletransportation is not the same as uploading: it preserves physical organization where uploading preserves only functional organization in a different physical substrate. But at least once one grants that uploads are conscious, the issues raised by the two cases are closely related.

In both cases, the choice between optimistic and pessimistic views is a question about personal identity: under what circumstances does a person persist over time? Here there is a range of possible views. An extreme view on one end (perhaps held by no-one) is that exactly the same matter is required for survival (so that when a single molecule in the brain is replaced, the original person ceases to exist). An extreme view on the other end is that merely having the same sort of conscious states suffices for survival (so that from my perspective there is no important difference between killing this body and killing my twin's body). In practice, most theorists hold that a certain sort of *continuity* or *connectedness* over time is required for survival. But they differ on what sort of continuity or connectedness is required.

There are a few natural hypotheses about what sort of connection is required. *Biological* theories of identity hold that survival of a person requires the intact survival of a brain or a biological organism. *Psychological* theories of identity hold that survival of a person requires the right sort of psychological continuity over time (preservation of memories, causally related mental states, and so on). *Closest-continuer* theories hold that a person survives as the most closely related subsequent entity, subject to various constraints.

Biological theorists are likely to hold the pessimistic view of teletransportation, and are even more likely to hold the pessimistic view of uploading. Psychological theorists are more likely to hold the optimistic view of both, at least if they accept that an upload can be conscious. Closest-continuer theorists are likely to hold that the answer depends on whether the uploading is destructive, in which case the upload will be the closest continuer, or nondestructive (in which case the biological system will be the closest continuer.4

I do not have a settled view about these questions of personal identity and find them very puzzling. I am more sympathetic with a psychological view of the conditions under which survival obtains than with a biological view, but I am unsure of this, for reasons I will elaborate later. Correspondingly, I am genuinely unsure whether to take an optimistic or a pessimistic view of destructive uploading. I am most inclined to be optimistic, but I am certainly unsure enough that I would hesitate before undergoing destructive uploading.

To help clarify the issue, I will present an argument for the pessimistic view and an argument for the optimistic view, both of which run parallel to related arguments that can be given concerning teletransportation.

The argument from nondestructive uploading

Suppose that yesterday Dave was uploaded into a computer. The original brain and body were not destroyed, so there are now two conscious beings: BioDave and DigiDave. BioDave's natural attitude will be that he is the original system and that DigiDave is at best some sort of branchline copy. DigiDave presumably has some rights, but it is natural to hold that he does not have BioDave's rights. For example, it is natural to hold that BioDave has certain rights to Dave's possessions, his friends, and so on, where DigiDave does not. And it is natural to hold that this is because BioDave is Dave: that is, Dave has survived as BioDave and not as DigiDave.

If we grant that, in a case of nondestructive uploading, DigiDave is not identical to Dave, then it is natural to question whether destructive uploading is any different. If Dave did not survive as DigiDave when the biological system was preserved, why should he survive as DigiDave when the biological system is destroyed?

We might put this in the form of an argument for the pessimistic view, as follows:

1. In nondestructive uploading, DigiDave is not identical to Dave.

2. If in nondestructive uploading, DigiDave is not identical to Dave, then in destructive uploading, DigiDave is not identical to Dave.

———————

3. In destructive uploading, DigiDave is not identical to Dave.

Various reactions to the argument are possible. A pessimist about uploading will accept the conclusion. An optimist about uploading will presumably deny one of the premises. One option is to deny premise 2, perhaps because one accepts a closest-continuer theory: when BioDave exists, he is the closest continuer, but when he does not, DigiDave is the closest continuer. Some will find that this makes one's survival and status an unacceptably extrinsic matter, though.

Another option is to deny premise 1, holding that even in nondestructive uploading DigiDave is identical to Dave. Now, in this case it is hard to deny that BioDave is at least as good a candidate as DigiDave, so this option threatens to have the consequence that DigiDave is also identical to BioDave. This consequence is hard to swallow as BioDave and DigiDave may be qualitatively distinct conscious beings, with quite different physical and mental states by this point.

