The Singularity: A Reply

David J. Chalmers

1 Introduction

I would like to thank the authors of the 26 contributions to this symposium on my article "The Singularity: A Philosophical Analysis". I learned a great deal from the reading their commentaries. Some of the commentaries engaged my article in detail, while others developed ideas about the singularity in other directions. In this reply I will concentrate mainly on those in the first group, with occasional comments on those in the second.

A singularity (or an intelligence explosion) is a rapid increase in intelligence to superintelligence (intelligence of far greater than human levels), as each generation of intelligent systems creates more intelligent systems in turn. The target article argues that we should take the possibility of a singularity seriously, and argues that there will be superintelligent systems within centuries unless certain specific defeating conditions obtain.

I first started thinking about the possibility of an intelligence explosion as a graduate student in Doug Hofstadter's AI lab at Indiana University in the early 1990s. Like many, I had the phenomenology of having thought up the idea myself, though it is likely that in fact I was influenced by others. I had certainly been exposed to Hans Moravec's 1988 book *Mind Children* in which the idea is discussed, for example. I advocated the possibility vigorously in a discussion with the AI researchers Rod Brooks and Doug Lenat and the journalist Maxine McKew on the Australian TV show *Lateline* in 1996. I first discovered the term "singularity" on Eliezer Yudkowsky's website in 1997, where I also encountered the idea of a combined intelligence and speed explosion for the first time. I was fascinated by the idea that all of human history might converge to a single point, and took that idea to be crucial to the singularity per se; I have been a little disappointed that this idea has receded in later discussions.

Since those early days I have always thought that the intelligence explosion is a topic that is both practically and philosophically important, and I was pleased to get a chance to develop these

ideas in a talk at the 2009 Singularity Summit and then in this paper for *JCS*. Of course the main themes in the target article (the intelligence explosion, negotiating the singularity, uploading) have all been discussed at length before, but often in non-academic forums and often in non-rigorous ways. One of my aims in the target article was to put the discussion on a somewhat clearer and more rigorous analytic footing than had been done in previously published work. Another aim was to help bring the issues to an audience of academic philosophers and scientists who may well have much to contribute.

In that respect I am pleased with the diversity of the commentators. There are nine academic philosophers (Nick Bostrom, Selmer Bringsjord, Richard Brown, Joseph Corabi, Barry Dainton, Daniel Dennett, Jesse Prinz, Susan Schneider, Eric Steinhart) and eight AI researchers (Igor Aleksander, Ben Goertzel, Marcus Hutter, Ray Kurzweil, Drew McDermott, Jurgen Schmidhuber, Murray Shanahan, Roman Yampolskiy). There are also representatives from cultural studies (Arkady Plotnitsky), cybernetics (Francis Heylighen), economics (Robin Hanson), mathematics (Burton Voorhees), neuroscience (Susan Greenfield), physics (Frank Tipler), psychiatry (Chris Nunn), and psychology (Susan Blackmore), along with two writers (Damien Broderick and Pamela McCorduck) and a researcher at the Singularity Institute (Carl Shulman).

Of the 26 articles, about four are wholeheartedly pro-singularity, in the sense of endorsing the claim that a singularity is likely: those by Hutter, Kurzweil, Schmidhuber, and Tipler. Another eleven or so seem to lean in that direction or at least discuss the possibility of a singularity sympathetically: Blackmore, Broderick, Corabi and Schneider, Dainton, Goertzel, McCorduck, Shanahan, Steinhart, Shulman and Bostrom, Voorhees, and Yampolskiy. Three come across as mildly skeptical, expressing a deflationary attitude toward the singularity without quite saying that it will not happen: Dennett, Hanson, and Prinz. And about seven express wholehearted skepticism: Aleksander, Bringsjord, Greenfield, Heylighen, McDermott, Nunn, and Plotnitsky.

About twelve of the articles focus mainly on whether there will or will not be a singularity or whether there will or will not be AI: the seven wholehearted skeptics along with McCorduck, Prinz, Schmidhuber, Shulman and Bostrom, and Tipler. Three articles focus mainly on how best to negotiate the singularity: Goertzel, Hanson, and Yampolskiy. Three focus mainly on the character and consequences of a singularity: Hutter, Shanahan, and Voorhees. Three focus mainly on consciousness: Brown, Dennett, and Kurzweil. Three focus mainly on personal identity: Blackmore, Corabi and Schneider, and Dainton. Two focus on connections to other fields: Broderick and Steinhart. Numerous other issues are discussed along the way: for example, uploading (Greenfield, Corabi and Schneider, Plotnitsky) and whether we are in a simulation (Dainton, Prinz, Shulman

and Bostrom).

