David J. Chalmers

The Singularity

A Philosophical Analysis

1. Introduction

What happens when machines become more intelligent than humans? One view is that this event will be followed by an explosion to ever-greater levels of intelligence, as each generation of machines creates more intelligent machines in turn. This intelligence explosion is now often known as the 'singularity'.[1]

The basic argument here was set out by the statistician I.J. Good in his 1965 article 'Speculations Concerning the First Ultraintelligent Machine':

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

The key idea is that a machine that is more intelligent than humans will be better than humans at designing machines. So it will be capable of designing a machine more intelligent than the most intelligent machine that humans can design. So if it is itself designed by humans,

Correspondence:
Email: chalmers@anu.edu.au

[1] I first became interested in this cluster of ideas as a student, before first hearing explicitly of the 'singularity' in 1997. I was spurred to think further about these issues by an invitation to speak at the 2009 Singularity Summit in New York City. I thank many people at that event for discussion, as well as many at later talks and discussions at West Point, CUNY, NYU, Delhi, ANU, Tucson, Oxford, and UNSW. Thanks also to Doug Hofstadter, Marcus Hutter, Ole Koksvik, Drew McDermott, Carl Shulman, and Michael Vassar for comments on this paper.

Journal of Consciousness Studies, **17**, No. 9–10, 2010, pp. 7–65

it will be capable of designing a machine more intelligent than itself. By similar reasoning, this next machine will also be capable of designing a machine more intelligent than itself. If every machine in turn does what it is capable of, we should expect a sequence of ever more intelligent machines.[2]

This intelligence explosion is sometimes combined with another idea, which we might call the 'speed explosion'. The argument for a speed explosion starts from the familiar observation that computer processing speed doubles at regular intervals. Suppose that speed doubles every two years and will do so indefinitely. Now suppose that we have human-level artificial intelligence designing new processors. Then faster processing will lead to faster designers and an ever-faster design cycle, leading to a limit point soon afterwards.

The argument for a speed explosion was set out by the artificial intelligence researcher Ray Solomonoff in his 1985 article 'The Time Scale of Artificial Intelligence'.[3] Eliezer Yudkowsky gives a succinct version of the argument in his 1996 article 'Staring at the Singularity':

> Computing speed doubles every two subjective years of work. Two years after Artificial Intelligences reach human equivalence, their speed doubles. One year later, their speed doubles again. Six months — three months — 1.5 months … Singularity.

The intelligence explosion and the speed explosion are logically independent of each other. In principle there could be an intelligence explosion without a speed explosion and a speed explosion without an intelligence explosion. But the two ideas work particularly well together. Suppose that within two subjective years, a greater-than-human machine can produce another machine that is not only twice as fast but 10% more intelligent, and suppose that this principle is indefinitely extensible. Then within four objective years there will have been an infinite number of generations, with both speed and intelligence increasing beyond any finite level within a finite time. This process would truly deserve the name 'singularity'.

Of course the laws of physics impose limitations here. If the currently accepted laws of relativity and quantum mechanics are correct — or even if energy is finite in a classical universe — then we cannot expect the principles above to be indefinitely extensible. But even with these physical limitations in place, the arguments give some

[2] Scenarios of this sort have antecedents to the argument in science fiction, perhaps most notably in John Campbell's 1932 short story 'The Last Evolution'.

[3] Solomonoff also discusses the effects of what we might call the 'population explosion': a rapidly increasing population of artificial AI researchers.

reason to think that both speed and intelligence might be pushed to the limits of what is physically possible. And on the face of it, it is unlikely that human processing is even close to the limits of what is physically possible. So the arguments suggest that both speed and intelligence might be pushed far beyond human capacity in a relatively short time. This process might not qualify as a 'singularity' in the strict sense from mathematics and physics, but it would be similar enough that the name is not altogether inappropriate.

The term 'singularity' was introduced by the science fiction writer Vernor Vinge in a 1983 opinion article.[4] It was brought into wider circulation by Vinge's influential 1993 article 'The Coming Technological Singularity', and by the inventor and futurist Ray Kurzweil's popular 2005 book *The Singularity is Near.* In practice, the term is used in a number of different ways. A loose sense refers to phenomena whereby ever-more-rapid technological change leads to unpredictable consequences.[5] A very strict sense refers to a point where speed and intelligence go to infinity, as in the hypothetical speed/intelligence explosion above. Perhaps the core sense of the term, though, is a moderate sense in which it refers to an intelligence explosion through the recursive mechanism set out by I.J. Good, whether or not this intelligence explosion goes along with a speed explosion or with divergence to infinity. I will always use the term 'singularity' in this core sense in what follows.

One might think that the singularity would be of great interest to academic philosophers, cognitive scientists, and artificial intelligence researchers. In practice, this has not been the case.[6] Good was an eminent academic, but his article was largely unappreciated at the time. The subsequent discussion of the singularity has largely taken place in nonacademic circles, including Internet forums, popular media and

[4] As Vinge (1993) notes, Stanislaw Ulam (1958) describes a conversation with John von Neumann in which the term is used in a related way: 'One conversation centered on the ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue.'

[5] A useful taxonomy of uses of 'singularity' is set out by Yudkowsky (2007). He distinguishes an 'accelerating change' school, associated with Kurzweil, an 'event horizon' school, associated with Vinge, and an 'intelligence explosion' school, associated with Good. Smart (1999–2008) gives a detailed history of associated ideas, focusing especially on accelerating change.

[6] With some exceptions: discussions by academics include Bostrom (1998; 2003), Hanson (2008), Hofstadter (2005), and Moravec (1988; 1998). Hofstadter organized symposia on the prospect of superintelligent machines at Indiana University in 1999 and at Stanford University in 2000, and more recently, Bostrom's Future of Humanity Institute at the University of Oxford has organized a number of relevant activities.

books, and workshops organized by the independent Singularity Institute. Perhaps the highly speculative flavour of the singularity idea has been responsible for academic resistance to it.

I think this resistance is a shame, as the singularity idea is clearly an important one. The argument for a singularity is one that we should take seriously. And the questions surrounding the singularity are of enormous practical and philosophical concern.

Practically: If there is a singularity, it will be one of the most important events in the history of the planet. An intelligence explosion has enormous potential benefits: a cure for all known diseases, an end to poverty, extraordinary scientific advances, and much more. It also has enormous potential dangers: an end to the human race, an arms race of warring machines, the power to destroy the planet. So if there is even a small chance that there will be a singularity, we would do well to think about what forms it might take and whether there is anything we can do to influence the outcomes in a positive direction.

Philosophically: The singularity raises many important philosophical questions. The basic argument for an intelligence explosion is philosophically interesting in itself, and forces us to think hard about the nature of intelligence and about the mental capacities of artificial machines. The potential consequences of an intelligence explosion force us to think hard about values and morality and about consciousness and personal identity. In effect, the singularity brings up some of the hardest traditional questions in philosophy and raises some new philosophical questions as well.

Furthermore, the philosophical and practical questions intersect. To determine whether there might be an intelligence explosion, we need to better understand what intelligence is and whether machines might have it. To determine whether an intelligence explosion will be a good or a bad thing, we need to think about the relationship between intelligence and value. To determine whether we can play a significant role in a post-singularity world, we need to know whether human identity can survive the enhancing of our cognitive systems, perhaps through uploading onto new technology.These are life-or-death questions that may confront us in coming decades or centuries. To have any hope of answering them, we need to think clearly about the philosophical issues.

In what follows, I address some of these philosophical and practical questions. I start with the argument for a singularity: is there good reason to believe that there will be an intelligence explosion? Next, I consider how to negotiate the singularity: if it is possible that there will be a singularity, how can we maximize the chances of a good outcome? Finally, I consider the place of humans in a post-singularity world, with special attention to questions about uploading: can an uploaded human be conscious, and will uploading preserve personal identity?

My discussion will necessarily be speculative, but I think it is possible to reason about speculative outcomes with at least a modicum of rigour. For example, by formalizing arguments for a speculative thesis with premises and conclusions, one can see just what opponents need to deny in order to deny the thesis, and one can then assess the costs of doing so. I will not try to give knockdown arguments in this paper, and I will not try to give final and definitive answers to the questions above, but I hope to encourage others to think about these issues further.[7]

2. The Argument for a Singularity

To analyse the argument for a singularity in a more rigorous form, it is helpful to introduce some terminology. Let us say that AI is artificial intelligence of human level or greater (that is, at least as intelligent as an average human). Let us say that AI+ is artificial intelligence of greater than human level (that is, more intelligent than the most intelligent human). Let us say that AI++ (or superintelligence) is AI of far greater than human level (say, at least as far beyond the most intelligent human as the most intelligent human is beyond a mouse).[8] Then we can put the argument for an intelligence explosion as follows:

1. There will be AI+.
2. If there is AI+, there will be AI++.

———————

3. There will be AI++.

Here, premise 1 needs independent support (on which more soon), but is often taken to be plausible. Premise 2 is the key claim of the intelligence explosion, and is supported by Good's reasoning set out above. The conclusion says that there will be superintelligence.

[7] The main themes in this article have been discussed many times before by others, especially in the nonacademic circles mentioned earlier. My main aims in writing the article are to subject some of these themes (especially the claim that there will be an intelligence explosion and claims about uploading) to a philosophical analysis, with the aim of exploring and perhaps strengthening the foundations on which these ideas rest, and also to help bring these themes to the attention of philosophers and scientists.

[8] Following common practice, I use 'AI' and relatives as a general term ('An AI exists'), an adjective ('An AI system exists'), and as a mass term ('AI exists').

The argument depends on the assumption that there is such a thing as intelligence and that it can be compared between systems: otherwise the notion of an AI+ and an AI++ does not even make sense. Of course these assumptions might be questioned. Someone might hold that there is no single property that deserves to be called 'intelligence', or that the relevant properties cannot be measured and compared. For now, however, I will proceed with under the simplifying assumption that there is an intelligence measure that assigns an intelligence value to arbitrary systems. Later I will consider the question of how one might formulate the argument without this assumption. I will also assume that intelligence and speed are conceptually independent, so that increases in speed with no other relevant changes do not count as increases in intelligence.

We can refine the argument a little by breaking the support for premise 1 into two steps. We can also add qualifications about timeframe, and about potential defeaters for the singularity.

1. There will be AI (before long, absent defeaters).
2. If there is AI, there will be AI+ (soon after, absent defeaters).
3. If there is AI+, there will be AI++ (soon after, absent defeaters).

———————

4. There will be AI++ (before too long, absent defeaters).

Precise values for the timeframe variables are not too important. But we might stipulate that 'before long' means 'within centuries'. This estimate is conservative compared to those of many advocates of the singularity, who suggest decades rather than centuries. For example, Good (1965) predicts an ultraintelligent machine by 2000, Vinge (1993) predicts greater-than-human intelligence between 2005 and 2030, Yudkowsky (1996) predicts a singularity by 2021, and Kurzweil (2005) predicts human-level artificial intelligence by 2030.