A third and related option holds that nondestructive uploading should be regarded as a case of fission. A paradigmatic fission case is one in which the left and right hemispheres of a brain are separated into different bodies, continuing to function well on their own with many properties of the original. In this case it is uncomfortable to say that both resulting systems are identical to the original, for the same reason as above. But one might hold that they are nevertheless on a par. For example, Parfit (1984) suggests although the original system is not identical to the left-hemisphere system or to the right-hemisphere system, it stands in a special relation R (which we might call survival) to both of them, and he claims that this relation rather than numerical identity is what matters. One could likewise hold that in a case of nondestructive uploading, Dave survives as both BioDave and DigiDave (even if he is not identical to them), and hold that survival is what matters. Still, if survival is what matters, this option does raise uncomfortable questions about whether DigiDave has the same rights as BioDave when both survive.

The argument from gradual uploading

Suppose that 1% of Dave's brain is replaced by a functionally isomorphic silicon circuit. Next suppose that another 1% is replaced, and then another 1%. We can continue the process for 100 months, after which a wholly uploaded system will result. We can suppose that functional isomorphism preserves consciousness, so that the system has the same sort of conscious states throughout.

Let Dave_n be the system after *n* months. Will Dave_1 , the system after one month, be Dave? It is natural to suppose so. The same goes for Dave_2 and Dave_3. Now consider Dave_{100} , the wholly uploaded system after 100 months. Will Dave_{100} be Dave? It is at least very natural to hold that it will be. We could turn this into an argument as follows.

1. For all $n < 100$, Dave_{n+1} is identical to Dave_n.

2. If for all $n < 100$, Dave_{n+1} is identical to Dave_n, then Dave_{100} is identical to Dave.

———————

3. Dave_{100} is identical to Dave.

On the face of it, premise 2 is hard to deny: it follows from repeated application of the claim that when $a = b$ and $b = c$, then $a = c$. On the face of it, premise 1 is hard to deny too: it is hard to see how changing 1% of a system will change its identity. Furthermore, if someone denies premise 1, we can repeat the thought-experiment with ever smaller amounts of the brain being replaced, down to single neurons and even smaller. Maintaining the same strategy will require holding that replacing a single neuron can in effect kill a person. That is a hard conclusion to accept. Accepting it would raise the possibility that everyday neural death may be killing us without our knowing it. One could resist the argument by noting that it is a sorites or slippery-slope argument, and by holding that personal identity can come in degrees or can have indeterminate cases. One could also drop talk of identity and instead hold that survival can come in degrees. For example, one might hold that each Dave_n survives to a large degree as Dave_{n+1} but to a smaller degree as later systems.

On this view, the original person will gradually be killed by the replacement process. This view requires accepting the counterintuitive view that survival can come in degrees or be indeterminate in these cases, though. Perhaps more importantly, it is not clear why one should accept that Dave is gradually killed rather than existing throughout. If one were to accept this, it would again raise the question of whether the everyday replacement of matter in our brains over a period of years is gradually killing us also.

My own view is that in this case, it is very plausible that the original system survives. Or at least, it is plausible that insofar as we ordinarily survive over a period of many years, we could survive gradual uploading too. At the very least, as in the case of consciousness, it seems that if gradual uploading happens, most people will become convinced that it is a form of survival. Assuming the systems are isomorphic, they will say that everything seems the same and that they are still present. It will be very unnatural for most people to believe that their friends and families are being killed by the process. Perhaps there will be groups of people who believe that the process either suddenly or gradually kills people without them or others noticing, but it is likely that this belief will come to seem faintly ridiculous.

Once we accept that gradual uploading over a period of years might preserve identity, the obvious next step is to speed up the process. Suppose that Dave's brain is gradually uploaded over a period of hours, with neurons replaced one at a time by functionally isomorphic silicon circuits. Will Dave survive this process? It is hard to see why a period of hours should be different in principle from a period of years, so it is natural to hold that Dave will survive.