I will not say much about the connections to other fields: Broderick's connections to science fiction and Steinhart's connections to theology. These connections are fascinating, and it is clear that antecedents of many key ideas have been put forward long ago. Still, it is interesting to note that very few of the science fiction works discussed by Broderick (or the theological works discussed by Steinhart) focus on a singularity in the sense of a recursive intelligence explosion. Perhaps Campbell's short story "The Last Evolution" comes closest here: here humans defend themselves from aliens by designing systems that design ever smarter systems that finally have the resources to win the war. There is an element of this sort of recursion in some works by Vernor Vinge (originator of the term "singularity"), although that element is smaller than one might expect. Most of the other works discussed by Broderick focus simply on greater-than-human intelligence, an important topic that falls short of a full singularity as characterized above.

At least two of the articles say that it is a bad idea to think or talk much about the singularity as other topics are more important: environmental catastrophe followed by nuclear war (McDermott) and our dependence on the Internet (Dennett). The potential fallacy here does not really need pointing out. That it is more important to talk about topic B than topic A does not entail that it is unimportant to talk about topic A. It is a big world, and there are a lot of important topics and a lot of people to think about them. If there is even a 1% chance that there will be a singularity in the next century, then it is pretty clearly a good idea for at least a hundred people (say) to be thinking hard about the possibility now. Perhaps this thinking will not significantly improve the outcome, but perhaps it will; we will not even be in a position to make a reasoned judgment about that question without doing a good bit of thinking first. That still leaves room for thousands to think about the Internet and for millions to think about the environment, as is already happening.

This reply will largely follow the shape of the original article. After starting with general considerations, I will spend the most time on the argument for an intelligence explosion, addressing various objections and analyses. In later sections I discuss issues about negotiating the singularity, consciousness, uploading, and personal identity.

2 The Argument for an Intelligence Explosion

The target article set out an argument for the singularity as follows.

(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).

Here AI is human-level artificial intelligence, AI+ is greater-than-human-level artificial intelligence, and AI++ is far-greater-than-human-level artificial intelligence (as far beyond smartest humans as humans are beyond a mouse). "Before long" is roughly "within centuries" and "soon after" is "within decades", though tighter readings are also possible. Defeaters are anything that prevents intelligent systems from manifesting their capacities to create intelligent systems, including situational defeaters (catastrophes and resource limitations) and motivational defeaters (disinterest or deciding not to create successor systems).

The first premise is an equivalence premise, the second premise is an extension premise, and the third premise is an amplification premise. The target article gave arguments for each premise: arguments from brain emulation and from evolution for the first, from extendible technologies for the second, and from a proportionality thesis for the third. The goal was to ensure that if someone rejects the claim that there is a singularity, they would have to be clear about which arguments and which premises they are rejecting.

This goal was partially successful. Of the wholehearted singularity skeptics, three (Bringsjord, McDermott, and Plotnitsky) engage these arguments in detail. The other four (Aleksander, Greenfield, Heylighen, and Nunn) express their skepticism without really engaging these arguments. The three mild skeptics (Dennett, Hanson, and Prinz) all engage the arguments at least a little.

Greenfield, Heylighen, and Nunn all suggest that intelligence is not just a matter of information-processing, and focus on crucial factors in human intelligence that they fear will be omitted in AI: understanding, embodiment, and culture respectively. Here it is worth noting that nothing in the original argument turns on equating intelligence with information-processing. For example, one can equate intelligence with understanding, and the argument will still go through. The emulation and evolution arguments still give reason to think that we can create AI with human-level understanding, the extendibility point gives reason to think that AI can go beyond that, and the explosion point gives reasons to think that systems with greater understanding will be able to create further systems with greater understanding still.

As for embodiment and culture, insofar as these are crucial to intelligence, AI can simply build them in. The arguments apply equally to embodied AI in a robotic body surrounded by other

intelligent systems. Alternatively, one can apply the emulation argument not just to an isolated system but to an embedded system, simulating its physical and cultural environment. This may require more resources than simulating a brain alone, but otherwise the arguments go through as before. Heylighen makes the intriguing point that an absence of values may serve as a resource limitation that slows any purported intelligence explosion to a convergence; but the only reason he gives is that values require a rich environmental context, so this worry is not a worry for AI that exists in a rich environmental context.

