Aspects of mind uploading

Olle Häggström

Chalmers University of Technology and the Institute for Future Studies

Abstract. Mind uploading is the hypothetical future technology of transferring human minds to computer hardware using whole-brain emulation. After a brief review of the technological prospects for mind uploading, a range of philosophical and ethical aspects of the technology are reviewed. These include questions about whether uploads will have consciousness and whether uploading will preserve personal identity, as well as what impact on society a working uploading technology is likely to have and whether these impacts are desirable. The issue of whether we ought to move forwards towards uploading technology remains as unclear as ever.

1. Introduction

According to transhumanism, the current form of *homo sapiens* should not be thought of as the end product of evolution, but rather as a transitionary state on the path towards posthuman life forms that we can achieve by enhancing ourselves, e.g., pharmacologically, genetically or by direct brain-computer interfaces. Transhumanists claims that such a development is not only possible but desirable. In this they are opposed by so-called bioconservatives, who maintain that it is undesirable or even forbidden, pointing at a variety of reasons, religious as well as secular. See Bostrom and Savulescu (2009) and Hansell and Grassie (2011) for collections of papers representing both sides of the transhumanism vs bioconservatism debate, or Chapter 3 of Häggström (2016) for my own attempt at a fair and balanced summary of the key issues.

The ultimate dream for many transhumanists is *uploading*, defined here as the transferring of our minds to computer hardware using *whole-brain emulation*, the idea being that a good enough simulation of a human brain simultaneously gives a simulation of the human mind, and that if the simulation is sufficiently detailed and accurate, then it goes from being a mere simulation to being in all relevant aspects an *exact replica* of the mind – an emulation. What precisely is the correct meaning of “all relevant aspects” is open to debate, and while the present paper will touch upon this issue, no pretention is made to settle it.

There are several reasons why we might want to upload. An uploaded mind will, at any given time, consist of a finite string of 0’s and 1’s, making long-distance travel as easy as the transfer of computer files. We will no longer be stuck with our fragile flesh-and-blood bodies, but can migrate between robot bodies at will, or into virtual worlds. Furthermore, uploading allows us to easily make backup copies of ourselves. A sufficient number of such backup copies of ourselves will not make us literally immortal (because, e.g., a civilization-scale catastrophe might destroy all of them), but it vastly improves the prospects for astronomically long lives, and will likely make us much less concerned about being killed in accidents or violent attacks, because all one loses from such an event is the memories acquired since last time one made a backup copy. An even more mind-blowing variant of this is the idea of making copies not meant for idle storage but for being up and running in parallel with the original.

These are some of the ideas that make uploading an attractive idea to many (but not all) transhumanists and futurologists. There is, however, a variety of further concerns regarding whether uploading technology can be successfully developed – and if so, whether it *should* be done. The purpose of the present paper is to review what I consider to be the key issues in this discussion, drawing heavily (but far from exclusively) on Sections 3.8 and 3.9 of Häggström (2016).

I'll begin in Section 2 with questions about what is needed for an *operationally successful* uploading technology, meaning that if I upload, the upload should correctly reproduce my behavior: it should look to any outside observer as if the upload is actually me. Is it reasonable to expect that such a technology is forthcoming? If yes, when? Can we even hope for *non-destructive* uploading, meaning that the procedure leaves my brain (and body) intact, or would we need to settle for *destructive* uploading?

Then, in Sections 3 and 4, which can jointly be viewed as the core of this paper, I'll discuss ways in which an operationally successful uploading technology might nevertheless be a *philosophical failure*, in the sense that although if I upload everything looks good from the outside, uploading nevertheless does not give *me* what I want. One way in which this might happen is if the upload fails to have consciousness (a possibility treated in Section 3). Another is that even if the upload has consciousness, it might fail to give me what I want by not being *me*, but merely someone else who shares my memories, my psychological traits, and so on (Section 4).

Among the various ethical concerns involved in whether or not we should move forward towards an uploading technology, I collect some of them in Section 5, except for what is perhaps the biggest one, namely what we can expect a society with a widely established uploading technology to be like, and whether such a society is worth wanting. That issue is treated separately in Section 6, based heavily on Hanson's (2016a) recent book *Age of Em,* which offers a generous range of predictions about what to expect from a society of uploads, or of *ems* (short for emulated minds) as he calls them. Finally, in Section 7, I wrap up with some concluding remarks.

2. Is it technically doable?

Timelines for the emergence of human- or superhuman-level artificial general intelligence (AGI) are so uncertain that epistemically reasonable predictions need to smear out their probability distributions pretty much all over the future time axis; see, e.g., Bostrom (2014), Grace et al. (2017) and Yudkowsky (2017). In contrast, we know enough about how complicated and involved whole brain emulation is to at least be able to say with confidence that a technology for mind uploading is not forthcoming in the next five or ten years (the only scenario I can think of that might reasonably falsify this prediction is if we suddenly discover how to create superintelligent AGI and this AGI decides to engage in developing mind uploading technology). While we might consider mind uploading to be the logical endpoint if we extrapolate two of today's largest and most prestigious ongoing research projects (the European Union's Human Brain Project and the White House BRAIN Initiative), neither of these has mind uploading among its explicit goals.

Kurzweil (2005) has done much to popularize the idea of mind uploading, and true to his faiblesse for unabashedly precise predictions of technological timelines, he states with confidence that uploading will have its breakthrough in the 2030s. This prediction is, however, based on a variety of highly uncertain assumptions. Sandberg and Bostrom (2008) are better at admitting these uncertainties, and their sober report still stands out as the main go-to place on the (future) technology of whole brain emulation. They systematically explore a wide range of possible scenarios, and break down much of their analysis into what they consider the three main ingredients in whole brain emulation, namely *scanning*, *translation* and *simulation*. Besides these, there is also the need to either give the emulation a robotic embodiment with suitable audiovisual and motoric input/output channels, or to embed it in some virtual reality environment, but this seems like a relatively easy task compared to the main three (Bostrom, 2014). Briefly, for what is involved in these three, we have the following summaries adapted from Häggström (2016).