To make the best case for gradual uploading, we can suppose that the system is active throughout, so that there is consciousness through the entire process. Then we can argue: (i) consciousness will be continuous from moment to moment (replacing a single neuron or a small group will not disrupt continuity of consciousness), (ii) if consciousness is continuous from moment to moment, it will be continuous throughout the process, (iii) if consciousness is continuous throughout the process, there will be a single stream of consciousness throughout, (iv) if there is a single stream of consciousness throughout, then the original person survives throughout. One could perhaps deny one of the premises, but denying any of them is uncomfortable. My own view is that continuity of consciousness (especially when accompanied by other forms of psychological continuity) is an extremely strong basis for asserting continuation of a person.

We can then imagine speeding up the process from hours to minutes. The issues here do not seem different in principle. One might then speed it up to seconds. At a certain point, one will arguably start replacing large enough chunks of the brain from moment to moment that the case for continuity of consciousness between moments is less secure. Still, once we grant that uploading over a period of minutes preserves identity, it is at least hard to see why uploading over a period of seconds should not.

As we upload faster and faster, the limit point is instant destructive uploading, where the whole brain is replaced at once. Perhaps this limit point is different from everything that came before it, but this is at least unobvious. We might formulate this as an argument for the optimistic view of destructive uploading. Here it is to be understood that both the gradual uploading and instant uploading are destructive in that they destroy the original brain.

1. Dave survives as Dave_{100} in gradual uploading.

2. If Dave survives as Dave_{100} in gradual uploading, Dave survives as DigiDave in instant uploading.

____

3. Dave survives as DigiDave in instant uploading.

I have in effect argued for the first premise above, and there is at least a prima facie case for the second premise, in that it is hard to see why there is a difference in principle between uploading over a period of seconds and doing so instantly. As before, this argument parallels a corresponding argument about teletransportation (gradual matter replacement preserves identity, so instant matter replacement preserves identity too), and the considerations available are similar. An opponent could resist this argument by denying premise 1 along the lines suggested earlier, or perhaps better, by denying premise 2. A pessimist about instant uploading, like a pessimist about teletransportation, might hold that intermediate systems play a vital role in the transmission of identity from one system to another.

This is a common view of the ship of Theseus, in which all the planks of a ship are gradually replaced over years. It is natural to hold that the result is the same ship with new planks. It is plausible that the same holds even if the gradual replacement is done within days or minutes. By contrast, building a duplicate from scratch without any intermediate cases arguably results in a new ship. Still, it is natural to hold that the question about the ship is in some sense a verbal question or a matter for stipulation, while the question about personal survival runs deeper than that. So it is not clear how well one can generalize from the ship case to the case of persons.

Where things stand

We are in a position where there are at least strongly suggestive arguments for both the optimistic and pessimistic views of destructive uploading. The arguments have diametrically opposed conclusions, so they cannot both be sound. My own view is that the optimist's best reply to the argument from nondestructive uploading is the fission reply, and the pessimist's best reply to the argument from gradual uploading is the intermediate-case reply. My instincts favor optimism, but as before I cannot be certain which view is correct.

Still, I am confident that the safest form of uploading is gradual uploading, and I am reasonably confident that gradual uploading is a form of survival. So if at some point in the future I am faced with the choice between uploading and continuing in an increasingly slow biological embodiment, then as long as I have the option of gradual uploading, I will be happy to do so. Unfortunately, I may not have that option. It may be that gradual uploading technology will not be available in my lifetime. It may even be that no adequate uploading technology will be available at all in my lifetime. This raises the question of whether there might still be a place for me, or for any currently existing humans, in a future of artificial intelligence.