Aleksander makes a different argument against the singularity. First, knowledge of existing AI suggests that it is far off, and second, designing a system with greater than human intelligence requires complete self-knowledge and a complete cognitive psychology. On the first point, the distance between current AI and human-level AI may cast doubt on claims about human-level AI within years and perhaps within decades, but it does not do much to cast doubt on my arguments for the premise of AI within centuries. On the second point, it is far from clear that complete self-knowledge is required here. Brute force physical emulation could produce human-level AI without much theoretical understanding; the theoretical understanding could come later once one can experiment easily with the emulation. And paths to AI such as artificial evolution and machine learning have the potential to take a route quite different from that of human intelligence, so that again self-knowledge is not required.

Dennett engages with my arguments just for a moment. He expresses a good deal of skepticism about the singularity, but his only concrete objection is that any measure of intelligence that humans devise will be so anthropocentric that it will distort what it contrives to measure. To which one can respond: an anthropocentric measure will at least capture something that we humans care about, so that if the argument is sound, the conclusion will still be one of enormous significance. One can also note that even we use a less anthropocentric measure of intelligence that humans do not devise, the argument may still go through. Either way, Dennett does not give any reason to deny any of the argument's premises.

Hanson expresses a different deflationary attitude to the argument, saying that its conclusion is too weak to be significant (!). The reason he gives is that there exist other sources of intelligence growth in our environment—the Flynn effect (IQ scores increase by three points per decade) and cultural development—and left to their own devices these will themselves lead to arbitrary increases in intelligence and to AI++. Here I think Hanson ignores the crucial temporal element in the conclusion. Perhaps the Flynn effect might lead to AI++ given enough time, but within centuries the level of AI+ (90 IQ points in 300 years?) is the best we can hope for. Economic

and cultural development is perhaps more powerful, but again there is little reason to think it can yield human:mouse increases in community intelligence over centuries. Hanson himself observes that we have seen that sort of increase over the last 100,000 years. He goes on to suggest that that faster growth combined with the Flynn effect can be expected to yield the same sort of increase over centuries, but he gives no argument for this enormous speed-up. Prima facie, these source of growth can at best be expected to lead to AI+ levels of human intelligence within centuries, not to AI++ levels.

Prinz invokes Bostrom's simulation argument to suggest that if a singularity is likely, it has probably happened already: we are probably simulated beings ourselves (as there will be many more simulated beings than nonsimulated beings), and these will be created by superintelligent beings. He then suggests that our creators are likely to destroy us before we reach AI++ and threaten them—so it is predictable that one of my defeaters (catastrophe) will occur. I am not unsympathetic with Bostrom's argument, but I think there are some holes in Prinz's use of it. In particular, it is quite likely that the great majority of simulations in the history of the universe will be unobserved simulations. Just as one finds with existing simulations, it is reasonable to expect that superintelligent beings will run millions of universe simulations at once (leaving them to run overnight, as it were) in order to gather statistics at the end and thereby and do science. They may well be able to devise anti-leakage mechanisms that make unobserved simulations cause little danger to them (the greatest danger of leakage comes from observed simulations, as discussed in the target paper). If so, our future is unthreatened. Prinz also suggests that we may destroy ourselves for relatively ordinary reasons (disease, environmental destruction, weapons) before reaching a singularity. Perhaps so, perhaps not, but here in any case I am happy enough with the limited conclusion that *absent defeaters* there will be a singularity.

Plotnitsky suggests that complexity may undermine my arguments from emulation and evolution for premise 1. If the brain is sufficiently complex, we may not be able to emulate it before long. And if evolution requires sufficient complexity, artificial evolution may be impossible in the relevant timeframe.

Here again, I was counting on the centuries time frame to make the premises of the arguments more plausible. In a recent report grounded in neurobiological evidence, Sandberg and Bostrom suggest that brain emulation with the needed degree of complexity may be possible within decades. I am a little skeptical of that prediction, but I do not see much reason to think that the intelligent processes in the brain involve complexity that cannot be emulated within centuries. I think it is more than likely that by the end of this century we will be able to simulate individual cells, their

connections and their plasticity very well. Another century should be more than enough to capture any remaining crucial internal structure in neurons and remaining features of global architecture relevant to intelligent behaviour. Throw in another century for embodiment and environmental interactions, and I think we reach the target on conservative assumptions. Of course these seat-of-the-pants estimates are not science, but I think they are reasonable all the same.