- **Scanning** is the high-resolution microscopy needed to detect and register all the relevant details of the brain. While huge uncertainty remains as to the level of resolution needed for this, it seems likely that the necessary level has already been attained, since techniques for seeing individual atoms are becoming increasingly routine. Still, the speed and parallelization likely needed to avoid astronomical scanning times is not there yet. Sandberg and Bostrom (2008) focus mainly on techniques that involve cutting up the brain in thin slices that are scanned separately using one or other of the microscopy technologies that may be available (magnetic resonance imaging, sub-diffraction optics, X-ray, electron microscopy, and so on). This seems like a clear case of destructive scanning, so whatever uploading procedure it forms part of will obviously have to be destructive as well. Non-destructive scanning is even more challenging. A speculative suggestion, advocated by Kurzweil (2005) and building on the work of Freitas (1999), is to use nanobots that enter the brain in huge numbers, observe the microscopic structures in there and report back.
- **Translation** is the image analysis and other information processing needed to turn the scanned data into something that can be used as the initial state for the simulation. The amount needed seems to be huge; Sandberg and Bostrom (2008) list some of the stuff that needs to be done: "Cell membranes must be traced, synapses identified, neuron volumes segmented, distribution of synapses, organelles, cell types and other anatomical details (blood vessels, glia), identified".
- **Simulation** requires large amounts of hardware capability in terms of memory, bandwidth and CPU, presumably in a massively parallel architecture. Exactly how large these amounts need to be depends, of course, on the level of microscopic detail required for a satisfactory emulation. It may also depend on the amount of understanding we have of the higher-level workings of the brain; the less we have of that, the more critical is it to get the microscopic details exactly right, requiring more computer power. Whole brain emulation is often envisioned as a pure bottom-up (and brute force) project with no need for such higher-level understanding, in which case it will seem to us more or less as a black box (although no more than our brain already does). A more nuanced view of future technologies should allow for the possibility of significant amounts of top-down design, based on advances in the neuroscientific understanding of how our minds work at various levels above the microscopic details.

Another possibility, interlacing the three main ingredients listed above, is the piece-by-piece replacement of the brain by electronic devices maintaining the same functionality and interfacing with the rest of the brain as the replaced tissue; this idea was pioneered by Moravec (1988). Eventually all that remains is electronics, and the information stored can then be moved at will. As far as the end result is concerned, this should probably count as destructive uploading, but some thinkers intuit the gradualness of the procedure as a less scary form of destruction and are more willing to believe in the survival of personal identity (whatever that means; see Section 4).

While alternatives are not obviously impossible, it seems likely that the first uploading method to become technically available will be destructive, and based on scanning brain slices. As to time lines, Bostrom (2014) sticks to the (tentative) verdict of Sandberg and Bostrom (2008), which he summarizes as "the prerequisite capabilities might be available around mid-century, though with a large uncertainty interval".

3. Consciousness

Consider the case of destructive uploading, in which case my brain is destroyed, and replaced by a whole brain emulation on computer hardware. It might be that, no matter how well the technology discussed in the previous section is perfected, if I upload I will still not be around to experience the thoughts and the doings of the upload. It is hardly inconceivable that if my brain is frozen and cut up in slices, I simply die, no matter what is done to the information stored in my brain. This would make the prospect of uploading a lot less attractive than if I could look forward to a rich life, full of experiences, in my new existence as an upload. (Still, it might not *always* make it *entirely* unattractive, as for instance I might nevertheless choose to upload if I am about to die anyway in some incurable disease while still having some life projects that I would like to complete, such as proving the Riemann hypothesis or raising my children.)

One way in which destructive uploading would cut me off from future experiences is if the upload simply lacks consciousness. Perhaps it is just not possible to harbor consciousness on a digital computer. Whether or not this is the case is an open question, along with the more general issues of what consciousness really is, and how it arises in a physical universe. Philosophers of mind cannot be expected to settle these matters anytime soon; see, e.g., Chalmers (1996), McGinn (2004), Searle (2004) and Dennett (2017) for some of the different and partly conflicting views held by leading contemporary philosophers. Until this is sorted out, any discussion of whether uploads will be conscious or not will have to be speculative to a considerable extent. But the topic is important, so we should not let this deter us – perhaps we can make progress!

My personal inclination, while far from being an unshakable belief, is towards accepting uploads as conscious beings. My reason is based on the following thought experiment (Häggström, 2016). I judge most or all of my human friends to be conscious beings, including my friend Johan. Let us imagine him a few decades from now taking me by surprise by removing the top of his skull and demonstrating that it does not contain the expected jellyish object known as a brain, but is instead full of electronic computer hardware. Note (crucially for my argument) that if we accept my position, based on the work reviewed in Section 2, that mind uploading is in principle operationally feasible, then something like this is a fairly plausible scenario. Johan's display would not change my verdict that he is conscious, because my (current) conviction that he is conscious is based not on beliefs about his inner anatomy, but on his speech and behavior, which will be as irresistibly conscious-seeming in the thought experiment as it is today. It seems to me that this is how, in general, we judge others to be conscious. If those reasons are right, then we should grant consciousness also to uploads behaving like us, such as Johan in the thought experiment.