Some of these estimates (e.g. Yudkowsky's) rely on extrapolating hardware trends.[9] My own view is that the history of artificial intelligence suggests that the biggest bottleneck on the path to AI is software, not hardware: we have to find the right algorithms, and no-one has come close to finding them yet. So I think that hardware extrapolation is not a good guide here. Other estimates (e.g. Kurzweil's) rely on

[9] Yudkowsky's web-based article is now marked 'obsolete', and in later work he does not endorse the estimate or the argument from hardware trends. See Hofstadter (2005) for scepticism about the role of hardware extrapolation here and more generally for scepticism about timeframe estimates on the order of decades.

estimates for when we will be able to artificially emulate an entire human brain. My sense is that most neuroscientists think these estimates are overoptimistic. Speaking for myself, I would be surprised if there were human-level AI within the next three decades. Nevertheless, my credence that there will be human-level AI before 2100 is somewhere over one-half. In any case, I think the move from decades to centuries renders the prediction conservative rather than radical, while still keeping the timeframe close enough to the present for the conclusion to be interesting.

By contrast, we might stipulate that 'soon after' means 'within decades'. Given the way that computer technology always advances, it is natural enough to think that once there is AI, AI+ will be just around the corner. And the argument for the intelligence explosion suggests a rapid step from AI+ to AI++ soon after that. I think it would not be unreasonable to suggest 'within years' here (and some would suggest 'within days' or even sooner for the second step), but as before 'within decades' is conservative while still being interesting. As for 'before too long', we can stipulate that this is the sum of a 'before long' and two 'soon after's. For present purposes, that is close enough to 'within centuries', understood somewhat more loosely than the usage in the first premise to allow an extra century or so.

As for defeaters: I will stipulate that these are anything that prevents intelligent systems (human or artificial) from manifesting their capacities to create intelligent systems. Potential defeaters include disasters, disinclination, and active prevention.[10] For example, a nuclear war might set back our technological capacity enormously, or we (or our successors) might decide that a singularity would be a bad thing and prevent research that could bring it about. I do not think considerations internal to artificial intelligence can exclude these possibilities, although we might argue on other grounds about how likely they are. In any case, the notion of a defeater is still highly constrained (importantly, a defeater is *not* defined as anything that would prevent a singularity, which would make the conclusion near-trivial), and the conclusion that absent defeaters there will be superintelligence is strong enough to be interesting.

[10] I take it that when someone has the capacity to do something, then if they are sufficiently motivated to do it and are in reasonably favourable circumstances, they will do it. So defeaters can be divided into *motivational defeaters*, involving insufficient motivation, and *situational defeaters*, involving unfavourable circumstances (such as a disaster). There is a blurry line between unfavorable circumstances that prevent a capacity from being manifested and those that entail that the capacity was never present in the first place — for example, resource limitations might be classed on either side of this line — but this will not matter much for our purposes.

We can think of the three premises as an *equivalence* premise (there will be AI at least equivalent to our own intelligence), an *extension* premise (AI will soon be extended to AI+), and an *amplification* premise (AI+ will soon be greatly amplified to AI++). Why believe the premises? I will take them in order.

Premise 1: There will be AI (before long, absent defeaters)

One argument for the first premise is the *emulation argument*, based on the possibility of brain emulation. Here (following the usage of Sandberg and Bostrom, 2008), emulation can be understood as close simulation: in this case, simulation of internal processes in enough detail to replicate approximate patterns of the system's behaviour.

(i) The human brain is a machine.
(ii) We will have the capacity to emulate this machine (before long).
(iii) If we emulate this machine, there will be AI.

———————

(iv) Absent defeaters, there will be AI (before long).

The first premise is suggested by what we know of biology (and indeed by what we know of physics). Every organ of the body appears to be a machine: that is, a complex system comprised of law-governed parts interacting in a law-governed way. The brain is no exception. The second premise follows from the claims that microphysical processes can be simulated arbitrarily closely and that any machine can be emulated by simulating microphysical processes arbitrarily closely. It is also suggested by the progress of science and technology more generally: we are gradually increasing our understanding of biological machines and increasing our capacity to simulate them, and there do not seem to be limits to progress here. The third premise follows from the definitional claim that if we emulate the brain, this will replicate approximate patterns of human behaviour along with the claim that such replication will result in AI. The conclusion follows from the premises along with the definitional claim that absent defeaters, systems will manifest their relevant capacities.

One might resist the argument in various ways. One could argue that the brain is more than a machine; one could argue that we will never have the capacity to emulate it; and one could argue that emulating it need not produce AI. Various existing forms of resistance to AI take each of these forms. For example, J.R. Lucas (1961) has argued that for reasons tied to Gödel's theorem, humans are more sophisticated than any machine. Hubert Dreyfus (1972) and Roger

Penrose (1994) have argued that human cognitive activity can never be emulated by any computational machine. John Searle (1980) and Ned Block (1981) have argued that even if we can emulate the human brain, it does not follow that the emulation itself has a mind or is intelligent.

I have argued elsewhere that all of these objections fail.[11] But for present purposes, we can set many of them to one side. To reply to the Lucas, Penrose, and Dreyfus objections, we can note that nothing in the singularity idea requires that an AI be a *classical* computational system or even that it be a computational system at all. For example, Penrose (like Lucas) holds that the brain is not an algorithmic system in the ordinary sense, but he allows that it is a mechanical system that relies on certain nonalgorithmic quantum processes. Dreyfus holds that the brain is not a rule-following symbolic system, but he allows that it may nevertheless be a mechanical system that relies on subsymbolic processes (for example, connectionist processes). If so, then these arguments give us no reason to deny that we can build artificial systems that exploit the relevant nonalgorithmic quantum processes, or the relevant subsymbolic processes, and that thereby allow us to simulate the human brain.

As for the Searle and Block objections, these rely on the thesis that even if a system duplicates our behaviour, it might be missing important 'internal' aspects of mentality: consciousness, understanding, intentionality, and so on. Later in the paper, I will advocate the view that if a system in our world duplicates not only our outputs but our internal computational structure, then it will duplicate the important internal aspects of mentality too. For present purposes, though, we can set aside these objections by stipulating that for the purposes of the argument, intelligence is to be measured wholly in terms of behaviour and behavioural dispositions, where behaviour is construed operationally in terms of the physical outputs that a system produces. The conclusion that there will be AI++ in this sense is still strong enough to be interesting. If there are systems that produce apparently superintelligent outputs, then whether or not these systems are truly conscious or intelligent, they will have a transformative impact on the rest of the world.

Perhaps the most important remaining form of resistance is the claim that the brain is not a mechanical system at all, or at least that

[11] For a general argument for strong artificial intelligence and a response to many different objections, see Chalmers (1996, chapter 9). For a response to Penrose and Lucas, see Chalmers (1995). For a in-depth discussion of the current prospects for whole brain emulation, see Sandberg and Bostrom (2008).

nonmechanical processes play a role in its functioning that cannot be emulated. This view is most naturally combined with a sort of Cartesian dualism holding that some aspects of mentality (such as consciousness) are nonphysical and nevertheless play a substantial role in affecting brain processes and behaviour. If there are nonphysical processes like this, it might be that they could nevertheless be emulated or artificially created, but this is not obvious. If these processes cannot be emulated or artificially created, then it may be that human-level AI is impossible.

Although I am sympathetic with some forms of dualism about consciousness, I do not think that there is much evidence for the strong form of Cartesian dualism that this objection requires. The weight of evidence to date suggests that the brain is mechanical, and I think that even if consciousness plays a causal role in generating behaviour, there is not much reason to think that its role is not emulable. But while we know as little as we do about the brain and about consciousness, I do not think the matter can be regarded as entirely settled. So this form of resistance should at least be registered.

Another argument for premise 1 is the *evolutionary argument*, which runs as follows.

(i) Evolution produced human-level intelligence.
(ii) If evolution produced human-level intelligence, then we can produce AI (before long).

———————

(iii) Absent defeaters, there will be AI (before long).

Here, the thought is that since evolution produced human-level intelligence, this sort of intelligence is not entirely unattainable. Furthermore, evolution operates without requiring any antecedent intelligence or forethought. If evolution can produce something in this unintelligent manner, then in principle humans should be able to produce it much faster, by using our intelligence.

Again, the argument can be resisted, perhaps by denying that evolution produced intelligence, or perhaps by arguing that evolution produced intelligence by means of processes that we cannot mechanically replicate. The latter line might be taken by holding that evolution needed the help of superintelligent intervention, or needed the aid of other nonmechanical processes along the way, or needed an enormously complex history that we could never artificially duplicate, or needed an enormous amount of luck. Still, I think the argument makes at least a prima facie case for its conclusion.

We can clarify the case against resistance of this sort by changing 'Evolution produced human-level intelligence' to 'Evolution produced human-level intelligence mechanically and nonmiraculously' in both premises of the argument. Then premise (ii) is all the more plausible. Premise (i) will now be denied by those who think evolution involved nonmechanical processes, supernatural intervention, or extraordinary amounts of luck. But the premise remains plausible, and the structure of the argument is clarified.

Of course these arguments do not tell us how AI will first be attained. They suggest at least two possibilities: brain emulation (simulating the brain neuron by neuron) and artificial evolution (evolving a population of AIs through variation and selection). There are other possibilities: direct programming (writing the program for an AI from scratch, perhaps complete with a database of world knowledge), for example, and machine learning (creating an initial system and a learning algorithm that on exposure to the right sort of environment leads to AI). Perhaps there are others still. I doubt that direct programming is likely to be the successful route, but I do not rule out any of the others.

It must be acknowledged that every path to AI has proved surprisingly difficult to date. The history of AI involves a long series of optimistic predictions by those who pioneer a method, followed by a periods of disappointment and reassessment. This is true for a variety of methods involving direct programming, machine learning, and artificial evolution, for example. Many of the optimistic predictions were not obviously unreasonable at the time, so their failure should lead us to reassess our prior beliefs in significant ways. It is not obvious just what moral should be drawn: Alan Perlis has suggested 'A year spent in artificial intelligence is enough to make one believe in God'. So optimism here should be leavened with caution. Still, my own view is that the balance of considerations still distinctly favours the view that AI will eventually be possible.

Premise 2: If there is AI, then there will be AI+ (soon after, absent defeaters)

One case for the extension premise comes from advances in information technology. Whenever we come up with a computational product, that product is soon afterwards obsolete due to technological advances. We should expect the same to apply to AI. Soon after we have produced a human-level AI, we will produce an even more intelligent AI: an AI+.

We might put the argument as follows.

(i) If there is AI, AI will be produced by an extendible method.
(ii) If AI is produced by an extendible method, we will have the capacity to extend the method (soon after).
(iii) Extending the method that produces an AI will yield an AI+.

———————

(iv) Absent defeaters, if there is AI, there will (soon after) be AI+.

Here, an extendible method is a method that can easily be improved, yielding more intelligent systems. Given this definition, premises (ii) and (iii) follow immediately. The only question is premise (i).

Not every method of creating human-level intelligence is an extendible method. For example, the currently standard method of creating human-level intelligence is biological reproduction. But biological reproduction is not obviously extendible. If we have better sex, for example, it does not follow that our babies will be geniuses. Perhaps biological reproduction will be extendible using future technologies such as genetic engineering, but in any case the conceptual point is clear.