Uploading after brain preservation

One possibility is that we can preserve our brains for later uploading. Cryonic technology offers the possibility of preserving our brains in a low-temperature state shortly after death, until such time as the technology is available to reactivate the brain or perhaps to upload the information in it. Of course much information may be lost in death, and at the moment we do not know whether cryonics preserves information sufficient to reactivate or reconstruct anything akin to a functional isomorph of the original. But one can at least hope that, after an intelligence explosion, extraordinary technology might be available.

If there is enough information for reactivation or reconstruction, will the resulting system be me? In the case of reactivation, it is natural to hold that the reactivated system will be akin to a person waking up after a long coma, so that the original person will survive. One might then gradually upload the brain and integrate the result into a technologically advanced world. Alternatively, one might create an uploaded system from the brain without ever reactivating it. Whether one counts this as survival will depend on one's attitude to ordinary destructive and nondestructive uploading. If one is an optimist about these, then one might also be an optimist about uploading from a preserved brain.

Another possible outcome is that there will be first a series of uploads from a preserved brain, using better and better scanning technology, and eventually reactivation of the brain. Here, an optimist about uploading might see this as a case of fission, while a pessimist might hold that only the reactivated system is identical to the original.

In these cases, our views of the philosophical issues about uploading affect our decisions not just in the distant future but in the near term. Even in the near term, anyone with enough money or suitable insurance has the option of having their brain or whole body cryonically preserved, and of leaving instructions about how to deal with the brain as technology develops. Our philosophical views about the status of uploading may well make a difference to the instructions that we should leave.

Of course most people do not preserve their brains, and even those who choose to do so may die in a way that renders preservation impossible. Are there other routes to survival in an advanced future world shared with superintelligent AIs?

Reconstructive uploading

The final alternative is reconstruction of the original system from records, and especially reconstructive uploading, in which an upload of the original system is reconstructed from records. Here, the records might include brain scans and other medical data; any available genetic material; audio and video records of the original person; their writings; and the testimony of others about them. These records may seem limited, but it is not out of the question that a superintelligent AI could go a long way with them. Given constraints on the structure of a human system, even limited information might make a good amount of reverse engineering possible. And detailed information, as might be available in extensive video recordings and in detailed brain images, might in principle make it possible for a superintelligence to reconstruct something close to a functional isomorph of the original system.

The question then arises: is reconstructive uploading a form of survival? If we reconstruct a functional isomorph of Einstein from records, will it be Einstein? Here, the pessimistic view says that this is at best akin to a copy of Einstein surviving. The optimistic view says that it is akin to having Einstein awaken from a long coma.

Reconstructive uploading from brain scans is closely akin to ordinary (nongradual) uploading from brain scans, with the main difference being the time delay, and perhaps the continued existence in the meantime of the original person. One might see it as a form of delayed destructive or nondestructive uploading. If one regards nondestructive uploading as survival (perhaps through fission), one will naturally regard reconstructive uploading the same way. If one regards destructive but not nondestructive uploading as survival because one embraces a closest continuer theory, one might also regard reconstructive uploading as survival (at least if the original biological system is gone). If one regards neither as survival, one will probably take the same attitude to reconstructive uploading. Much the same options plausibly apply to reconstructive uploading from other sources of information.

Upshot

I think that gradual uploading is certainly the safest method of uploading.

A number of further questions about uploading remain. Of course there are any number of social, legal, and moral issues that I have not begun to address. Here I address just two further questions.

One question concerns cognitive enhancement. Suppose that before or after uploading, our cognitive systems are enhanced to the point that they use a wholly different cognitive architecture. Would we survive this process? Again, it seems to me that the answers are clearest in the case where the enhancement is gradual. If my cognitive system is overhauled one component at a time, and if at every stage there is reasonable psychological continuity with the previous stage, then I think it is reasonable to hold that the original person survives.

Another question is a practical one. If reconstructive uploading will eventually be possible, how can one ensure that it happens? There have been billions of humans in the history of the planet. It is not clear that our successors will want to reconstruct every person that ever lived, or even every person of which there are records. So if one is interested in immortality, how can one maximize the chances of reconstruction? One might try keeping a bank account with compound interest to pay them for doing so, but it is hard to know whether our financial system will be relevant in the future, especially after an intelligence explosion.