I admit that this argument is nowhere near a demonstration that uploads will be conscious, and perhaps it should not even count as a good reason. Perhaps it just means that our spontaneous intuitive reasons for judging others as conscious are no good. But if we stick to those reasons, then the consistent thing to do is to treat uploads the same (analogously to how already today we grant consciousness to, say, people whose gender or ethnicity differs from our own). To me, this line of thinking is further supported by a moral argument: we do not know who is conscious or who is not (heck, I don't even know that my friend Johan is conscious), but it seems to me that anyone who acts conscious is entitled to our benefit of the doubt. This is in line with Sandberg's (2014) Principle of Assuming the Most: "Assume that any emulated system could have the same mental properties as the original system and treat it accordingly". In contrast, Fukuyama (2002) simply denies that uploads can be conscious, and goes on to say (in a way that is clearly meant not only descriptively but also normatively) that "if we double-crossed [an upload] we would feel no guilt […] and if circumstances forced us to kill him […] we would feel no more regret than if we lost any other valuable asset, like a car or a teleporter”.

The main idea to be found in the literature is favor of uploads being conscious is the so-called *computational theory of mind* (CTOM). Roughly, it states that what matters for consciousness is not the material substance itself but its organization (“it ain’t the meat, it’s the motion”, as Sharvy, 1985, put it), and that the organization that produces consciousness is the right kind of information processing; what precisely “the right kind of” means is typically left unspecified for the time being. Of the philosophers mentioned above, Dennett is a proponent of CTOM, Chalmers looks agnostically but fairly favorably upon it, and Searle thinks (for reasons we will have a look at a few paragraphs down) it badly mistaken, while McGinn adheres to the so-called mysterian view that understanding the fundamentals of consciousness is not within reach of the human cognitive machinery in much the same way that a dog can never understand Fermat’s last theorem. Fukuyama (2002) rejects CTOM for no other reason than the (unsupported and probably false) statement that “it works only by denying the existence of what you and I and everyone else understand consciousness to be (that is, subjective feelings)”.

Arguments like the thought experiment about Johan above tilt me towards a favorable view of CTOM. But there are also arguments against it. Fairly typical is the argument by Pigliucci (2014) who, in his contribution to the collection by Blackford and Broderick (2014), repeatedly asserts that “consciousness is a biological phenomenon”, and complains that Chalmers (2014) in the same volume “proceeds *as if* we had a decent theory of consciousness, and by that I mean a decent *neurobiological theory*” (emphasis in the original). Since CTOM is not a neurobiological theory, it doesn’t pass Pigliucci’s muster and must therefore be wrong.

Or so the argument goes. As first explained in Häggström (2016), I don’t buy it. To expose the error in Pigliucci’s argument, I need to spell it out a bit more explicitly than he does. Pigliucci knows of exactly one conscious entity, namely himself, and he has some reasons to conjecture that most other humans are conscious as well, and furthermore that in all these cases the consciousness resides in the brain (at least to a large extent). Hence, since brains are neurobiological objects, consciousness must be a (neuro-)biological phenomenon. This is how I read Pigliucci’s argument. The problem with it is that brains have more in common than being neurobiological objects. For instance, they are also material objects, and they are computing devices. So rather than saying something like “brains are neurobiological objects, so a decent theory of consciousness is neurobiological”, Pigliucci could equally well say “brains are material objects, hence panpsychism”, or he could say “brains are computing devices, hence CTOM”, or he might even admit the uncertain nature of his attributions of consciousness to others and say “the only case of consciousness I know of is my own, hence solipsism”. So what is the right level of generality? Any serious discussion of the pros and cons of CTOM ought to start with the admission that this is an open question. By simply postulating from the outset what the right answer is to this question, Pigliucci short-circuits the discussion, and we see that his argument is not so much an argument as a naked claim.

There are somewhat better anti-CTOM arguments around than Pigliucci’s. The most famous one is *the Chinese room argument* of Searle (1980). The following summary and discussion of the argument is mostly taken from Häggström (2016). The argument is a *reduction ad absurdum*: assuming correctness of CTOM, Searle deduces the in-principle possibility a scenario that is so crazy that the assumption (CTOM) must be rejected. Here is how he reasons:

Assume CTOM. Then a computer program can be written that *really* thinks, that *really* understands, and that *really* is conscious, as opposed to merely giving the outwards appearance of doing so.

Suppose, for concreteness, that the program speaks and understands Chinese. (Since there are humans who can do this, CTOM implies that there are such computer programs as well.) Let us now implement this program in a slightly unusual way – as opposed to on an electronic computer. Instead of a computer, there is a room containing John Searle himself (who does not understand Chinese), plus a number of large stacks of paper. The first stack gives precise instructions in English (corresponding to the computer program) for what Searle (corresponding to the CPU) should do, while the other stacks play the role of computer memory. The room furthermore has two windows, one where people outside can provide strings of Chinese symbols as input to the system, and another where Searle delivers other such strings as output – strings that he produces by following the step-by-step instructions in the first stack. The same information processing is going on here as would have been the case in the computer, so if CTOM is true, then Searle understands Chinese, which is crazy, because all he knows is how to mindlessly follow those step-by-step instructions, so CTOM must be wrong.

There are several ways in which a CTOM proponent might respond to this attempt at a *reductio*. Personally, I am inclined to side with Hofstadter and Dennett (1981) in holding forth what Searle calls *the systems reply*, which is to say that in the proposed scenario, it is not Searle who understands Chinese, but the whole system, consisting of the room, plus the stacks of paper, plus Searle. Searle is just one of the components of the system, and does not understand Chinese any more than a single neuron can be found in my brain that understands the statement of Fermat's last theorem.

Searle (1980) does have a retort to the systems reply, namely that the thought experiment can be modified so as to get rid of the room. Instead, he himself memorizes the entire rulebook and carries out the step-by-step manipulations in his head. Again we have the same information processing going on, so CTOM implies that Searle understands Chinese, a consequence that the systems reply can no longer explain away, because this time the system consists of nothing but Searle. But still Searle does not understand Chinese – all he does is mindless step-by-step symbol manipulation – so CTOM must be wrong.