The role of complexity in the evolutionary argument is a little trickier, due to the enormous time scales involved: hundreds of millions of years of evolution culminating in a brain certainly involve much more complexity than the brain itself! On the other hand, unlike the emulation argument, the evolutionary argument certainly does not require an emulation of the entire history of evolution. It just requires a process that is relevantly similar to that history in that it produces an intelligent system. But it can certainly do so via a quite different route. The question now is whether the evolution of intelligence *essentially* turns on complexity of a level that we cannot hope to replicate artificially within centuries. I do not have a knockdown argument that this is impossible (Shulman and Bostrom's discussion of evolutionary complexity is worth reading here), but I would be surprised. Artificial evolution already has some strong accomplishments within its first few decades.

Shulman and Bostrom focus on a different objection to the evolutionary argument. It may be that evolution is extraordinarily hard, so that intelligence evolves a very small number of times in the universe. Due to observer selection effects, we are among them. But we are extraordinary lucky. Normally one can reasonably say that extraordinary luck is epistemically improbable, but not so in this case. Because of this, the inference from the fact that evolution produced intelligence to the claim that artificial evolution can be expected to produce intelligence is thrown into question.

I do not have much to add to Shulman and Bostrom's thorough discussion of this issue. I am fairly sympathetic with their "self-indication assumption", which raises the rational probability of that the evolution of intelligence is easy, on the grounds that there will be many more intelligent beings under the hypothesis that the evolution of intelligence is easy. I also agree with them that parallel evolution of intelligence (e.g. in octopuses) gives evidence for "evolution is easy" that is not undercut by observer effects, although as they note, the issues are not cut and dried.

All in all, I think that the evolutionary argument is worth taking seriously. But the worries about complexity and about observer effects raise enough doubts about it that it is probably best to give it a secondary role, with the argument from emulation playing the primary role. After all, one sound argument for the equivalence premise is enough to ensure the truth of a conclusion.

Bringsjord appeals to computational theory to argue that my argument is fatally flawed. His main argument is that Turing-level systems (M_2) can never create super-Turing-level systems (M_3), so that starting from ordinary AI we can never reach AI++, and one of my premises must be false. A secondary argument is that humans are at level M_3 and AI is restricted to level M_2 so that AI can never reach human level. I think there are multiple problems with his arguments, two problems pertaining especially to the first argument and two problems pertaining especially to the second.

First: intelligence does not supervene on computational class. So if we assume humans are themselves in M_2, it does not follow that systems in M_3 are required for AI++. We know that there is an enormous range of intelligence (as ordinarily conceived) within class M_2: from mice to apes to ordinary people to Einstein. Bringsjord gives no reason to think that any human is close to the upper level of M_2. So there is plenty of room within M_2 for AI+ and AI++, or at least Bringsjord's argument gives no reason to think not. His argument in effect assumes a conception of intelligence (as computational class) that is so far from our ordinary notion that it has no bearing on arguments involving relatively ordinary notions of intelligence.

Second: if level M_3 processes are possible in our world, then it is far from obvious that level-M_2 AI could not create it. Computational theory establishes only that a very limited sort of "creation" is impossible: roughly, creation that exploits only the AI's internal recourses plus digital inputs and outputs. But if the AI has the ability to observe and manipulate noncomputational processes in nature, then there is nothing to stop the AI from producing a series of Turing-computable outputs that themselves lead (via the mediation of external physical processes) to the assembly of a super-Turing machine.

Third: There is little reason to think that humans are at level M_3. Bringsjord gives no argument for that claim here. Elsewhere he has appealed to Gödelian arguments to make the case. I think that Gödelian arguments fail; for an analysis, see Chalmers 1995.

Fourth: even if humans are at level M_3, there is then little reason to think that AI must be restricted to level M_2. If we are natural systems, then presumably there are noncomputational processes in nature that undergird our intelligence. There is no obvious reasons why we could not exploit those in an artificial system, thereby leading to AI at level M_3. Bringsjord appeals to an unargued premise saying that our AI-creating resources are limited to level M_2, but if we are at M_3 there is little reason to believe this premise.