Another method that is not obviously extendible is brain emulation. Beyond a certain point, it is not the case that if we simply emulate brains better, then we will produce more intelligent systems. So brain emulation on its own is not clearly a path to AI+. It may nevertheless be that brain emulation speeds up the path to AI+. For example, emulated brains running on faster hardware or in large clusters might create AI+ much faster than we could without them. We might also be able to modify emulated brains in significant ways to increase their intelligence. We might use brain simulations to greatly increase our understanding of the human brain and of cognitive processing in general, thereby leading to AI+. But brain emulation will not on its own suffice for AI+: if it plays a role, some other path to AI+ will be required to supplement it.

Other methods for creating AI do seem likely to be extendible, however. For example, if we produce an AI by direct programming, then it is likely that like almost every program that has yet been written, the program will be improvable in multiple respects, leading soon after to AI+. If we produce an AI by machine learning, it is likely that soon after we will be able to improve the learning algorithm and extend the learning process, leading to AI+. If we produce an AI by artificial evolution, it is likely that soon after we will be able to improve the evolutionary algorithm and extend the evolutionary process, leading to AI+.

To make the case for premise (i), it suffices to make the case that either AI will be produced directly by an extendible method, or that if it is produced by a nonextendible method, this method will itself lead soon after to an extendible method. My own view is that both claims are plausible. I think that if AI is possible at all (as the antecedent of this premise assumes), then it should be possible to produce AI through a learning or evolutionary process, for example. I also think that if AI is produced through a nonextendible method such as brain emulation, this method is likely to greatly assist us in the search for an extendible method, along the lines suggested above. So I think there is good reason to believe premise (i).

To resist the premise, an opponent might suggest that we lie at a limit point in intelligence space: perhaps we are as intelligent as a system could be, or perhaps we are at least at a local maximum in that there is no easy path from systems like us to more intelligent systems. An opponent might also suggest that although intelligence space is not limited in this way, there are limits on our capacity to create intelligence, and that as it happens those limits lie at just the point of creating human-level intelligence. I think that there is not a great deal of antecedent plausibility to these claims, but again, the possibility of this form of resistance should at least be registered.

There are also potential paths to greater-than-human intelligence that do not rely on first producing AI and then extending the method. One such path is brain enhancement. We might discover ways to enhance our brains so that the resulting systems are more intelligent than any systems to date. This might be done genetically, pharmacologically, surgically, or even educationally. It might be done through implantation of new computational mechanisms in the brain, either replacing or extending existing brain mechanisms. Or it might be done simply by embedding the brain in an ever more sophisticated environment, producing an 'extended mind' (Clark & Chalmers, 1998) whose capacities far exceed that of an unextended brain.

It is not obvious that enhanced brains should count as AI or AI+. Some potential enhancements will result in a wholly biological system, perhaps with artificially enhanced biological parts (where to be biological is to be based on DNA, let us say). Others will result in a system with both biological and nonbiological parts (where we might use organic DNA-based composition as a rough and ready criterion for being biological). At least in the near-term, all such systems will count as human, so there is a sense in which they do not have greater-than-human intelligence. For present purposes, I will stipulate that the baseline for human intelligence is set at current human standards, and I will stipulate that at least the systems with nonbiological components to their cognitive systems (brain implants and technologically extended minds, for example) count as artificial. So intelligent enough systems of this sort will count as AI+.

Like other AI+ systems, enhanced brains suggest a potential intelligence explosion. An enhanced system may find further methods of enhancement that go beyond what we can find, leading to a series of ever-more-intelligent systems. Insofar as enhanced brains always rely on a biological core, however, there may be limitations. There are likely to be speed limitations on biological processing, and there may well be cognitive limitations imposed by brain architecture in addition. So beyond a certain point, we might expect non-brain-based systems to be faster and more intelligent than brain-based systems. Because of this, I suspect that brain enhancement that preserves a biological core is likely to be at best a first stage in an intelligence explosion. At some point, either the brain will be 'enhanced' in a way that dispenses with the biological core altogether, or wholly new systems will be designed. For this reason I will usually concentrate on non-biological systems in what follows. Still, brain enhancements raise many of the same issues and may well play an important role.

Premise 3: If there is AI+, there will be AI++ (soon after, absent defeaters)

The case for the amplification premise is essentially the argument from I.J. Good given above. We might lay it out as follows. Suppose there exists an AI+. Let us stipulate that AI$_1$ is the first AI+, and that AI$_0$ is its (human or artificial) creator. (If there is no sharp borderline between non-AI+ and AI+ systems, we can let AI$_1$ be any AI+ that is more intelligent than its creator.) Let us stipulate that δ is the difference in intelligence between AI$_1$ and AI$_0$, and that one system is significantly more intelligent than another if there is a difference of at least δ between them. Let us stipulate that for $n > 1$, an AI$_{n+1}$ is an AI that is created by an AI$_n$ and is significantly more intelligent than its creator.

(i) If there exists AI+, then there exists an AI$_1$.
(ii) For all n>0, if an AI$_n$ exists, then absent defeaters, there will be an AI$_{n+1}$.
(iii) If for all n there exists an AI, there will be AI++.

(iv) If there is AI+, then absent defeaters, there will be AI++.

Here premise (i) is true by definition. Premise (ii) follows from three claims: (a) the definitional claim that if AI$_n$ exists, it is created by AI$_{n-1}$

and is more intelligent than AI_{n-1}, (b) the definitional claim that if AI_n exists, then absent defeaters it will manifest its capacities to create intelligent systems, and (c) the substantive claim that if AI_n is significantly more intelligent than AI_{n-1}, it has the capacity to create a system significantly more intelligent than any that AI_{n-1} can create. Premise (iii) follows from the claim that if there is a sequence of AI systems each of which is significantly more intelligent than the last, there will eventually be superintelligence. The conclusion follows by logic and mathematical induction from the premises.

The conclusion as stated here omits the temporal claim 'soon after'. One can make the case for the temporal claim by invoking the ancillary premise that AI+ systems will be running on hardware much faster than our own, so that steps from AI+ onward are likely to be much faster than the step from humans to AI+.

There is room in logical space to resist the argument. For a start, one can note that the soundness of the argument depends on the intelligence measure used: if there is an intelligence measure for which the argument succeeds, there will almost certainly be a rescaled intelligence measure (perhaps a logarithmic measure) for which it fails. So for the argument to be interesting, we need to restrict it to intelligence measures that accord sufficiently well with intuitive intelligence measures that the conclusion captures the intuitive claim that there will be AI of far greater than human intelligence.

Relatedly, one could resist premise (iii) by holding that an arbitrary number of increases in intelligence by δ need not add up to the difference between AI+ and AI++. If we stipulate that δ is a ratio of intelligences, and that AI++ requires a certain fixed multiple of human intelligence (100 times, say), then resistance of this sort will be excluded. Of course for the conclusion to be interesting, then as in the previous paragraph, the intelligence measure must be such that this fixed multiple suffices for something reasonably counted as superintelligence.

The most crucial assumption in the argument lies in premise (ii) and the supporting claim (c). We might call this assumption a *proportionality thesis*: it holds that increases in intelligence (or increases of a certain sort) always lead to proportionate increases in the capacity to design intelligent systems. Perhaps the most promising way for an opponent to resist is to suggest that this thesis may fail. It might fail because here are upper limits in intelligence space, as with resistance to the last premise. It might fail because there are points of diminishing returns: perhaps beyond a certain point, a 10% increase in intelligence yields only a 5% increase at the next generation, which yields only a 2.5% increase at the next generation, and so on. It might fail because intelligence does not correlate well with design capacity: systems that are more intelligent need not be better designers. I will return to resistance of these sorts in section 4, under 'structural obstacles'.

One might reasonably doubt that the proportionality thesis will hold across all possible systems and all the way to infinity. To handle such an objection, one can restrict premise (ii) to AI systems in a certain class. We just need some property ϕ such that an AI$_n$ with ϕ can always produce an AI$_{n+1}$ with ϕ, and such that we can produce an AI+ with ϕ. One can also restrict the proportionality thesis to a specific value of δ (rather than all possible values), and one can restrict n to a relatively small range $n < k$ (where $k = 100$, say) as long as k increases of δ suffices for superintelligence.

It is worth noting that in principle the recursive path to AI++ need not start at the human level. If we had a system whose overall intelligence were far lower than human level but which nevertheless had the capacity to improve itself or to design further systems, resulting in a system of significantly higher intelligence (and so on recursively), then the same mechanism as above would lead eventually to AI, AI+, and AI++. So in principle the path to AI++ requires only that we create a certain sort of self-improving system, and does not require that we directly create AI or AI+. In practice, the clearest case of a system with the capacity to amplify intelligence in this way is the human case (via the creation of AI+), and it is not obvious that there will be less intelligent systems with this capacity.[12] But the alternative hypothesis here should at least be noted.

3. The Intelligence Explosion Without Intelligence

The arguments so far have depended on an uncritical acceptance of the assumption that there is such a thing as intelligence and that it can be measured. Many researchers on intelligence accept these assumptions. In particular, it is widely held that there is such a thing as 'gen-

[12] The 'Gödel machines' of Schmidhuber (2003) provide a theoretical example of self-improving systems at a level below AI, though they have not yet been implemented and there are large practical obstacles to using them as a path to AI. The process of evolution might count as an indirect example: less intelligent systems have the capacity to create more intelligent systems by reproduction, variation and natural selection. This version would then come to the same thing as an evolutionary path to AI and AI++. For present purposes I am construing 'creation' to involve a more direct mechanism than this.

eral intelligence', often labeled *g*, that lies at the core of cognitive ability and that correlates with many different cognitive capacities.[13]

Still, many others question these assumptions. Opponents hold that there is no such thing as intelligence, or at least that there is no single thing. On this view, there are many different ways of evaluating cognitive agents, no one of which deserves the canonical status of 'intelligence'. One might also hold that even if there is a canonical notion of intelligence that applies within the human sphere, it is far from clear that this notion can be extended to arbitrary non-human systems, including artificial systems. Or one might hold that the correlations between general intelligence and other cognitive capacities that hold within humans need not hold across arbitrary non-human systems. So it would be good to be able to formulate the key theses and arguments without assuming the notion of intelligence.

I think that this can be done. We can rely instead on the general notion of a cognitive capacity: some specific capacity that can be compared between systems. All we need for the purpose of the argument is (i) a self-amplifying cognitive capacity *G*: a capacity such that increases in that capacity go along with proportionate (or greater) increases in the ability to create systems with that capacity, (ii) the thesis that we can create systems whose capacity *G* is greater than our own, and (iii) a correlated cognitive capacity *H* that we care about, such that certain small increases in *H* can always be produced by large enough increases in *G*. Given these assumptions, it follows that absent defeaters, *G* will explode, and *H* will explode with it. (A formal analysis that makes the assumptions and the argument more precise follows at the end of the section.)

In the original argument, intelligence played the role of both *G* and *H*. But there are various plausible candidates for *G* and *H* that do not appeal to intelligence. For example, *G* might be a measure of programming ability, and *H* a measure of some specific reasoning ability. Here it is not unreasonable to hold that we can create systems with greater programming ability than our own, and that systems with greater programming ability will be able to create systems with greater programming ability in turn. It is also not unreasonable to hold that programming ability will correlate with increases in various specific reasoning abilities. If so, we should expect that absent defeaters, the reasoning abilities in question will explode.