I recommend the discussion that followed in Hofstadter and Dennett (1981), Searle (1982a), Dennett (1982) and Searle (1982b) to those readers who enjoy a good fight, and as a fascinating example of the amount of heat and anger a philosophical discussion can generate. Hofstadter and Dennett suggest that Searle has failed to imagine the astounding amount of learning needed to "swallow up the description of another human being" and whether that can be done without thereby achieving an understanding of whatever it is that this other human being understands. With a particularly inflammatory phrase, they stress that "a key part of [Searle's] argument is in glossing over these questions of orders of magnitude". Searle (1982a, 1982b) replies, but offers no amendment to this alleged glossing-over, and no further clue as to exactly *why*, under these extreme circumstances, the individual will not understand Chinese. Maybe he will, maybe he won't.

When I try to imagine the outcome of this extreme thought experiment, the most natural interpretation that comes to mind is in terms of multiple personality disorder (Häggström, 2016). To us outside observers, it will look like we are faced with two persons inhabiting the same body: English-speaking-Searle who tells us (in English) that he does not understand a word of Chinese, and Chinese-speaking-Searle who insists (in Chinese) that he does understand Chinese. Why in the world should we trust only the former and not the latter?

Bergström (2016) will have none of this, and says that in this situation, "Searle knows Chinese only in the same sense that I 'know' how to play grandmaster-level chess, namely if I have a grandmaster next to me telling me which moves to make" (my translation). Here, when speaking of "Searle", Bergström obviously refers to English-speaking-Searle, whose perspective he is eager to entertain, while he entirely ignores that of Chinese-speaking-Searle. This asymmetry strikes me as terribly arbitrary. It's almost as if he is taking for granted that Chinese-speaking-Searle doesn't even *have* a perspective, a shocking violation of the benefit-of-the-doubt moral principle I suggested earlier in this section.

These examples from the literature only scratch the surface of what has been said about CTOM, but they serve to illustrate how little we know about the origins and role of consciousness in the physical universe. CTOM may come across as counterintuitive, something that the Chinese room argument does make some way towards showing, but also something that Schwitzgebel (2014) argues is bound to hold for *any* theory of consciousness – counterintuitive enough to warrant in his opinion the term *crazy*. And while CTOM may well even be wrong, it survives quite easily against currently available attempts to shoot it down, and it strikes me as sufficiently elegant and consistent to be a plausible candidate for a correct theory about how consciousness arises in the physical world. This is encouraging for those who hope for conscious uploads, but the jury is likely to remain out there for a long time still.

4. Personal identity

Even if it turns out uploads are conscious, I might still hesitate to undergo destructive uploading, because if the upload is not *me*, but merely a very precise copy of me (my personality traits, my memories, and so on), then destructive uploading implies that I die. So will the upload be me? This is the problem of personal identity. Throughout most of this section, I will (mostly for the sake of simplicity) discuss personal identity in the context of *teleportation* rather than uploading; the cases are very similar and the arguments can be easily transferred from one case to the other.

Teleportation is a hypothetical (and controversial) future means of transportation. An individual's body (including the brain) is scanned, and the information thus obtained is sent to the destination, where the body is reassembled – or, if you prefer to view it that way, a new identical body is assembled. (Whether this distinction is substantial or merely terminological is part of the problem of personal identity.) It is an open problem whether it will ever be doable, but we will assume for the sake of the argument that we have a teleportation technology good enough so that the reassembled individual, including his or her behavior, is indistinguishable from the original. As with uploading, we distinguish between *destructive teleportation* (where the scanning procedure involves the destruction of the original body) and *non-destructive teleportation* (where the original body remains intact).

Teleportation has been commonplace in science fiction for more than half a century, with occasional appearances before that, such as in Richard Wagner's 1874 opera *Der Ring des Nibelungen*. Amongst pioneering philosophical treatments of the issue of whether personal identity survives destructive teleportation we find Lem (1957) and Parfit (1984). Resolving the issue empirically seems difficult (as anticipated already by Lem) and may perhaps even be impossible: even in a world where such a technology is widely established, it does not help to ask people who have undergone teleportation whether or not they are the same person as before they teleported, because no matter what the true answer is, they will all *feel* as if they are the same person as before, and respond accordingly (provided they are honest). Faced with these bleak prospects for resolving the issue empirically, one way out, in the logical positivist tradition, is to rule it out as ill-posed or meaningless. Let us still have a go at a philosophical analysis. The following formalism, tentatively distinguishing two kinds of survival, is taken from Häggström (2016).

Imagine that today is Tuesday, I am in Gothenburg, but need to attend a meeting in New York on Thursday. Tomorrow (Wednesday) is travel day, and I need to choose between (destructive) teleportation and old-fashioned air travel. The former is far more convenient, but if it means that I die then I prefer to go by air. Thus: if I take the teleporter, will I survive until Thursday? That may depend on the exact meaning of survival, and as a preliminary step towards sorting out this thorny issue, let us distinguish *σurvival* from *Σurvival*, as follows:

> **σurvival**: I σurvive until Thursday if, on Thursday, there exists a person who has the same personality traits and memories and so on, as I have today, allowing for some wiggling corresponding to the amount of change we expect a person to go through over the course of a few days.
>
> **Σurvival**: I Σurvive until Thursay if (a) I σurvive until Thursday, and (b) the situation satisfies whatever extra condition is needed for the guy in New York to really be me.