McDermott holds that my argument for premise 3 fails and that more generally there is little reason to think that AI++ is possible. He says first "the argument is unsound, because a series of increases from AI_n to AI_{n+1}, each exponentially smaller than the previous one, will reach a limit".

Here McDermott appears to overlook the definitions preceding the argument, which stipulate that there is a positive δ such that the difference in intelligence between AI_n and AI_{n+1} is at least δ for all n. Of course it is then possible to question the key premise saying that if there is AI_n there will be AI_{n+1} (itself a consequence of the proportionality theses), but McDermott's initial formal point gives no reason to do so.

McDermott goes on to question the key premise by saying that it relies on an appeal to extendible methods, and that no method is extendible without limits. Here he misconstrues the role of extendible methods in my argument. They play a key role in the case for premise 2, where extendibility is used to get from AI to AI+. Indefinite extendibility is not required here, however: small finite extendibility is enough. And in the case for premise 3, extendibility plays no role. So McDermott's doubts about indefinite extendibility have no effect on my argument.

I certainly do not think that a single extendible method is likely to get us from AI all the way to AI++. It is likely that as systems get more intelligent, they will come up with new and better methods all the time, as a consequence of their greater intelligence. In the target article, McDermott's worries about convergence are discussed under the label "diminishing returns" (a "structural obstacle") in section 4. Here I argue that small differences in design capacity tend to lead to much greater differences in the systems designed, and that a "hill-leaping" process through intelligence space can get us much further than mere hill-climbing (of which ordinary extendibility is an instance). McDermott does not address this discussion.

The issues here are certainly nontrivial. The key issue is the "proportionality thesis" saying that among systems of certain class, an increase of δ in intelligence will yield an increase of δ in the intelligence of systems that these systems can design. The evaluation of that thesis requires careful reflection on the structure of intelligence space. It is a mild disappointment that none of the commentators focused on the proportionality thesis and the structure of intelligence space. I am inclined to think that the success of my arguments (and indeed the prospects for a full-blown singularity) may turn largely on those issues.

Because these issues are so difficult, I do not have a knockdown argument that AI++ is possible. But just as McDermott says that it would be surprising if the minimal level of intelligence required for civilization were also the maximal possible level, I think it would be surprising if it were close to the maximal possible level. Computational space is vast, and it would be extremely surprising if the meanderings of evolution had come close to exhausting its limits.

Furthermore: even if we are close to the algorithmic limits, speed and population explosions alone might produce a system very close to AI++. Consider a being that could achieve in a single

second as much as the Manhattan project achieved with hundreds of geniuses over five years. Such a being would be many levels beyond us: if not human:mouse, than at least human:dog! If we assume hundreds of AI+ (which McDermott allows) rather than hundreds of geniuses, then the difference is all the greater. McDermott's reasons give little reason to doubt that this sort of system is possible. So even on McDermott's assumption, something quite close to a singularity is very much in prospect.

On the other side, two of the articles vigorously advocate a singularity using arguments different from mine. Tipler argues that physics (and in particular the correct theory of quantum gravity) makes a singularity inevitable. An indefinitely expanding universe leads to contradictions, as it requires black hole evaporations, which violate quantum-mechanical unitarity. So the universe must end in collapse. A collapsing universe contradicts the second law of thermodynamics unless event horizons are absent, which requires a spatially closed universe and an infinite series of "Kasner crushings". No unintelligent process could produce this series, and no carbon-based life could survive so close to the collapse. So artificial intelligence is required. Physics requires a singularity.

I do not have the expertise in physics to assess Tipler's argument. I take it that certain key claims will be questioned by most physicists, however: for example, the claim that black hole evaporation violates unitarity, the claim that the universe must end in collapse, and the claim that a universe with an initial singularity or final collapse must be spatially closed. I also note that the last two steps of the argument in the previous paragraph seem questionable. First, to produce the Kasner crushings, non-artificial but non-carbon-based intelligence would presumably serve as well as AI. Second, for all Tipler has said, an AI process at or below human-level intelligence will be able to bring about the crushings. The more limited conclusion that the laws of physics require non-carbon-based intelligence would still be a strong one, but it falls short of requiring a full-blown singularity.