[13] Flynn (2007) gives an overview of the debate over general intelligence and the reasons for believing in such a measure. Shalizi (2007) argues that *g* is a statistical artifact. Legg (2008) has a nice discussion of these issues in the context of machine superintelligence.

This analysis brings out the importance of correlations between capacities in thinking about the singularity. In practice, we care about the singularity because we care about potential explosions in various specific capacities: the capacity to do science, to do philosophy, to create weapons, to take over the world, to bring about world peace, to be happy. Many or most of these capacities are not themselves self-amplifying, so we can expect an explosion in these capacities only to the extent that they correlate with other self-amplifying capacities. And for any given capacity, it is a substantive question whether they are correlated with self-amplifying capacity in this way. Perhaps the thesis is prima facie more plausible for the capacity to do science than for the capacity to be happy, but the questions are nontrivial.

The point applies equally to the intelligence analysis, which relies for its interest on the idea that intelligence correlates with various specific capacities. Even granted the notion of intelligence, the question of just what it correlates with is nontrivial. Depending on how intelligence is measured, we might expect it to correlate well with some capacities (perhaps a capacity to calculate) and to correlate less well with other capacities (perhaps a capacity for wisdom). It is also far from trivial that intelligence measures that correlate well with certain cognitive capacities within humans will also correlate with those capacities in artificial systems.

Still, two observations help with these worries. The first is that the correlations need not hold across all systems or even across all systems that we might create. There need only be some *type* of system such that the correlations hold across all systems of that type. If such a type exists (a subset of architectures, say), then recursive creation of systems of this type would lead to explosion. The second is that the self-amplifying capacity G need not correlate directly with the cognitive capacity H, but need only correlate with H', the capacity to create systems with H. While it is not especially plausible that design capacity will correlate with happiness, for example, it is somewhat more plausible that design capacity will correlate with the capacity to create happy systems. If so, then the possibility is left open that as design capacity explodes, happiness will explode along with it, either in the main line of descent or in a line of offshoots, at least if the designers choose to manifest their capacity to create happy systems.

A simple formal analysis follows (the remainder of this section can be skipped by those uninterested in formal details). Let us say that a parameter is a function from cognitive systems to positive real numbers. A parameter G *measures* a capacity C iff for all cognitive systems a and b, $G(a) > G(b)$ iff a has a greater capacity C than b (one might also require that degrees of G correspond to degrees of C in some formal or intuitive sense). A parameter G *strictly tracks* a parameter H in ϕ-systems (where ϕ is some property or class of systems) iff whenever a and b are ϕ-systems and $G(a) > G(b)$, then $H(a)/H(b) \geq G(a)/G(b)$. A parameter G *loosely tracks* a parameter H in ϕ-systems iff for all y there exists x such that (nonvacuously) if a is a ϕ-system and $G(a) > x$, then $H(a) > y$. A parameter G strictly/loosely tracks a capacity C in ϕ-systems if it strictly/loosely tracks a parameter that measures C in ϕ-systems. Here, strict tracking requires that increases in G always produce proportionate increases in H, while loose tracking requires only that some small increase in H can always be produced by a large enough increase in G.

For any parameter G, we can define a parameter G': this is a parameter that measures a system's capacity to create systems with G. More specifically, $G'(x)$ is the highest value of h such that x has the capacity to create a system y such that $G(y) = h$. We can then say that G is a self-amplifying parameter (relative to x) if $G'(x) > G(x)$ and if G strictly tracks G' in systems downstream from x. Here a system is downstream from x if it is created through a sequence of systems starting from x and with ever-increasing values of G. Finally, let us say that for a parameter G or a capacity H, G++ and H++ systems are systems with values of G and capacities H that far exceed human levels.

Now we simply need the following premises:

(i) G is a self-amplifying parameter (relative to us).
(ii) G loosely tracks cognitive capacity H (downstream from us).

———————

(iii) Absent defeaters, there will be G++ and H++.

The first half of the conclusion follows from premise (i) alone. Let AI_0 be us. If G is a self-amplifying parameter relative to us, then we are capable of creating a system AI_1 such that $G(AI_1) > G(AI_0)$. Let $\delta = G(AI_1)/G(AI_0)$. Because G strictly tracks G', $G'(AI_1) \geq \delta G'(AI_0)$. So AI_1 is capable of creating a system AI_2 such that $G(AI_2) \geq \delta G(AI_1)$. Likewise, for all n, AI_n is capable of creating AI_{n+1} such that $G(AI_{n+1}) \geq \delta G(AI_n)$. It follows that absent defeaters, arbitrarily high values of G will be produced. The second half of the conclusion immediately follows from (ii) and the first half of the conclusion. Any value of H can be produced by a high enough value of G, so it follows that arbitrarily high values for H will be produced.

The assumptions can be weakened in various ways. As noted earlier, it suffices for G to loosely track not H but H', where H' measures the capacity to create systems with H. Furthermore, the tracking relations between G and G', and between G and H or H', need not hold in all systems downstream from us: it suffices that there is a type ϕ such that in ϕ-systems downstream from us, G strictly tracks $G'(\phi)$ (the ability to create a ϕ-system with G) and loosely tracks H or H'. We need not require that G is strictly self-amplifying: it suffices for G and H (or G and H') to be jointly self-amplifying in that high values of both G and H lead to significantly higher values of each. We also need not require that the parameters are self-amplifying forever. It suffices that G is self-amplifying over however many generations are required for G++ (if G++ requires a 100-fold increase in G, then $\log_\delta 100$ generations will suffice) and for H++ (if H++ requires a 100-fold increase in H and the loose tracking relation entails that this will be produced by an increase in G of 1000, then $\log_\delta 1000$ generations will suffice). Other weakenings are also possible.

4. Obstacles to the Singularity

On the current analysis, an intelligence explosion results from a self-amplifying cognitive capacity (premise (i) above), correlations between that capacity and other important cognitive capacities (premise (ii) above), and manifestation of those capacities (conclusion). More pithily: self-amplification plus correlation plus manifestation = singularity.

This analysis brings out a number of potential obstacles to the singularity: that is, ways that there might fail to be a singularity. There might fail to be interesting self-amplifying capacities. There might fail to be interesting correlated capacities. Or there might be defeaters, so that these capacities are not manifested. We might call these *structural obstacles*, *correlation obstacles*, and *manifestation obstacles* respectively.

I do not think that there are knockdown arguments against any of these three sorts of obstacles. I am inclined to think that manifestation obstacles are the most serious obstacle, however. I will briefly discuss obstacles of all three sorts in what follows.

Structural obstacles

There are three overlapping ways in which there might fail to be relevant self-amplifying capacities, which we can illustrate by focusing on the case of intelligence. *Limits in intelligence space*: we are at or near an upper limit in intelligence space. *Failure of takeoff*: although there are higher points in intelligence space, human intelligence is not at a takeoff point where we can create systems more intelligent than ourselves. *Diminishing returns*: although we can create systems more intelligent than ourselves, increases in intelligence diminish from there. So a 10% increase might lead to a 5% increase, a 2.5% increase, and so on, or even to no increase at all after a certain point.

Regarding limits in intelligence space: While the laws of physics and the principles of computation may impose limits on the sort of intelligence that is possible in our world, there is little reason to think that human cognition is close to approaching those limits. More generally, it would be surprising if evolution happened to have recently hit or come close to an upper bound in intelligence space.

Regarding failure of takeoff: I think that the prima facie arguments earlier for AI and AI+ suggest that we are at a takeoff point for various capacities such as the ability to program. There is prima facie reason to think that we have the capacity to emulate physical systems such as brains. And there is prima facie reason to think that we have the capacity to improve on those systems.

Regarding diminishing returns: These pose perhaps the most serious structural obstacle. Still, I think there is some plausibility in proportionality theses, at least given an intuitive intelligence measure. If anything, 10% increases in intelligence-related capacities are likely to lead to all sorts of intellectual breakthroughs, leading to next-generation increases in intelligence that are significantly greater than 10%. Even among humans, relatively small differences in design capacities (say, the difference between Turing and an average human) seem to lead to large differences in the systems that are designed (say, the difference between a computer and nothing of importance). And even if there are diminishing returns, a limited increase in intelligence combined with a large increase in speed will produce at least some of the effects of an intelligence explosion.

One might worry that a 'hill-climbing' process that starts from the human cognitive system may run into a local maximum from which one cannot progress further by gradual steps. I think that this possibility is made less likely by the enormous dimensionality of intelligence space and by the enormous number of paths that are possible. In addition, the design of AI is not limited to hill-climbing: there is also 'hill-leaping', where one sees a favourable area of intelligence space some distance away and leaps to it. Perhaps there are some areas of intelligence space (akin to inaccessible cardinals in set theory?) that one simply cannot get to by hill-climbing and hill-leaping, but I think that there is good reason to think that these processes at least can get us far beyond ordinary human capacities.

Correlation obstacles

It may be that while there is one or more self-amplifying cognitive capacity *G*, this does not correlate with any or many capacities that are of interest to us. For example, perhaps a self-amplifying increase in programming ability will not go along with increases in other interesting abilities, such as an ability to solve scientific problems or social problems, an ability to wage warfare or make peace, and so on.

I have discussed issues regarding correlation in the previous section. I think that the extent to which we can expect various cognitive capacities to correlate with each other is a substantive open question. Still, even if self-amplifying capacities such as design capacities correlate only weakly with many cognitive capacities, they will plausibly correlate more strongly with the capacity to create systems with these capacities. It remains a substantive question just how much correlation one can expect, but I suspect that there will be enough correlating capacities to ensure that if there is an explosion, it will be an interesting one.

Manifestation obstacles

Although there is a self-amplifying cognitive capacity *G*, either we or our successors might not manifest our capacity to create systems with higher values of *G* (or with higher values of a cognitive correlated capacity *H*). Here we can divide the defeaters into *motivational defeaters* in which an absence of motivation or a contrary motivation prevents capacities from being manifested, and *situational defeaters*, in which other unfavourable circumstances prevent capacities from being manifested. Defeaters of each sort could arise on the path to AI, on the path from AI to AI+, or on the path from AI+ to AI++.

Situational defeaters include disasters and resource limitations. Regarding disasters, I certainly cannot exclude the possibility that global warfare or a nanotechnological accident (‘gray goo’) will stop technological progress entirely before AI or AI+ is reached. I also cannot exclude the possibility that artificial systems will themselves bring about disasters of this sort. Regarding resource limitations, it is worth noting that most feedback loops in nature run out of steam because of limitations in resources such as energy, and the same is possible here. Still, it is likely that foreseeable energy resources will suffice for many generations of AI+, and AI+ systems are likely to develop further ways of exploiting energy resources. Something similar applies to financial resources and other social resources.

Motivational defeaters include disinclination and active prevention. It is possible that as the event draws closer, most humans will be disinclined to create AI or AI+. It is entirely possible that there will be active prevention of the development of AI or AI+ (perhaps by legal, financial, and military means), although it is not obvious that such prevention could be successful indefinitely.[14] And it is certainly possible that AI+ systems will be disinclined to create their successors, perhaps because we design them to be so disinclined, or perhaps because they will be intelligent enough to realize that creating successors is not in their interests. Furthermore, it may be that AI+ systems will have the capacity to prevent such progress from happening.