It is clear that I will σurvive the teleportation, but will I also Σurvive? That is unclear, due to the definition of Σurvival being incomplete, or one might even say fuzzy – what exactly *is* that condition in (b)? The fuzziness is intentional, because I honestly do not have any good idea for what that property might be. In my view, any proposal for a nonvacuous condition in (b) needs to be backed up by a good argument, because otherwise Occam's razor (Baker, 2010) should compel us to simply drop the condition, and conclude that Σurvival coincides with σurvival – in which case teleportation is safe.

We may find a hint about what the condition might be in what is perhaps the most common objection to the idea of survival of personal identity under destructive teleportation, namely the comparison with non-destructive teleportation summarized by Yudkowsky (2008) as follows:

> Ah, but suppose an improved Scanner were invented, which scanned you non-destructively, but still transmitted the same information to Mars. Now, *clearly*, in this case, *you, the original* have simply stayed on Earth, and the person on Mars is *only a copy*. Therefore [teleportation without the improved scanner] is actually murder and birth, not *travel* at all – it destroys the original, and constructs a copy! [Italics in the original]

While Yudkowsky's purpose here (just like mine) is to state the argument in order to shoot it down, I don't think his eloquent summary of it (including his sarcastic "*clearly*") is at all unfair. Chalmers (2014) and Pigliucci (2014) are among the many writers who consider it; Chalmers attaches some weight to it but considers it inconclusive, while Pigliucci regards it as conclusive and claims that "if it is possible to do the transporting or uploading in a non-destructive manner, *obviously* we are talking about duplication, not preservation of identity" (italics in the original). Note the convenient ambiguity of Pigliucci's "*obviously*" (echoing Yudkowsky's "*clearly*"), and that the term is warranted if it refers to "duplication" but not at all so if it refers to "preservation of identity". If σurvival is all there is to survival, then there is no contradiction in surviving in two bodies.

Pigliucci fails to spell out his argument, but it seems safe to assume that he (along with other proponents of the destructive vs non-destructive comparison argument) does think there is some nonvacuous condition in (b). Since a major asymmetry between the two bodies in the non-destructive teleportation case is that only one of them has a continuous space-time trajectory from the original (pre-teleportation) body, my best guess at what condition he implicitly refers to is CBTST, short for *continuity of my body's trajectory through space-time* (Häggström, 2016). If CBTST is the property in (b), then clearly I will *not* Σurvive the teleportation to New York. But is CBTST really crucial to personal identity? Perhaps if preservation of personal identity over time requires that I consist of the same elementary particles as before, but this seems not to be supported by fundamental physics, according to which elementary particles do not have identities, so that claims like “these electrons are from the original body, while those over there are not” simply do not make sense (Bach, 1988, 1997). Since our bodies consist of elementary particles, postulating the corresponding asymmetry between the two bodies coming out of the non-destructive teleportation procedure is just as nonsensical. And even if physics did admit elementary particles with individual identities, the idea of tying one’s personal identity to a particular collection of elementary particles seem indefensible in view of the relentless flux of matter on higher levels. Says Kurzweil (2005):

> The specific set of particles that my body and brain comprise are in fact completely different from the atoms and molecules that I comprised only a short while ago. We know that most of our cells are turned over in a matter of weeks, and even our neurons, which persist as distinct cells for a relatively long time, nonetheless change all of their constituent molecules within a month. [...] The half-life of a microtubule is about ten minutes. [...]
>
> So I am a completely different set of stuff than I was a month ago, and all that persists is the pattern of organization of that stuff. The pattern changes also, but slowly and in a continuum. I am rather like the pattern that water makes in a stream as it rushes past the rocks in its path. The actual molecules of water change every millisecond, but the pattern persists for hours or even years.

Proponents of CBTST – or of any other nonempty condition – as being crucial to (b) need to spell out good arguments for why this would be the case. Until they do, the reasonable stance seems to be that survival is simply σurvival, in which case teleportation is safe. The view that survival is just σurvival may seem counterintuitive or even appalling – why should I care as much as I do about the person tomorrow who claims to be me if all that connects his personal identity to mine is mere σurvival? See Blackmore (2012) for a beautiful attempt to come to grips with this in our day-to-day existence; also Parfit (1984) arrives at a similar position.

In order to connect this teleportation discussion back to the issue of uploading, assume now that we accept that destructive teleportation is safe in the sense of preserving personal identity, and assume furthermore that after careful reflection of the issues discussed in Section 3 we also accept that uploads are conscious. Should we then conclude that we would survive destructive uploading? While the two assumptions do offer some support in this direction, the conclusion does not quite follow, because condition (b) for Σurvival might for instance be that (the physical manifestation of) the person on Thursday is made of the same kind of substance as the one on Tuesday. The desired conclusion does however seem to follow if we strengthen the survivial-under-teleportation assumption to Σurvival=σurvival.

5. Ethical concerns

With the development of unloading technology, we will encounter difficult ethical problems, and the corresponding legal ones. For instance, Bostrom (2014) notes that, just like with any other technology, “before we would get things to work perfectly, we would probably get things to work imperfectly”. This might well amount to creating a person with (an upload analogue of) severe brain damage resulting in horrible suffering. Would it be morally permissible to do so? Metzinger (2003) says no:

> What would you say if someone came along and said, “Hey, we want to genetically engineer mentally retarded human infants! For reasons of scientific progress we need infants with certain cognitive and emotional deficits in order to study their postnatal psychological development.” […] You would certainly think this was not only an absurd and appalling but also a dangerous idea. It would hopefully not pass any ethics committee in the democratic world. However, what today's ethics committees *don’t* see is how the first machines satisfying a minimally sufficient set of constraints for conscious experience could be just *like* such mentally retarded infants.