My own strategy is to write about a future of artificial intelligence and about uploading. Perhaps this will encourage our successors to reconstruct me, if only to prove me wrong.

Notes

[1] See Sandberg and Bostrom 2008 and Strout 2006 for detailed discussion of potential uploading technology. See Egan 1994 and Sawyer 2005 for fictional explorations of uploading.

[2] The further-fact claim here is simply that facts about consciousness are epistemologically further facts, so that knowledge of these facts is not settled by reasoning from microphysical knowledge alone. This claim is compatible with materialism about consciousness. A stronger claim is that facts about consciousness are ontologically further facts, involving some distinct elements in nature—e.g. fundamental properties over and above fundamental physical properties. In the framework of Chalmers (2003), a type-A materialist (e.g., Daniel Dennett) denies that consciousness involves epistemologically further facts, a type-B materialist (e.g., Ned Block) holds that consciousness involves

epistemologically but not ontologically further facts, while a property dualist (e.g., me) holds that consciousness involves ontologically further facts. It is worth noting that the majority of materialists (at least in philosophy) are type-B materialists and hold that there are epistemologically further facts.

3 It will be obvious to anyone who has read Derek Parfit's *Reasons and Persons* that the current discussion is strongly influenced by Parfit's discussion there. Parfit does not discuss uploading, but his discussion of related phenomena such as teletransportation can naturally be seen to generalize. In much of what follows I am simply carrying out aspects of the generalization.

4 In the 2009 PhilPapers survey of 931 professional philosophers [philpapers.org/surveys, 34% accepted or leaned toward a psychological view, 17% a biological view, and 12% a further-fact view (others were unsure, unfamiliar with the issue, held that there is no fact of the matter, and so on). Respondents were not asked about uploading, but on the closely related question of whether teletransportation (with new matter) is survival or death, 38% accepted or leaned toward survival and 31% death. Advocates of a psychological view broke down 67/22% for survival/death, while advocates of biological and further-fact views broke down 12/70% and 33/47% respectively.

Bibliography

Block, N. 1981. Psychologism and behaviorism. Philosophical Review 90:5-43.

Bostrom, N. 1998. How long before superintelligence? International Journal of Future Studies 2. http://www.nickbostrom.com/superintelligence.html

Bostrom, N. 2003. Ethical issues in advanced artificial intelligence. In (I. Smit, ed) Cognitive,

Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2.

International Institute of Advanced Studies in Systems Research and Cybernetics.

Bostrom, N. 2006. Quantity of experience: Brain-duplication and degrees of consciousness.

Minds and Machines 16:185-200.

Campbell, J.W. 1932. The last evolution. Amazing Stories.

Chalmers, D.J. 1990. How Cartesian dualism might have been true. http://consc.net/notes/dualism.html

Chalmers, D.J. 1995. Minds, machines, and mathematics. Psyche 2:11-20.

Chalmers, D.J. 1996. The Conscious Mind. Oxford University Press.

Chalmers, D.J. 2003. Consciousness and its place in nature. In (S. Stich and F. Warfield, eds)

Blackwell Guide to the Philosophy of Mind. Blackwell.

Chalmers, D.J. 2005. The Matrix as metaphysics. In (C. Grau, ed.) Philosophers Explore the Matrix. Oxford University Press.

Chalmers, D.J. 2006. Perception and the fall from Eden. In (T. Gendler and J. Hawthorne, eds) Perceptual Experience. Oxford University Press.

Clark, A. and Chalmers, D. 1998. The extended mind. Analysis 58:7-19.

Dainton, B. 2008. The Phenomenal Self. Oxford University Press.

Dreyfus, H. 1972. What Computers Can't Do.

Egan, G. 1994. Permutation City. Orion/Millenium.