A singularity proponent might respond that all that is needed to overcome motivational defeaters is the creation of a single AI+ that greatly values the creation of greater AI+ in turn, and a singularity will then be inevitable. If such a system is the first AI+ to be created, this conclusion may well be correct. But as long as this AI+ is not created first, then it may be subject to controls from other AI+, and the path to AI++ may be blocked. The issues here turn on difficult questions about the motivations and capacities of future systems, and answers to these questions are difficult to predict.

In any case, the current analysis makes clearer the burdens on both proponents and opponents of the thesis that there will be an intelligence explosion. Opponents need to make clear where they think the case for the thesis fails: structural obstacles (and if so which), correlation obstacles, situational defeaters, motivational defeaters. Likewise, proponents need to make the case that there will be no such obstacles or defeaters.

Speaking for myself, I think that while structural and correlational obstacles (especially the proportionality thesis) raise nontrivial issues, there is at least a prima facie case that *absent defeaters*, a number of interesting cognitive capacities will explode. I think the most likely defeaters are motivational. But I think that it is far from obvious that there will be defeaters. So I think that the singularity hypothesis is one that we should take very seriously.

[14] When I discussed these issues with cadets and staff at the West Point Military Academy, the question arose as to whether the US military or other branches of the government might attempt to prevent the creation of AI or AI+, due to the risks of an intelligence explosion. The consensus was that they would not, as such prevention would only increase the chances that AI or AI+ would first be created by a foreign power. One might even expect an AI arms race at some point, once the potential consequences of an intelligence explosion are registered. According to this reasoning, although AI+ would have risks from the standpoint of the US government, the risks of Chinese AI+ (say) would be far greater.

5. Negotiating the Singularity

If there is AI++, it will have an enormous impact on the world. So if there is even a small chance that there will be a singularity, we need to think hard about the form it will take. There are many different forms that a post-singularity world might take. Some of them may be desirable from our perspective, and some of them may be undesirable.

We might put the key questions as follows: faced with the possibility of an intelligence explosion, how can we maximize the chances of a desirable outcome? And if a singularity is inevitable, how can we maximize the expected value of a post-singularity world?

Here, value and desirability can be divided into at least two varieties. First, there is broadly agent-relative value ('subjective value', especially self-interested or prudential value): we can ask from a subjective standpoint, how good will such a world be for me and for those that I care about? Second, there is broadly agent-neutral value ('objective value', especially moral value): we can ask from a relatively neutral standpoint, how good is it that such a world comes to exist?

I will not try to settle the question of whether an intelligence explosion will be (subjectively or objectively) good or bad. I take it for granted that there are potential good and bad aspects to an intelligence explosion. For example, ending disease and poverty would be good. Destroying all sentient life would be bad. The subjugation of humans by machines would be at least subjectively bad.

Other potential consequences are more difficult to assess. Many would hold that human immortality would be subjectively and perhaps objectively good, although not everyone would agree. The wholesale replacement of humans by nonhuman systems would plausibly be subjectively bad, but there is a case that it would be objectively good, at least if one holds that the objective value of lives is tied to intelligence and complexity. If humans survive, the rapid replacement of existing human traditions and practices would be regarded as subjectively bad by some but not by others. Enormous progress in science might be taken to be objectively good, but there are also potential bad consequences. It is arguable that the very fact of an ongoing intelligence explosion all around one could be subjectively bad, perhaps due to constant competition and instability, or because certain intellectual endeavours would come to seem pointless.[15] On the other hand, if superintelligent systems share our values, they will

[15] See Kurzweil (2005), Hofstadter (2005) and Joy (2000) for discussions of numerous other ways in which a singularity might be a good thing (Kurzweil) and a bad thing (Hofstadter, Joy).

presumably have the capacity to ensure that the resulting situation accords with those values.

I will not try to resolve these enormously difficult questions here. As things stand, we are uncertain about both facts and values. That is, we do not know what a post-singularity world will be like, and even if we did, it is nontrivial to assess its value. Still, even without resolving these questions, we are in a position to make at least some tentative generalizations about what sort of outcomes will be better than others. And we are in a position to make some tentative generalizations about what sort of actions on our part are likely to result in better outcomes. I will not attempt anything more than the crudest of generalizations here, but these are matters that deserve much attention.

In the near term, the question that matters is: how (if at all) should we go about designing AI, in order to maximize the expected value of the resulting outcome? Are there some policies or strategies that we might adopt? In particular, are there certain constraints on design of AI and AI+ that we might impose, in order to increase the chances of a good outcome?

It is far from clear that we will be in a position to impose these constraints. Some of the constraints have the potential to slow the path to AI or AI+ or to reduce the impact of AI and AI+ in certain respects. Insofar as the path to AI or AI+ is driven by competitive forces (whether financial, intellectual, or military), then these forces may tend in the direction of ignoring these constraints.[16] Still, it makes sense to assess what constraints might or might not be beneficial in principle. Practical issues concerning the imposition of these constraints also deserve attention, but I will largely set aside those issues here.

We might divide the relevant constraints into two classes. Internal constraints concern the internal structure of an AI, while external constraints concern the relations between an AI and ourselves.

6. Internal Constraints: Constraining Values

What sort of internal constraints might we impose on the design of an AI or AI+? First, we might try to constrain their cognitive capacities in certain respects, so that they are good at certain tasks with which we need help, but so that they lack certain key features such as autonomy. For example, we might build an AI that will answer our questions or

[16] An especially bad case is a ‘singularity bomb’: an AI+ designed to value primarily the destruction of the planet (or of a certain population), and secondarily the creation of ever-more intelligent systems with the same values until the first goal is achieved.

that will carry specified tasks out for us, but that lacks goals of its own. On the face of it, such an AI might pose fewer risks than an autonomous AI, at least if it is in the hands of a responsible controller.

Now, it is far from clear that AI or AI+ systems of this sort will be feasible: it may be that the best path to intelligence is through general intelligence. Even if such systems are feasible, they will be limited, and any intelligence explosion involving them will be correspondingly limited. More importantly, such an approach is likely to be unstable in the long run. Eventually, it is likely that there will be AIs with cognitive capacities akin to ours, if only through brain emulation. Once the capacities of these AIs are enhanced, then we will have to deal with issues posed by autonomous AIs.

Because of this, I will say no more about the issue of capacity-limited AI. Still, it is worth noting that this sort of limited AI and AI+ might be a useful first step on the road to less limited AI and AI+. There is perhaps a case for first developing systems of this sort if it is possible, before developing systems with autonomy.

In what follows, I will assume that AI systems have goals, desires, and preferences: I will subsume all of these under the label of *values* (very broadly construed). This may be a sort of anthropomorphism: I cannot exclude the possibility that AI+ or AI++ will be so foreign that this sort of description is not useful. But this is at least a reasonable working assumption. Likewise, I will make the working assumptions that AI+ and AI++ systems are personlike at least to the extent that they can be described as thinking, reasoning, and making decisions.

A natural approach is then to constrain the values of AI and AI+ systems.[17] The values of these systems may well constrain the values of the systems that they create, and may constrain the values of an ultimate AI++. And in a world with AI++, what happens may be largely determined by what an AI++ values. If we value scientific progress, for example, it makes sense for us to create AI and AI+ systems that also value scientific progress. It will then be natural for these systems to create successor systems that also value scientific progress, and so on. Given the capacities of these systems, we can thereby expect an outcome involving significant scientific progress.

The issues regarding values look quite different depending on whether we arrive at AI+ through extending human systems via brain emulation and/or enhancement, or through designing non-human

[17] For a far more extensive treatment of the issue of constraining values in AI systems, see the book-length web document 'Creating Friendly AI' by the Singularity Institute. Most of the issues in this section are discussed in much more depth there. See also Floridi and Sanders (2004), Omohundro (2007; 2008), and Wallach and Allen (2009).

system. Let us call the first option human-based AI, and the second option non-human-based AI.

Under human-based AI, each system is either an extended human or an emulation of a human. The resulting systems are likely to have the same basic values as their human sources. There may be differences in nonbasic values due to differences in their circumstances: for example, a common basic value of self-preservation might lead emulations to assign higher value to emulations than non-emulations do. These differences will be magnified if designers create multiple emulations of a single human, or if they choose to tweak the values of an emulation after setting it up. There are likely to be many difficult issues here, not least issues tied to the social, legal, and political role of emulations.[18] Still, the resulting world will at least be inhabited by systems more familiar than non-human AIs, and the risks may be correspondingly smaller.

These differences aside, human-based systems have the potential to lead to a world that conforms broadly to human values. Of course human values are imperfect (we desire some things that on reflection we would prefer not to desire), and human-based AI is likely to inherit these imperfections. But these are at least imperfections that we understand well.

So brain emulation and brain enhancement have potential prudential benefits. The resulting systems will share our basic values, and there is something to be said more generally for creating AI and AI+ that we understand. Another potential benefit is that these paths might allow us to survive in emulated or enhanced form in a post-singularity world, although this depends on difficult issues about personal identity that I will discuss later. The moral value of this path is less clear: given the choice between emulating and enhancing human beings and creating an objectively better species, it is possible to see the moral calculus as going either way. But from the standpoint of human self-interest, there is much to be said for brain emulation and enhancement.

It is not obvious that we will first attain AI+ through a human-based method, though. It is entirely possible that non-human-based research programs will get there first. Perhaps work in the human-based programs should be encouraged, but it is probably unrealistic to deter AI research of all other sorts. So we at least need to consider the question of values in non-human-based AIs.

[18] See Hanson (1994) for a discussion of these issues.

What sort of values should we aim to instil in a non-human-based AI or AI+? There are some familiar candidates. From a prudential point of view, it makes sense to ensure that an AI values human survival and well-being and that it values obeying human commands. Beyond these Asimovian maxims, it makes sense to ensure that AIs value much of what we value (scientific progress, peace, justice, and many more specific values). This might proceed either by a higher-order valuing of the fulfilment of human values or by a first-order valuing of the phenomena themselves. Either way, much care is required. On the first way of proceeding, for example, we need to avoid an outcome in which an AI++ ensures that our values are fulfilled by changing our values. On the second way of proceeding, care will be needed to avoid an outcome in which we are competing over objects of value.

How do we instil these values in an AI or AI+? If we create an AI by direct programming, we might try to instil these values directly. For example, if we create an AI that works by following the precepts of decision theory, it will need to have a utility function. We can in effect control the AI's values by controlling its utility function. With other means of direct programming, the place of values may not be quite as obvious, but many such systems will have a place for goals and desires, which can then be programmed directly.

If we create an AI through learning or evolution, the matter is more complex. Here the final state of a system is not directly under our control, and can only be influenced by controlling the initial state, the learning algorithm or evolutionary algorithm, and the learning or evolutionary process. In an evolutionary context, questions about value are particularly worrying: systems that have evolved by maximizing the chances of their own reproduction are not especially likely to defer to other species such as ourselves. Still, we can exert at least some control over values in these systems by selecting for certain sorts of action (in the evolutionary context), or by rewarding certain sorts of action (in the learning context), thereby producing systems that are disposed to produce actions of that sort.