Obviously, the weight of this argument hinges to a large extent on the wide-open question of whether or not uploads are conscious, but here I do think Sandberg’s (2014) Principle of Assuming the Most, discussed in Section 3 and serving as a kind of precautionary principle, should guide our decisions. And even if (contrary to our expectations) reliable arguments that computer consciousness is impossible would be established, we need not end up in Fukuyama’s (2002) position that we lack moral reason to treat uploads well. We could for instance reason analogously to how Kant does about animals: he does not grant them status as beings to whom we can have moral duties, but claims nevertheless about man that “if he is not to stifle his human feelings, he must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men” (Kant 1784-5, quoted in Gruen, 2017). This kind of argument already appears in the current debate over permissible human treatment of sexbots (Danaher, 2017).

The concerns about malfunctioning and suffering uploads raised by Bostrom and Metzinger above can be mitigated to some extent – but hardly eliminated – by experimenting with whole brain emulations of animals ranging from nematode worms and fruit flies to mice and primates, before moving on to humans. This seems likely to happen. Still, assuming (plausibly) that destructive uploading comes first, the first humans to upload will be taking an enormous risk. While it would probably not be hard to find heroic volunteers willing to risk their lives for a scientific and technological breakthrough, the project is likely to encounter legal obstacles. We can probably not expect the law to grant personhood to uploads before the technology has already been established, whence at such a pioneering stage even an operationally successful upload will count legally as assisted suicide. Sandberg (2014) points out that “suicide is increasingly accepted as a way of escaping pain, but suicide for science is not regarded as an acceptable reason”, and cites the Nuremberg code as support for the last statement.

Sandberg then goes on to suggest that the following might be way around this legal obstacle. Consider cryonics: the low-temperature preservation of legally dead people who hope to be restored to life in a future with far more powerful medical procedures than today’s. The fact that cryonics is simultaneously

(a) legal, and

(b) a plausible hope for eventual successful awakening,

is based on the (conjectured) existence of a gap between legal death and so-called information-theoretic death, defined as a state where a brain is so ramshackle that the person (his personality, memories and so on) can no longer be reconstructed even in principle: the necessary information is simply gone (Merkle, 1992). If part of this gap can be exploited for uploading purposes by doing the destructive brain scanning post-mortem, then the legal (as well as the ethical) situation may be less problematic than the case of doing it to a living person.

Further ethical obstacles to uploading in the early stages of the technology concern the situation of an upload at that time, with no legal status as a person but stuck in a "legal limbo" (Sandberg, 2014). At a later time, when uploading is more widespread, along with copying of uploads, legal and ethical concerns multiply. If I survive as two uploads, will just one of them be bound by contracts I've signed pre-uploading, or both, or neither? Who inherits my marriage, and who inherits my belongings? As to the latter, one might suggest that the uploads share them equally, but then what about inactive backup copies? Sandberg mentions all of these concerns (and others, involving vulnerability and privacy), as well as a challenge to our democratic system: surely uploads ought to count as full citizens, but then what if on election day I create 1000 copies of myself – does that give me 1000 votes? This brings us to the subject of the next section: what can we expect a society with widespread uploading technology to be like?

6. A society of uploads

A question of huge import to the issue of whether or not we ought to move forward towards an operationally successful uploading technology is what impact it is likely to have on society: what can we expect a society with uploading technology to be like, and would that be preferable to how society is likely to develop without uploading? Here factual and normative issues are closely entangled, but in order to try to separate out the normative aspect, one may ask what kind of future society we want. A common answer is a society in which we have good lives, where "good" may for instance mean "happy" or "meaningful" (or some combination thereof) – terms that call for further definition. This quickly gets thorny, but there is a more concrete and well-defined wish that is at the focus in the growing research area of existential risk studies, namely the avoidance of extinction of humanity (Bostrom, 2013; Häggström, 2016; Torres, 2017). To be able to lead happy or meaningful lives, we first of all need to avoid going extinct.

Sandberg (2014) mentions several ways in which a mind uploading technology may help reducing the risk for human extinction (provided that we decide that uploads qualify as human beings). One is that uploads are better suited for space travel than are biological humans, and so they facilitate space colonization in order not to "carry all eggs in one planetary basket". Another one is that if biological humans and uploads live side by side, a biological pandemic might wipe out the biological humans and a computer virus might wipe out the uploads, but it would take both events to exterminate humanity altogether.

Not going extinct is however far from a good enough specification of what might go into a desirable future for humanity. It has been suggested that, in the presence of suffering, survival could in fact be *worse* than extinction, and this has recently led to an increased attention to a possible tension between actions meant to reduce the risk for human extinction and those meant to reduce the risk for creating astronomical amounts of suffering; see Althaus and Gloor (2016).

Hanson, in his *Age of Em: Work, Love and Life when Robots Rule the Earth* (2016a), which is the (so far) unchallenged masterpiece in the niche of non-fiction accounts of futures societies with uploads, summarizes more than two decades of his thinking on the topic. He avoids the normative issue almost entirely, settling instead for unsentimental predictions based on ideas from economics, evolutionary biology, physics and other fields. The main assumption he postulates is concerned with is that uploading technology becomes feasible before the creation of superhuman AGI, and before we attain much understanding of how thoughts and the brain's other high-level processes supervene on the lower-level processes that are simulated. The second part of the assumption prevents us from boosting the intelligence of the uploads to superhuman levels, other than in terms of speed by transferring them to faster hardware. The speedup becomes possible by what is perhaps Hanson's second most central assumption, namely that hardware will continue to become faster and cheaper far beyond today's level. This assumption also opens up the possibility to increase population size to trillions or more.

Hanson describes a society of uploads in fascinating detail on topics ranging from city planning, gender imbalance, management practices, surveillance and military operations to loyalty, status, leisure, swearing and sex. Changes compared to the way we live today are predicted to be in many ways large and disruptive.