Floridi, L. and Sanders, J.W. 2004. On the morality of artificial agents. Minds and Machines 14:349-79.

Flynn, J.R. 2007. What is Intelligence? Cambridge University Press.

Good, I.J. 1965. Speculations concerning the first ultraintelligent machine. In (F. Alt & M. Rubino, eds) Advances in Computers, vol 6.

Hanson, R. 1994. If uploads come first: The crack of a future dawn. Extropy 6:2. http://hanson.gmu.edu/uploads.html

Hanson, R. 2008. Economics of the singularity. IEEE Spectrum, June, 37-43.

Hanson, R, 2009. Prefer law to values. http://www.overcomingbias.com/2009/10/prefer-law-tovalues.html

Hofstadter, D.R. 2005. Moores law, artificial evolution, and the fate of humanity. In (L. Booker, S. Forrest, M. Mitchell, and R. Riolo, eds) Perspectives on Adaptation in Natural and Artificial Systems. Oxford University Press.

Joy, W. 2000. Why the future doesnt need us. Wired 8.04, July 2000.

Kurzweil, R. 2005. The Singularity is Near.

Legg, S. 2008. Machine Superintelligence.

Lucas, J.R. 1961. Minds, machines, and Gödel. Philosophy 36:112-27.

Moravec, H. 1988. Mind Children: The Future of Robot and Human Intelligence. Harvard University Press.

Moravec, H. 1998. Robots: Mere Machine to Transcendent Mind. Oxford University Press.

Omohundro, S. 2007. The nature of self-improving artificial intelligence. http://steveomohundro.com/scientificcontributions/

Omohundro, S. 2008. The basic AI drives. In (P. Wang, B. Goertzel, and S. Franklin, eds) Proceedings of the First AGI Conference. Frontiers in Artificial Intelligence and Applications, Volume 171. IOS Press.

Parfit, D.A. 1984. Reasons and Persons. Oxford University Press.

Penrose, R. 1994. Shadows of the Mind. Oxford University Press.

Sandberg, A. & Bostrom, N. 2008. Whole brain emulation: A roadmap. Technical report 2008-3, Future for Humanity Institute, Oxford University. http://www.fhi.ox.ac.uk/Reports/2008-3.pdf

Sawyer, R. 2000. Calculating God. Tor.

Sawyer, R. 2005. Mindscan. Tor.

Schmidhuber, J. 2003. G¨ odel machines: Self-referential universal problem solvers making provably optimal self-improvements. http://arxiv.org/abs/cs.LO/0309048

Searle, J. 1980. Minds, brains, and programs. Behavioral and Brain Sciences 3:417-57.

Shalizi, C. 2007. g, a statistical myth. http://bactra.org/weblog/523.html

Smart, J. 1999-2008. Brief history of intellectual discussion of accelerating change.
http://www.accelerationwatch.com/history brief.html

Solomonoff, F. 1985. The time scale of artificial intelligence: Reflections on social effects. North-Holland Human Systems Management 5:149-153. Elsevier.

Strout, J. 2006. The mind uploading home page. http://www.ibiblio.org/jstrout/uploading/

Ulam, S. 1958. John von Neumann 1903-1957. Bulletin of the American Mathematical Society 64 (number 3, part 2): 1-49.

Unger, P. 1990. Identity, Consciousness, and Value. Oxford University Press.

Vinge, V. 1983. First word. Omni, January 1983, p. 10.

Vinge, V. 1993. The coming technological singularity: How to survive in the post-human era. Whole Earth Review, winter 1993.

Wallach, W & Allen, C. 2009. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.

Yudkowsky, E. 1996. Staring at the singularity. http://yudkowsky.net/obsolete/singularity.html

Yudkowsky, E. 2002. The AI-box experiment. http://yudkowsky.net/singularity/aibox

Yudkowsky, E. 2007. Three major singularity schools. http://yudkowsky.net/singularity/schools

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. In (N. Bostrom, ed.) Global Catastrophic Risks. Oxford University Press.