Schmidhuber says that we should stop discussing the Singularity in such an abstract way, because AI research is nearly there. In particular, his Gödel machines are well on the path to self-improving intelligence that will lead to an intelligence explosion. Again, I lack the expertise to fully assess the argument, but past experience suggests that a certain amount of caution about bold claims by AI researchers advocating their own frameworks is advisable. I am certainly given pause by the fact that implementation of Gödel machines lags so far behind the theory. I would be interested to see the evidence that the sort of implementation that may lead to a full-blown intelligence explosion is itself likely to be practically possible within the next century, say.

Before moving on, I will note one of the most interesting responses to the singularity argument I have come across. In a discussion at Berkeley, a teacher in the audience noted that what goes for the design of artificial systems may also apply to the teaching of human systems. That suggests the follow analog (or parody?) of I.J. Good's argument for a singularity:

> Let a superteacher be defined as a teacher who we teach to surpass the teaching activities of any existing teacher however competent. Since the teaching of teachers is one of these teaching activities, a superteacher could teach even better teachers. There would then unquestionably be a "teaching explosion", and the results of current education would be left far behind.

It is an interesting exercise to evaluate this argument, consider possible flaws, and consider whether those apply to the argument for an intelligence explosion. I suppose that it is far from clear that we can teach a superteacher. We can certainly train ordinary teachers, but it is not obvious that any simple extension of these methods will yield a superteacher. It is also far from clear that a proportionality thesis applies to teaching: just because one teacher is 10% better than another, that does not mean that the former can teach their pupils to be 10activity) than the pupils of the latter. Even in ordinary cases it is arguable that this thesis fails, and as we move to extraordinary cases it may be that the capacity limits of the brain will inevitably lead to diminishing returns. Now, a proponent of the original intelligence explosion argument can argue that AI design differs from teaching in these respects. Still, the argument and the analogy here certainly repay reflection.

3 Negotiating the Singularity

Three of the articles (by Goertzel, Hanson, and Yampolskiy) concern how we should best negotiate the singularity, and three (by Hutter, Shanahan, and Voorhees) concern its character and consequences. Most of these do not engage my article in much depth (appropriately, as these were the areas in which my article had the least to say), so I will confine myself to a few comments on each.

Goertzel's very interesting article suggests that to maximize the chance of a favorable singularity, we should build an "AI Nanny": an AI+ system with the function of monitoring and preventing further attempts to build AI+ until we better understanding the processes and the risks. The idea is certainly worth thinking about, but I have a worry that differs from those that Goertzel considers. Even if building an AI Nanny is feasible, it is certainly much more difficult than building a

regular AI system at the same level of intelligence. So by the time we have built an AI nanny at level-n, we can expect that there will exist a regular AI system at level-$n+1$, thereby rendering the nanny system obsolete. Perhaps an enormous amount of co-ordination would avoid the problem (a worldwide AI-nanny Manhattan project with resources far outstripping any other project?), but it is far from clear that such co-ordination is feasible and the risks are enormous. Just one rogue breakaway project before the AI nanny is complete could lead to consequences worse than might have happened without the nanny. Still, the idea deserves serious consideration.

Hanson says that the human-machine conflict is similar in kind to ordinary intergenerational conflicts (the old generation wants to maintain power in face of the new generation) and is best handled by familiar social mechanisms, such as legal contracts whereby older generations pay younger generations to preserve certain loyalties. Two obvious problems arise in the application to AI+. Both arise from the enormous differences in power between AI+ systems and humans (a disanalogy with the old/young case). First, it is far from clear that humans will have enough to offer AI+ systems in payment to offset the benefits to AI+ systems in taking another path. Second, it is far from clear that AI+ systems will have much incentive to respect the existing human legal system. At the very least, it is clear that these two crucial matters depend greatly on the values and motives of the AI systems. If the values and motives are enough like ours, then we may have something to offer them and they may have reason to respect legal stability (though even this much is far from clear, as interactions between indigenous and invading human societies tends to bring out). But if their values and motives are different from ours, there is little reason to think that reasoning based on evidence from human values will apply. So even if we take Hanson's route of using existing social mechanisms, it will be crucial to ensure that the values and motives of AI systems fall within an appropriate range.