Central to Hanson's scenario is the transformation of labor markets. The ease of copying uploads in combination with lowered hardware costs will push down wages towards subsistence levels: why would an employer pay you $100,000 per year for your work when an upload can do it at a hardware cost of $1 per year? Today's society in large parts of the world has been able to escape the Malthusian trap (for the time being), because ever since the industrial revolution we have maintained an innovation rate that has allowed the economy to grow much faster than the population. Our current evolutionarily maladaptive reproductive behavior contributes to this gap, which Hanson notes is historically an exceptional situation. To emphasize its unsustainability (a view akin to the even darker but disturbingly convincing analysis by Alexander, 2014), he calls our present era *dreamtime*. While a society of uploads will innovate even faster than ours, the ease with which uploads admit copying is likely to cause a population increase that the innovation rate cannot match.

Among the many exotica in Hanson's future is that uploads will be able run on different hardware and thus at different speeds. This opens up the possibility to rationalize a company by removing some level of middle management and instead letting a single manager run at a thousand times the speed of the lower-level workers, giving him time to manage thousands of them individually; the lower-level workers can switch hardware and speed up in their one-on-one meetings with the manager. Adjustable speed has the further advantage of making it easier to meet project deadlines, and so on.

An even more alien (to the way we live our lives today) aspect of the world painted by Hanson is the idea that most work will be done by so-called *spurs*, copied from a template upload in order to work for a few hours or so and then be terminated. Many readers have imagined the predicament of being a spur unbearable and asked why this would not trigger a revolution. Hanson (2016c) explains, however, that the situation of the spur is not much different from that of a person at a party knowing that he has drunk so much that by next morning his memories of the party will be permanently lost to amnesia. This strikes me as correct, at least if spurs accept the Σurvival=σurvival view discussed in Section 3 – which they seem likely to do, as they feel psychologically contiguous with the template up to the time of copying.

More generally, readers of Hanson (2016a) tend to react to his future scenarios by describing them as dystopian. Hanson consistently rejects this label, maintaining that while the life of the uploads in his world may appear awful *to us*, it will actually be pretty good *to them*. Unfortunately, in Häggström (2016), I gave a misleading impression of Hanson's reasons for this, by suggesting that uploads "can be made to enjoy themselves, e.g., by cheap artificial stimulation of their pleasure centra". Such wireheading does not appear in his scenarios, and in Hanson (2016b) he corrects the misunderstanding, and gives his true reasons:

> Most ems work and leisure in virtual worlds of spectacular quality, and […] ems need never experience hunger, disease, or intense pain, nor ever see, hear, feel, or taste grime or anything ugly or disgusting. Yes they'd work most of the time but their jobs would be mentally challenging, they'd be selected for being very good at their jobs,

> and people can find deep fulfillment in such modes. We are very culturally plastic, and em culture would promote finding value and fulfillment in typical em lives.

In his thoughtful review of *Age of Em*, Alexander (2016) criticizes the absence of wireheading and the relative sparsity of (the digital upload equivalent of) related pharmacological brain manipulation. Such absence can of course be taken as part of Hanson's central assumption that brain emulations are essentially black boxes so that we do not have the understanding required to usefully manipulate the inner workings of the emulations, but Alexander holds this assumption to be unrealistic if taken too far. Already today we can use amphetamine-based stimulants to significantly boost our ability to focus, and their main drawbacks – addictiveness and health concerns – become trivial in a world where, in Alexander's words, "minds have no bodies, and where any mind that gets too screwed up can be reloaded from a backup copy", whence

> from employers' point of view there's no reason not to have all workers dosed with superior year 2100 versions of Adderall at all times. I worry that not only will workers not have any leisure time, but they'll be neurologically incapable of having their minds drift off while on the job.

This leads Alexander to envision a world in which the economy as a whole keeps rolling with relentless momentum, but whose inhabitants turn into mindless puppets. Hanson (2016d) replies that the employers' incentive to uses stimulants to boost their workforce may be tempered by the insight that "wandering minds may take away from the current immediate task, but they help one to search for hidden problems and opportunities", but the extent to which this fends off Alexander's dystopic logical conclusion seems to me uncertain.

Note finally that while this section has focused on a future with uploading technology, we also need to know something about the likely long-term outcome of a future *without* such technology in order to compare and to determine whether the technology is desirable. The latter kind of future falls outside the scope of the present paper, and I will just note that a key issue may be to figure out whether in such a future we have a way of avoiding the Moloch described by Alexander (2014).

7. Conclusion

I argued in Section 2 that mind uploading is likely doable eventually if we put sufficiently many decades of work into it. But should we do it? Sections 3-6 have featured a number of concerns that might cast doubt on that.

As to the concerns in Section 3 and 4 that uploading might be a philosophical dead-end, either by uploads not being conscious, or by personal identity not surviving the uploading procedure, there seems to be a reasonable case for cautious optimism. Furthermore, regardless of what the true answer to these difficult questions are, one plausible scenario is that if uploading ever becomes a wide-spread technology, the philosophical concerns will quickly be forgotten. Uploads will consistently testify to us (and tell themselves) that they are conscious and that they are the same person as pre-uploading, and when these testimonials have become so commonplace as to seem superfluous, the philosophical concerns about uploading will perhaps eventually appear as silly as asking "Will I cease to be conscious if I have a cup of coffee?" or "Does taking a nap cause a person to be replaced by a different person who just happens to have identical personality and memories?".

Yet, even if these philosophical questions are settled in favor of survival under uploading (or if they are deemed ill-posed and therefore discarded), there still remains the difficult bioethics-like concerns discussed in Section 5, as well as the question discussed in Section 6 on whether uploading technology is likely to lead to a society worth wanting. Whatever knowledge we have on these issues at present is highly tentative, and it is important that we put serious efforts into resolving them, rather than merely pressing ahead blindly with technology development. The future of humanity (or posthumanity) may be at stake.