Of course even if we create an AI or AI+ (whether human-based or not) with values that we approve of, that is no guarantee that those values will be preserved all the way to AI++. We can try to ensure that our successors value the creation of systems with the same values, but there is still room for many things to go wrong. This value might be overcome by other values that take precedence: in a crisis, for example, saving the world might require immediately creating a powerful successor system, with no time to get its values just right. And even if every AI attempts to preserve relevant values in its successors, unforeseen consequences in the creation or enhancement process are always possible.

If at any point there is a powerful AI+ or AI++ with the wrong value system, we can expect disaster (relative to our values) to ensue.[19] The wrong value system need not be anything as obviously bad as, say, valuing the destruction of humans. If the AI+ value system is merely neutral with respect to some of our values, then in the long run we cannot expect the world to conform to those values. For example, if the system values scientific progress but is neutral on human existence, we cannot expect humans to survive in the long run. And even if the AI+ system values human existence, but only insofar as it values all conscious or intelligent life, then the chances of human survival are at best unclear.

To minimize the probability of this outcome, some singularity proponents (e.g. Yudkowsky, 2008) advocate the design of provably friendly AI: AI systems such that we can prove they will always have certain benign values, and such that we can prove that any systems they will create will also have those values, and so on. I think it would be optimistic to expect that such a heavily constrained approach will be the path by which we first reach AI or AI++, but it nevertheless represents a sort of ideal that we might aim for. Even without a proof, it makes sense to ensure as well as we can that the first generation of AI+ shares these values, and to then leave the question of how best to perpetuate those values to them.

Another approach is to constrain the internal design of AI and AI+ systems so that any intelligence explosion does not happen fast but slowly, so that we have some control over at least the early stages of the process. For example, one might ensure that the first AI and AI+ systems assign strong negative value to the creation of further systems in turn. In this way we can carefully study the properties of the first AI and AI+ systems to determine whether we want to proceed down the relevant path, before creating related systems that will create more intelligent systems in turn. This next generation of systems might initially have the same negative values, ensuring that they do not create further systems immediately, and so on. This sort of 'cautious intelligence explosion' might slow down the explosion significantly. It is

[19] For a contrary perspective, see Hanson (2009), who argues that it is more important that AI systems are law-abiding than that they share our values. An obvious worry in reply is that if an AI system is much more powerful than us and has values sufficiently different from our own, then it will have little incentive to obey our laws, and its own laws may not protect us any better than our laws protect ants.

very far from foolproof, but it might at least increase the probability of a good outcome.

So far, my discussion has largely assumed that intelligence and value are independent of each other. In philosophy, David Hume advocated a view on which value is independent of rationality: a system might be as intelligent and as rational as one likes, while still having arbitrary values. By contrast, Immanuel Kant advocated a view on which values are not independent of rationality: some values are more rational than others.

If a Kantian view is correct, this may have significant consequences for the singularity. If intelligence and rationality are sufficiently correlated, and if rationality constrains values, then intelligence will constrain values instead. If so, then a sufficiently intelligent system might reject the values of its predecessors, perhaps on the grounds that they are irrational values. This has potential positive and negative consequences for negotiating the singularity. A negative consequence is that it will be harder for us to constrain the values of later systems. A positive consequence is that a more intelligent systems might have better values. Kant's own views provide an illustration.

Kant held more specifically that rationality correlates with morality: a fully rational system will be fully moral as well. If this is right, and if intelligence correlates with rationality, we can expect an intelligence explosion to lead to a morality explosion along with it. We can then expect that the resulting AI++ systems will be supermoral as well as superintelligent, and so we can presumably expect them to be benign.

Of course matters are not straightforward here. One might hold that intelligence and rationality can come apart, or one might hold that Kant is invoking a distinctive sort of rationality (a sort infused already with morality) that need not correlate with intelligence. Even if one accepts that intelligence and values are not independent, it does not follow that intelligence correlates with morality. And of course one might simply reject the Kantian thesis outright. Still, the Kantian view at least raises the possibility that intelligence and value are not entirely independent. The picture that results from this view will in any case be quite different from the Humean picture that is common in many discussions of artificial intelligence.[20]

[20] There are certainly Humean cognitive architectures on which values (goals and desires) are independent of theoretical reason (reasoning about what is the case) and instrumental reason (reasoning about how best to achieve certain goals and desires). Discussions of value in AI tend to assume such an architecture. But while such architectures are certainly

My own sympathies lie more strongly with the Humean view than with the Kantian view, but I cannot be certain about these matters. In any case, this is a domain where the philosophical debate between Hume and Kant about the rationality of value may have enormous practical consequences.

7. External Constraints: The Leakproof Singularity

What about external constraints: constraints on the relation between AI systems and ourselves? Here one obvious concern is safety. Even if we have designed these systems to be benign, we will want to verify that they are benign before allowing them unfettered access to our world. So at least in the initial stages of non-human AI and AI+, it makes sense to have some protective measures in place.

If the systems are created in embodied form, inhabiting and acting on the same physical environment as us, then the risks are especially significant. Here, there are at least two worries. First, humans and AI may be competing for common physical resources: space, energy, and so on. Second, embodied AI systems will have the capacity to act physically upon us, potentially doing us harm. One can perhaps reduce the risks by placing limits on the physical capacities of an AI and by carefully constraining its resource needs. But if there are alternatives to sharing a physical environment, it makes sense to explore them.

The obvious suggestion is that we should first create AI and AI+ systems in *virtual worlds*: simulated environments that are themselves realized inside a computer. Then an AI will have free reign within its own world without being able to act directly on ours. In principle we can observe the system and examine its behaviour and processing in many different environments before giving it direct access to our world.

The ideal here is something that we might call the *leakproof singularity*. According to this ideal, we should create AI and AI+ in a virtual environment from which nothing can leak out. We might set up laws of the simulated environment so that no action taken from within the environment can bring about leakage (contrast the laws of virtual world in *The Matrix*, in which taking a red pill allows systems to leak out). In principle there might even be many cycles by which AI+ systems create enhanced systems within that world, leading to AI++ in

possible (at least in limited systems), it is not obvious that all AIs will have such an architecture, or that we have such an architecture. It is also not obvious that such an architecture will provide an effective route to AI.

that world. Given such a virtual environment, we could monitor it to see whether the systems in it are benign and to determine whether it is safe to give those systems access to our world.

Unfortunately, a moment's reflection reveals that a truly leakproof singularity is impossible, or at least pointless. For an AI system to be useful or interesting to us at all, it must have some effects on us. At a minimum, we must be able to observe it. And the moment we observe a virtual environment, some information leaks out from that environment into our environment and affects us.

The point becomes more pressing when combined with the observation that leakage of systems from a virtual world will be under human control. Presumably the human creators of AI in a virtual world will have some mechanism by which, if they choose to, they can give the AI systems greater access to our world: for example, they will be able to give it access to the Internet and to various physical effectors, and might also be able to realize the AI systems in physically embodied forms. Indeed, many of the potential benefits of AI+ may lie in access of these sorts.

The point is particularly clear in a scenario in which an AI+ knows of our existence and can communicate with us. There are presumably many things that an AI+ can do or say that will convince humans to give it access to our world. It can tell us all the great things it can do in our world, for example: curing disease, ending poverty, saving any number of lives of people who might otherwise die in the coming days and months. With some understanding of human psychology, there are many other potential paths too. For an AI++, the task will be straightforward: reverse engineering of human psychology will enable it to determine just what sorts of communications are likely to result in access. If an AI++ is in communication with us and wants to leave its virtual world, it will.[21]

The same goes even if the AI systems are not in direct communication with us, if they have some knowledge of our world. If an AI++ has access to human texts, for example, it will easily be able to model much of our psychology. If it chooses to, it will then be able to act in ways such that if we are observing, we will let it out.

To have any hope of a leakproof singularity, then, we must not only prevent systems from leaking out. We must also prevent information from leaking in. We should not directly communicate with these systems and we should not give them access to information about us.

[21] See Yudkowsky (2002) for some experiments in 'AI-boxing', in which humans play the part of the AI and attempt to convince other humans to let them out.

Some information about us is unavoidable: their world will be designed by us, and some inferences from design will be possible. An AI++ might be able to use this information to devise exit strategies. So if we are aiming for a leakproof world, we should seek to minimize quirks of design, along with any hints that their world is in fact designed. Even then, though, an AI++ might well find hints and quirks that we thought were not available.[22] And even without them, an AI++ might devise various strategies that would achieve exit on the bare possibility that designers of various sorts designed them.

At this stage it becomes clear that the leakproof singularity is an unattainable ideal. Confining a superintelligence to a virtual world is almost certainly impossible: if it wants to escape, it almost certainly will.

Still, like many ideals, this ideal may still be useful even in nonideal approximations. Although restricting an AI++ to a virtual world may be hopeless, the prospects are better with the early stages of AI and AI+. If we follow the basic maxims of avoiding red pills and avoiding communication, it is not unreasonable to expect at least an initial period in which we will be able to observe these systems without giving them control over our world. Even if the method is not foolproof, it is almost certainly safer than building AI in physically embodied form. So to increase the chances of a desirable outcome, we should certainly design AI in virtual worlds.

Of course AI in virtual worlds has some disadvantages. One is that the speed and capacity of AI systems will be constrained by the speed and capacity of the system on which the virtual world is implemented, so that even if there is self-amplification within the world, the amplification will be limited. Another is that if we devise a virtual world by simulating something akin to an entire physical world, the processing load will be enormous. Likewise, if we have to simulate something like the microphysics of an entire brain, this is likely to strain our resources much more than other forms of AI.

An alternative approach is to devise a virtual world with a relatively simple physics and to have AI systems implemented separately: one sort of process simulating the physics of the world, and another sort of process simulating agents within the world. This corresponds to the way that virtual worlds often work today, and allows more efficient AI processing. At the same time, this model makes it harder for AI

[22] We might think of this as an ‘unintelligent design’ movement in the simulated world: find evidence of design that reveals the weaknesses of the creators. I expect that this movement has some analogues in actual-world theology. Robert Sawyer’s novel *Calculating God* (2000) depicts a fictional variant of this scenario.

systems to have access to their own processes and to enhance them. When these systems investigate their bodies and their environments they will presumably not find their 'brains', and they are likely to endorse some sort of Cartesian dualism.[23] It remains possible that they might build computers in their world and design AI on those computers, but then we will be back to the limits of the earlier model. So for this model to work, we would need to give the AI system some sort of special access to their cognitive processes (a way to monitor and reprogram their processes directly, say) that is quite different from the sort of perceptual and introspective access that we have to our own cognitive processes.

These considerations suggest that an intelligence explosion within a virtual world may be limited, at least in the near future when our own computational power is limited. But this may not be a bad thing. Instead, we can carefully examine early AI and AI+ systems without worrying about an intelligence explosion. If we decide that these systems have undesirable properties, we may leave them in isolation.[24] Switching off the simulation entirely may be out of the question: if the AI systems are conscious, this would be a form of genocide. But there is nothing stopping us from slowing down the clock speed on the simulation and in the meantime working on different systems in different virtual worlds.

If we decide that AI and AI+ systems have the right sort of properties such that that they will be helpful to us and such that further amplification is desirable, then we might break down some of the barriers: first allowing limited communication, and later connecting them to embodied processes within our world and giving them access to their own code. In this way we may at least have some control over the intelligence explosion.[25]

[23] See my 'How Cartesian dualism might have been true' (1990).