Yampolskiy's excellent article gives a thorough analysis of issues pertaining to the "leakproof singularity": confining an AI system, at least in the early stages, so that it cannot "escape". It is especially interesting to see the antecedents of this issue in Lampson's 1973 confinement problem in computer security. I do not have much to add to Yampolskiy's analysis. I am not sure that I agree with Yampolskiy's view that the AI should never be released, even if we have reason to believe it will be benign. As technology progresses, it is probably inevitable that someone will produce an unconfined AI+ system, and it is presumably better if the first unconfined AI+ is benign.

Likewise, I have little to add to Hutter's analysis of the character of an intelligence explosion. On his questions of whether it will be visible from the outside or the inside, I am somewhat more inclined to give positive answers. From the outside, even an "inward" explosion is likely to have

external effects, and an "outward" explosion may at least involve a brief period of interaction with outsiders (or alternatively, sudden death). Hutter is probably right that observation of the entire explosion process by outsiders is unlikely, however. From the inside, a uniform speed explosion might not be detectable, but there are likely to be sources of nonuniformity. It is likely that the world will contain unaccelerated processes to compare to, for example. And algorithmic changes are likely to lead to qualitative differences that will show up even if there is a speed explosion. Hutter is certainly right that it is not easy to draw a boundary between speed increases and intelligence increases, as the earlier example of an instant Manhattan project suggests. But perhaps we can distinguish speed improvements from algorithmic improvements reasonably well, and then leave it as a matter for stipulation which counts as an increase in "intelligence". Hutter's discussion of the potential social consequences of AI and uploading is well-taken.

Shanahan and Voorhees focus on the different sorts of intelligence that AI+ and AI++ systems might have. Voorhees discusses differences in both cognitive and sensory intelligence. Shahahan suggests that that a singularity may evolve an evolution from prereflective and reflective creatures (our current state) to postreflective creatures, and that this evolution may involve a sea change in our attitude to philosophical questions. It is striking that Shanahan's superintelligent postreflective beings appear to hold the Wittgensteinian and Buddhist views with which Shanahan is most sympathetic. Speaking for myself, I might suggest (by parity of reasoning) that AI++ systems will certainly be dualists! More likely, systems that are so much more intelligent than us will have philosophical insights of a character that we simply have not anticipated. Perhaps this is the best hope for making real progress on eternal philosophical problems such as the problem of consciousness. Even if the problem is too hard for humans to solve, our superintelligent descendants may be able to get somewhere with it.

4 Consciousness

Although the target article discussed consciousness only briefly, three of the commentators (Brown, Dennett, and Kurzweil) focus mainly on that issue, spurred by my earlier work on that topic or perhaps by the name of this journal.

Kurzweil advocates a view of consciousness with which I am highly sympathetic. He holds that there is a hard problem of consciousness distinct from the easy problems of explaining various functions; there is a conceptual and epistemological gap between physical processes and consciousness (requiring a "leap of faith" to ascribe consciousness to others); and that artificially

intelligent machines that appear to be conscious will almost certainly be conscious. As such, he appears to hold the (epistemological) further-fact view of consciousness discussed briefly in the target article, combined with a functionalist (as opposed to biological) view of the physical correlates of consciousness.

Dennett reads the target article as a mystery story, one that superficially is about the singularity but fundamentally is about my views about consciousness. I think that is a misreading: my views about consciousness (and especially the further-fact view) play only a marginal role in the article. Still, Dennett is not the only one to find some puzzlement in how someone could at once be so sympathetic to AI and functionalism on one hand and to further-fact views about consciousness on the other. I thought I addressed this in the target article by noting that the two issues are orthogonal: one concerns whether the physical correlates of consciousness are biological or functional, while the second concerns the relationship between consciousness and those physical correlates. Dennett has nothing to say about that distinction. But even without getting into philosophical technicalities, it is striking that someone like Kurzweil has the same combination of views as me. Perhaps this is evidence that the combination is not entirely idiosyncratic.

Dennett likes my fading and dancing qualia arguments for functionalism, and wonders why they do not lead me to embrace type-A materialism: the view that there is no epistemological gap between physical processes and consciousness. The answer is twofold. First, while fading and dancing qualia are strange, they are not logically impossible. Second, even if we grant logical impossibility here, the arguments establish only that biological and non-biological systems (when functionally equivalent in the same world) are on a par with respect to consciousness: if one is conscious, the other is conscious. If there is no epistemic gap between biological processes and consciousness (as type-A materialism suggests), it would follow that there is no epistemic gap between nonbiological processes and consciousness. But if there *is* an epistemic gap between biological processes and consciousness (as I think), the argument yields an equally large epistemic gap between nonbiological processes and consciousness. So the argument simply has no bearing on whether there is an epistemic gap and on whether type-A materialism is true. As far as this argument is concerned, we might say: type-A materialism in, type-A materialism out; epistemic gap in, epistemic gap out.