Acknowledgement. I am grateful to Dorna Behdadi and Björn Bengtsson for helpful discussions.

References

Alexander, S. (2014) Meditations on Moloch, *Slate Star Codex*, July 30.

Alexander, S. (2016) Book review: Age of Em, *Slate Star Codex*, May 28.

Althaus, D. and Gloor, L. (2016) *Reducing Risks of Astronomical Suffering: A Neglected Priority*, Foundational Research Institute, Berlin.

Bach, A. (1988) The concept of indistinguishable particles in classical and quantum physics, *Foundations of Physics* **18**, 639-649.

Bach, A. (1997) *Indistinguishable Classical Particles*, Springer, New York.

Bergström, L. (2016) Recension av Here Be Dragons, *Filosofisk Tidskrift*, no. 4, 39-48.

Blackford, R. and Broderick, D. (2014) *Intelligence Unbound: The Future of Uploads and Machine Minds*, Wiley Blackwell, Chichester.

Blackmore, S. (2012) She won't be me, *Journal of Consciousness Studies* **19**, 16-19.

Bostrom, N. (2013) Existential risk prevention as global priority, *Global Policy* **4**, 15-31.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Bostrom, N. and Savulescu, J. (2009) *Human Enhancement*, Oxford University Press, Oxford.

Chalmers, D. (1996) *The Conscious Mind*, Oxford University Press, Oxford.

Chalmers, D. (2014) Uploading: a philosophical analysis, in Blackford and Broderick (2014), pp 102-118.

Danaher, J. (2017) Robotic rape and robotic child sexual abuse: should they be criminalised?, *Criminal Law and Philosophy* **11**, 71-95.

Dennett, D. (1982) The myth of the computer: an exchange, *New York Review of Books*, June 24.

Dennett, D. (2017) *From Bacteria to Bach and Back: The Evolution of Minds*, W.W. Norton & Co, New York.

Freitas, R. (1999) *Nanomedicine, Vol. I: Basic Capabilities,* Landes Bioscience, Georgetown, TX.

Fukuyama, F. (2002) *Our Posthuman Future: Consequences of the Biotechnology Revolution*, Farrar, Straus and Giroux, New York.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2017) When will AI exceed human performance? Evidence from AI experts, *arXiv*:1705.08807.

Gruen, L. (2017) The moral status of animals, *The Stanford Encyclopedia of Philosophy,* Fall 2017 Edition, (ed. E. Zalta).

Häggström, O. (2016) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

Hansell, G. and Grassie, W. (2011) *H+/-: Transhumanism and Its Critics*, Metanexus Institute, Philadelphia.

Hanson, R. (2016a) *Age of Em: Work, Love and Life when Robots Rule the Earth*, Oxford University Press, Oxford.

Hanson, R. (2016b) Here be dragons, *Overcoming Bias*, January 16.

Hanson, R. (2016c) Is forgotten party death? *Overcoming Bias*, April 23.

Hanson, R. (2016d) Alexander on *Age of Em*, May 30.

Hofstadter, D. and Dennett, D. (1981) *The Mind's I*, Basic Books, New York.

Kant, I. (1784-5) Moral Philosophy: Collin's Lecture Notes, in *Lectures on Ethics*, (Cambridge Edition of the Works of Immanuel Kant) , P. Heath and J.B. Schneewind (ed. and trans.), Cambridge University Press, Cambridge, 1997, pp. 37-222. Original is *Anthropologie in pragmatischer Hinsicht*, published in the standard *Akademie der Wissenschaften* edition, vol. 27.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*, Viking, New York.

Lem, S. (1957) *Dialogi*, Wydawnictwo Literackie, Krakow. English translation by Frank Prengel at LEM.PL, http://english.lem.pl/index.php/works/essays/dialogs/106-a-look-inside-dialogs

McGinn, C. (2004) *Consciousness and its Objects*, Clarendon, Oxford.

Merkle, R. (1992) The technical feasibility of cryonics, *Medical Hypotheses* **39**, 6-16.

Moravec, H. (1988) *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, Cambridge, MA.

Parfit, D. (1984) *Reasons and Persons*, Oxford University Press, Oxford.

Pigliucci, M. (2014) Uploading: a philosophical counter-analysis, in Blackford and Broderick (2014), pp 119-130.

Sandberg, A. (2014) Ethics of brain emulations, *Journal of Experimental & Theoretical Artificial Intelligence* **26**, 439-457.

Sandberg, A. and Bostrom, N. (2008) Whole brain emulation: a roadmap, Future of Humanity Institute technical report \#2008-3.

Schwitzgebel, E. (2014) The crazyist metaphysics of mind, *Australian Journal of Philosophy* **92**, 665-682.

Searle, J. (1980) Minds, brains and programs, *Behavioral and Brain Sciences* **3**, 417-457.

Searle, J. (1982a) The myth of the computer, *New York Review of Books*, April 29.

Searle, J. (1982b) The myth of the computer: an exchange (reply to Dennett), *New York Review of Books*, June 24.

Searle, J. (2004) *Mind: A Brief Introduction*, Oxford University Press, Oxford.

Sharvy, R. (1985) It ain't the meat it's the motion, *Inquiry* **26**, 125-134.

Torres, P. (2017) *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*, Pitchstone Publishing, Durham, NC.

Yudkowsky, E. (2008) Timeless identity, *Less Wrong*, June 3.

Yudkowsky, E. (2017) *There's No Fire Alarm for Artificial General Intelligence*, Machine Intelligence Research Institute, Berkeley, CA.