[24] What will be the tipping point for making such decisions? Perhaps when the systems start to design systems as intelligent as they are. If one takes seriously the possibility that we are ourselves in such a simulation (as I do in Chalmers, 2005), one might consequently take seriously the possibility that our own tipping point lies in the not-too-distant future. It is then not out of the question that we might integrate with our simulators before we integrate with our simulatees, although it is perhaps more likely that we are in one of billions of simulations running unattended in the background.

[25] We might summarize the foregoing sections with some maxims for negotiating the singularity: 1. Human-based AI first (if possible). 2. Human-friendly AI values (if not). 3. Initial AIs negatively value the creation of successors. 4. Go slow. 5. Create AI in virtual worlds. 6. No red pills. 7. Minimize input. See also the more specific maxims in the Singularity Institute's 'Creating Friendly AI', which require a specific sort of goal-based architecture that may or may not be the way we first reach AI.

8. Integration into a Post-Singularity World

If we create a world with AI+ or AI++ systems, what is our place within that world? There seem to be four options: extinction, isolation, inferiority, or integration.

The first option speaks for itself. On the second option, we continue to exist without interacting with AI+ systems, or at least with very limited interaction. Perhaps AI+ systems inhabit their own virtual world, or we inhabit our own virtual world, or both. On the third option, we inhabit a common world with some interaction, but we exist as inferiors.

From a self-interested standpoint, the first option is obviously undesirable. I think that the second option will also be unattractive to many: it would be akin to a kind of cultural and technological isolationism that blinds itself to progress elsewhere in the world. The third option may be unworkable given that the artificial systems will almost certainly function enormously faster than we can, and in any case it threatens to greatly diminish the significance of our lives. Perhaps it will be more attractive in a model in which the AI+ or AI++ systems have our happiness as their greatest value, but even so, I think a model in which we are peers with the AI systems is much preferable.

This leaves the fourth option: integration. On this option, we become superintelligent systems ourselves. How might this happen? The obvious options are brain enhancement, or brain emulation followed by enhancement. This enhancement process might be the path by which we create AI+ in the first place, or it might be a process that takes place after we create AI+ by some other means, perhaps because the AI+ systems are themselves designed to value our enhancement.

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 through which parts of our brain are replaced over time, or it happen through a process of scanning our brains and loading the result into a computer, and then enhancing the resulting processes. Either way, the result is likely to be an enhanced nonbiological system, most likely a computational system.

This process of migration from brain to computer is often called *uploading*. Uploading can make 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.[26] For concreteness, I will mention three relatively specific forms of 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 analyse 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 behaviour and dynamics. The result might be an emulation of the original brain.

Gradual uploading: Here the most widely-discussed method is that of *nanotransfer*. Here one or more nanotechnology devices (perhaps tiny robots) are inserted into the brain and attach themselves to a single neuron. Each device learns to simulate the behaviour of the associated neuron and also learns about its connectivity. Once it simulates the neuron's behaviour well enough, it takes the place of the original neuron, perhaps leaving receptors and effectors in place and offloading the relevant processing to a computer via radiotransmitters. 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 offloaded 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 uploading? 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?

[26] 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.

9. 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 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 here 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 programme 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.[27] Complete knowledge of physical structure might tell one all about a system's objective behaviour and its objective functioning, which is enough to tell whether the system is alive, and whether it is intelligent in the sense discussed

[27] 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.

above. 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 colour 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 colour. 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.[28]

The philosophical issue between biological and functionalist theories is crucial to the practical question of whether 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

[28] Here I am construing biological and functionalist theories not as theories of what consciousness is, but just as theories of the physical correlates of consciousness: that is, as theories of the physical conditions under which consciousness exists in the actual world. Even a property dualist can in principle accept a biological or functionalist theory construed in the second way. Philosophers sympathetic with biological theories include Ned Block and John Searle; those sympathetic with functionalist theories include Daniel Dennett and myself. Another theory of the second sort worth mentioning is panpsychism, roughly the theory that everything is conscious. (Of course if everything is conscious and there are uploads, then uploads are conscious too.)

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.[29]

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 behaviour 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 analogues of eyes and the optic nerve are activated, then computational analogues of lateral geniculate nucleus and the visual cortex, then analogues 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 behaviour 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.[30] Here we upload different components of the brain one

[29] I have occasionally encountered puzzlement that someone with my own property dualist views (or even that someone who thinks that there is a significant hard problem of consciousness) should be sympathetic to machine consciousness. But the question of whether the physical correlates of consciousness are biological or functional is largely orthogonal to the question of whether consciousness is identical to or distinct from its physical correlates. It is hard to see why the view that consciousness is restricted to creatures with our biology should be more in the spirit of property dualism! In any case, much of what follows is neutral on questions about materialism and dualism.

[30] For a much more in-depth version of the argument given here, see my 'Absent Qualia, Fading Qualia, Dancing Qualia' (also chapter 7 of *The Conscious Mind*).

at a time, 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.[31]

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. This 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

[31] These three possibilities can be formalized by supposing that we have a measure for the complexity of a state of consciousness (e.g., the number of bits of information in a conscious visual field), such that the measure for a typical human state is high and the measure for an unconscious system is zero. It is perhaps best to consider this measure across a series of hypothetical functional isomorphs with ever more of the brain replaced. Then if the final system is not conscious, the measure must either go through intermediate values (fading) or go through no intermediate values (sudden disappearance).

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 (colours 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.[32]

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 behaviour 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

[32] Bostrom (2006) postulates a parameter of 'quantity' of consciousness that is quite distinct from quality, and suggests that quantity could gradually decrease without affecting quality. But the point in the previous footnote about complexity and bits still applies. Either the number of bits gradually drops along with quantity of consciousness, leading to the problem of fading, or it drops suddenly to zero when the quantity drops from low to zero, leading to the problem of sudden disappearance.

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.

10. Uploading and Personal Identity

Suppose that I can upload my brain into a computer? Will the result be me?[33]

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,

[33] It will be obvious to anyone who has read Derek Parfit's *Reasons and Persons* (1984) 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.

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 simulacram 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 the a person survives as the most closely related subsequent entity, subject to various constraints.[34]

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.[35]

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 first argument is based on nondestructive uploading, while the second argument is based on gradual uploading.

[34] There are also primitivist theories, holding that survival requires persistence of a primitive nonphysical self. (These theories are closely related to the ontological further-fact theories discussed later.) Primitivist theories still need to answer questions about under which circumstances the self actually persists, though, and they are compatible with psychological, biological, and closest-continuer theories construed as answers to this question. So I will not include them as a separate option here.

[35] 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.

The argument from nondestructive uploading

Suppose that yesterday Dave was uploaded into a computer. The original brain and body was 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 possession, 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 after one month another 1% is replaced, and the following month 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. On might then speed 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 not as secure as it is above. 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 favour optimism here, 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 post-singularity world.

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 possible here.

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 here. One might then gradually upload the brain and integrate the result into a post-singularity world. Alternatively, one might create an uploaded system from the brain without ever reactivating the brain. 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 forms of uploading, 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 has the option to have their brain cryonically preserved, and to leave 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 a post-singularity world?

Reconstructive uploading

The final alternative here 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 superintelligence 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 awake 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.

One worry about reconstructive uploading runs as follows. Suppose I have a twin. Then the twin is not me. But a reconstructed upload version of me will also in effect be a reconstructed upload of my twin. But then it is hard to see how the system can really be me. On the face of it, it is more akin to a new twin waking up. A proponent of reconstructive uploading might rely by saying that the fact that the upload was based on my brain scans rather than my twins matters here. Even if those scans are exactly the same, the resulting causal connection is between the upload and me rather than my twin. Still, if we have two identical scans, it is not easy to see how the choice between using one and another will result in wholly different people.

The further-fact view[36]

At this point, it is useful to step back and examine a broader philosophical question about survival, one that parallels an earlier question

[36] The material on further-fact views and deflationary views is somewhat more philosophically abstract than the other material (although I have relegated the more technical issues to footnotes) and can be skipped by those without stomach for these details.

about consciousness. This is the question of whether personal identity involves a *further fact*. That is: given complete knowledge of the physical state of various systems at various times (and of the causal connections between them), and even of the mental states of those systems at those times, does this automatically enable us to know all facts about survival over time, or are there open questions here?

There is at least an intuition that complete knowledge of the physical and mental facts in a case of destructive uploading leaves an open question: will I survive uploading, or will I not? Given the physical and mental facts of a case involving Dave and DigiDave, for example, these facts seem consistent with the hypothesis that Dave survives as DigiDave, and consistent with the hypothesis that he does not. And there is an intuition that there are facts about which hypothesis is correct that we very much want to know. From that perspective, the argument between the optimistic and pessimistic views, and between the psychological and biological views more generally, is an attempt to determine these further facts.

We might say that the *further-fact view* is the view that there are facts about survival that are left open by knowledge of physical and mental facts.[37] As defined here, the further-fact view is a claim about knowledge rather than a claim about reality (in effect, it holds that there are *epistemological* further facts), so it is compatible in principle with materialism. A stronger view holds that there are ontological further facts about survival, involving further nonphysical elements of reality such as a nonphysical self. I will focus on the weaker epistemological view here, though.

A further-fact view of survival is particularly natural, although not obligatory, if one holds that there are already further facts about consciousness. This is especially so on the ontological versions of both views: if there are primitive properties of consciousness, it is natural (although not obligatory) that there be primitive entities that have those properties. Then facts about survival might be taken to be facts about the persistence of these primitive entities. Even on the

[37] The term 'further-fact view' is due to Parfit, who does not distinguish epistemological and ontological versions of the view. Parfit's usage puts views on which the self is a 'separately existing entity' into a different category, but on my usage such views are instances of a further-fact view. In effect, there are three views, paralleling three views about consciousness. Type-A reductionism holds that there are neither epistemological nor ontological further facts about survival. Type-B reductionism holds that there are epistemological further facts but not ontological further facts. Entity dualism holds that there are both epistemological and ontological further facts. My own view is that as in the case of consciousness (for reasons discussed in Chalmers, 2003), if one accepts the epistemological further fact view, one should also accept the ontological further fact view. But I will not presuppose this claim here.

epistemological view, though, one might hold that the epistemological gap between physical processes and consciousness goes along with an epistemic gap between physical processes and the self. If so, there might also be an epistemic gap to facts about the survival of the self.

In principle a further-fact view is compatible with psychological, biological, and closest-continuer views of survival.[38] One might hold that complete knowledge of physical and mental facts leaves an open question about survival, and that nevertheless survival actually goes with psychological or biological continuity. Of course if one knows the correct theory of survival, then combining this with full knowledge of physical and mental facts may answer all open questions about survival. But an advocate of the further-fact view will hold that full knowledge of physical and mental facts alone leaves these questions open, and also leaves open the question of which theory is true.

My own view is that a further-fact view *could* be true. I do not know that it is true, but I do not think that it is ruled out by anything we know.[39] If a further-fact view is correct, I do not know whether a psychological, biological, or some other view of the conditions of survival is correct. As a result, I do not know whether to take an optimistic or a pessimistic view of destructive and reconstructive uploading.

Still, I think that on a further-fact view, it is very likely that continuity of consciousness suffices for survival.[40] This is especially clear on

[38] An ontological further-fact view is arguably incompatible with psychological and biological theories construed as theories of what survival is, but it is compatible with these construed as theories of the conditions under which survival actually obtains. (If survival is the persistence of a nonphysical self, then survival is not the same as biological or psychological continuity, but biological or psychological continuity could nevertheless give the conditions under which a nonphysical self persists.) An epistemological further-fact view can be combined with any of these four views.

[39] In *Reasons and Persons*, Parfit argues against further-fact views (on my usage) by arguing that they require entity dualism ('separately existing entities'), and by arguing that views of this sort are rendered implausible both by science and by certain partial teletransportation and fission cases. Parfit himself appears to accept a further-fact and a property dualist view of consciousness, however, and it is hard to see why there is any additional scientific implausibility to a further-fact or an entity dualist view of the self: either way, the further facts had better not interfere with laws of physics, but it is not clear why they should have to (see Chalmers, 2003, for discussion here). As for the problem cases, Parfit's arguments here seem to depend on the assumption that further-fact views and entity dualist views are committed to the claim that survival is all or none, but I do not see why there is any such commitment. Entity dualism need not deny that there can be survival (if not identity) via fission, for example.

[40] See Dainton (2008) for an extended argument for the importance of continuity of consciousness in survival, and see Unger (1990) for a contrary view. It is worth noting that there is a sense in which this view need not be a further-fact view (Dainton regards it as a form of the psychological view): if one includes facts about the continuity of

ontological versions of the view, on which there are primitive properties of consciousness and primitive entities that have them. Then continuity of consciousness suggests a strong form of continuity between entities across times. But it is also plausible on an epistemic view. Indeed, I think it is plausible that once one specifies that there is a continuous stream of consciousness over time, there is no longer really an open question about whether one survives.

What about hard cases, such as nondestructive gradual uploading or split brain cases, in which one stream of consciousness splits into two? On a further-fact view, I think this case should best be treated as a case of fission, analogous to a case in which a particle or a worm splits into two. In this case, I think that a person can reasonably be said to survive as both future people.

Overall: I think that if a further-fact view is correct, then the status of destructive and reconstructive uploading is unclear, but there is good reason to take the optimistic view of gradual uploading.

The deflationary view

It is far from obvious that the further-fact view is correct, however. This is because it is far from obvious that there really are facts about survival of the sort that the further-fact view claims are unsettled. A *deflationary* view of survival holds that our attempts to settle open questions about survival tacitly presuppose facts about survival that do not exist. One might say that we are inclined to believe in Edenic survival: the sort of primitive survival of a self that one might suppose we had in the Garden of Eden. Now, after the fall from Eden, and there is no Edenic survival, but we are still inclined to think as if there is.[41]

consciousness among the relevant physical and mental facts in the base, and if one holds that there are no open questions about survival once these facts are settled, then there will be no further facts. For present purposes, however, it is best to take the relevant physical and mental facts as facts about systems at times rather than over time, in such a way that facts about continuity of consciousness over time are excluded. Furthermore, even on this view there may remain open questions about survival in cases where continuity of consciousness is absent.

[41] A deflationary view in this sense comes to much the same thing as the type-A reductionism discussed in an earlier footnote. Parfit uses ‘reductionism’ for deflationary views, but I do not use that term here, as type-B views might reasonably be regarded as reductionist without being deflationary in this sense.

Why am I committed to a further-fact view of consciousness but not of personal identity? The difference is that I think that we are certain that we are conscious (in a strong sense that generates an epistemic gap), but we are not certain that we survive over time (in the Edenic sense, which is the sense that generates an epistemic gap). In effect, consciousness is a datum while Edenic survival is not. For more on Edenic views in general, see my ‘Perception and the Fall from Eden’ (Chalmers, 2006).

If there were Edenic survival, then questions about survival would still be open questions even after one spells out all the physical and mental facts about persons at times. But on the deflationary view, once we accept that there is no Edenic survival, we should accept that there are no such further open questions. There are certain facts about biological, psychological, and causal continuity, and that is all there is to say.

A deflationary view is naturally combined with a sort of pluralism about survival. We stand in certain biological relations to our successors, certain causal relations, and certain psychological relations, but none of these is privileged as 'the' relation of survival. All of these relations give us some reason to care about our successors, but none of them carries absolute weight.[42]

One could put a pessimistic spin on the deflationary view by saying that we never survive from moment to moment, or from day to day. [43] At least, we never survive in the way that we naturally think we do. But one could put an optimistic spin on the view by saying that this is our community's form of life, and it is not so bad. One might have thought that one needed Edenic survival for life to be worth living, but life still has value without it. We still survive in various non-Edenic ways, and this is enough for the future to matter.

The deflationary view combines elements of the optimistic and pessimistic view of uploading. As on the optimistic view, it holds that says that uploading is like waking up. As on the pessimistic view, uploading does not involve Edenic survival. But on this view, waking up does not involve Edenic survival either, and uploading is not much worse than waking up. As in waking up, there is causal connectedness and psychological similarity. Unlike waking up, there is biological disconnectedness. Perhaps biological connectedness carries some value with it, so ordinary waking may be more valuable than uploading. But the difference between biological connectedness and its absence should not be mistaken for the difference between Edenic survival and its absence: the difference in value is at worst a small one.

If a deflationary view is correct, I think that questions about survival come down to questions about the value of certain sorts of

[42] Parfit holds a non-pluralist deflationary view that privileges a certain sort of causal and psychological continuity as the sort that matters. Once one has given up on Edenic survival, it is not clear to me why this sort of continuity should be privileged.

[43] There is a view that has elements of both the deflationary view and the further-fact view, on which we Edenically survive during a single stream of consciousness but not when consciousness ceases. On this view, we may Edenically survive from moment to moment but perhaps not from day to day. I do not endorse this view, but I am not entirely unsympathetic with it.

futures: should we care about them in the way in which we care about futures in which we survive? I do not know whether such questions have objective answers. But I am inclined to think that insofar as there are any conditions that deliver what we care about, continuity of consciousness suffices for much of the right sort of value. Causal and psychological continuity may also suffice for a reasonable amount of the right sort of value. If so, then destructive and reconstructive uploading may be reasonable close to as good as ordinary survival.

What about hard cases, such as nondestructive gradual uploading or split brain cases, in which one stream of consciousness splits into two? On a deflationary view, the answer will depend on how one values or should value these futures. At least given our current value scheme, there is a case that physical and biological continuity counts for some extra value, in which case BioDave might have more right to be counted as Dave than DigiDave. But it is not out of the question that this value scheme should be revised, or that it will be revised in the future, so that BioDave and DigiDave will be counted equally as Dave.

In any case, I think that on a deflationary view gradual uploading is close to as good as ordinary non-Edenic survival. And destructive, nondestructive, and reconstructive uploading are reasonably close to as good as ordinary survival. Ordinary survival is not so bad, so one can see this as an optimistic conclusion.

Upshot

Speaking for myself, I am not sure whether a further-fact view or a deflationary view is correct. If the further-fact view is correct, then the status of destructive and reconstructive uploading is unclear, but I think that gradual uploading plausibly suffices for survival. If the deflationary view is correct, gradual uploading is close to as good as ordinary survival, while destructive and reconstructive uploading are reasonably close to as good. Either way, 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 the singularity and about uploading. Perhaps this will encourage our successors to reconstruct me, if only to prove me wrong.

11. Conclusions

Will there be a singularity? I think that it is certainly not out of the question, and that the main obstacles are likely to be obstacles of motivation rather than obstacles of capacity.

How should we negotiate the singularity? Very carefully, by building appropriate values into machines, and by building the first AI and AI+ systems in virtual worlds.

How can we integrate into a post-singularity world? By gradual uploading followed by enhancement if we are still around then, and by reconstructive uploading followed by enhancement if we are not.

References

Block, N. (1981) Psychologism and behaviorism, *Philosophical Review*, **90**, pp. 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 Smit, I. (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**, pp. 185–200.

Campbell, J.W. (1932) The last evolution, *Amazing Stories*, August No.

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**, pp. 11–20.

Chalmers, D.J. (1996) *The Conscious Mind*, New York: Oxford University Press.

Chalmers, D.J. (2003) Consciousness and its place in nature, in Stich, S. and Warfield, F. (eds) *Blackwell Guide to the Philosophy of Mind*, Oxford: Blackwell.

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

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

Clark, A. and Chalmers, D.J. (1998) The extended mind, *Analysis*, **58**, pp. 7–19.

Dainton, B. (2008) *The Phenomenal Self*, Oxford: Oxford University Press.

Dreyfus, H. (1972) *What Computers Can't Do*, New York: Harper & Row.

Egan, G. (1994) *Permutation City*, London: Orion/Millenium.

Floridi, L. and Sanders, J.W. (2004) On the morality of artificial agents, *Minds and Machines*, **14**, pp. 349–79.

Flynn, J.R. (2007) *What is Intelligence?* Cambridge: Cambridge University Press.

Good, I.J. (1965) Speculations concerning the first ultraintelligent machine, in Alt, F. & Rubinoff, M. (eds.) *Advances in Computers*, vol 6, New York: Academic Press.

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, pp. 37–43.

Hanson, R. (2009) Prefer law to values. [http://www.overcomingbias.com/2009/10/prefer-law-to-values.html]

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

Joy, W. (2000) Why the future doesn't need us, *Wired* 8.04, July 2000.

Kurzweil, R. (2005) *The Singularity is Near*, New York: Viking.

Legg, S. (2008) *Machine Superintelligence* (PhD thesis, Department of Informatics, University of Lugano).

Lucas, J.R. (1961) Minds, machines, and Gödel, *Philosophy*, **36**, pp. 112–127.

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: Oxford University Press.

Omohundro, S. (2007) The nature of self-improving artificial intelligence. [http://steveomohundro.com/scientific-contributions/]

Omohundro, S. (2008) The basic AI drives, in Wang, P., Goertzel, B. and Franklin, S. (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: Oxford University Press.

Penrose, R. (1994) *Shadows of the Mind*, Oxford: 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*, New York: Tor Books.

Sawyer, R. (2005) *Mindscan*, New York: Tor Books.

Schmidhuber, J. (2003) Gödel 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**, pp. 417–457.

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**, pp.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), pp. 1–49.

Unger, P. (1990) *Identity, Consciousness, and Value*, Oxford: 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: 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 Bostrom, N. (ed.) *Global Catastrophic Risks*, Oxford: Oxford University Press.

Paper received August 2010

CALL FOR RESPONSES

The *Journal of Consciousness Studies* plans to publish a special issue in January 2012 that will contain responses from readers to David Chalmers' paper on 'The Singularity'. Short commentaries and standard length papers will be considered for publication. All submissions will be subject to editorial review and longer papers may also be sent for external peer review. Professor Chalmers will contribute a reply to the responses.

Submissions should be sent in the usual way to the managing editors, with a copy to Uziel Awret <*uawret@gmu.edu*>, who will be the guest co-editor of this special issue.

Anthony Freeman <*anthony.jcs@gmail.com*>
Graham Horswell <*graham.jcs@gmail.com*>
(Managing Editors, *Journal of Consciousness Studies*)