Dennett says that mere logical possibilities are not to be taken seriously. As I see things, logical possibilities make the difference between type-A and type-B materialism. If zombies are so much as logically possible, for example, there is an epistemic gap of the relevant sort between physical processes and consciousness. Still, it is possible to put all this without invoking logical possibility

(witness Kurzweil, who thinks that zombies are "scientifically irrelevant" because unobservable but who nevertheless thinks there is an epistemic gap). The thought-experiment gives reason to think that nonbiological systems can be conscious; but as with our reasons for thinking that other people are conscious, these are nonconclusive reasons that are compatible with an epistemic gap. And as before, the argument shows at best that the epistemic gap for nonbiological systems is no larger than the epistemic gap for biological systems. It does nothing to show that the gap is nonexistent.

I will not go into all the very familiar reasons for thinking there is an epistemic gap: apart from matters of logical possibility, there is Mary in her black-and-white room who does not know what it is like to see red, and most fundamentally the distinction between the hard and the easy problems of consciousness. Dennett quotes me as saying that is no point presenting me with counterarguments as no argument could shake my intuition, but this is a misquotation. I very much like seeing and engaging with counterarguments, but non-question-begging arguments are required to get to first base. Like the arguments of Dennett (1995) that I responded to in Chalmers (1997), Dennett's arguments here work only if they presuppose their conclusion. So the arguments fail to meet this minimal standard of adequacy.

Toward the end of his article Dennett descends into psychoanalysis, offering seven purported reasons why I reject his type-A materialism: faith, fame, Freud, and so on. I am not against psychoanalysis in philosophy, but in this case the psychoanalyses are hopeless. Dennett must at least recognize that my reasons for holding that there is an epistemic gap are very widely shared, both inside and outside philosophy. Even such apparent archreductionists as Ray Kurzweil and Steven Pinker seem to share them. It is easy to see that Dennett's purported analyses do not have a hope of applying in these cases. Now, it is not out of question a deep psychoanalysis could reveal a very subtle mistake or illusion that huge numbers of intelligent people are subject to. In Dennett's more serious moments he has taken stabs at analyses of this sort. It would be nice to see him bring this seriousness to bear once more on the topic of consciousness. It would also be nice to see him bring it to bear on the topic of the singularity.[1]

Brown uses considerations about simulated worlds to raise problems for my view of conscious-

[1]I do apologize to Dennett, though, for not citing his 1978 article "Where am I?" in the context of uploading. It certainly had a significant influence on my discussion. Likewise, my question about a reconstruction of Einstein was intended as an obvious homage to Hofstadter's "A Conversation with Einstein's Brain". These articles are such classics that it is easy to simply to take them as part of the common background context, just as one might take Turing's work as part of the background context in discussing artificial intelligence. But I acknowledge my debt here.

ness. First, he cites my 1990 discussion piece "How Cartesian dualism might have been true", in which I argued that creatures who live in simulated environments with separated simulated cognitive processes would endorse Cartesian dualism. The cognitive processes that drive their behavior would be entirely distinct from the processes that govern their environment, and an investigation of the latter would reveal no sign of the former: they will not find brains inside their heads driving their behavior, for example. Brown notes that the same could apply even if the creatures are zombies, so this sort of dualism does not essentially involve consciousness. I think this is right: we might call it process dualism, because it is a dualism of two distinct sorts of processes. If the cognitive processes essentially involve consciousness, then we have something akin to traditional Cartesian dualism; if not, then we have a different sort of interactive dualism.

Brown goes on to argue that simulated worlds show how one can reconcile biological materialism with the conceivability and possibility of zombies. If biological materialism is true, a perfect simulation of a biological conscious being will not be conscious. But if it is a perfect simulation in a world that perfectly simulates our physics, it will be a physical duplicate of the original. So it will be a physical duplicate without consciousness: a zombie.

I think Brown's argument goes wrong at the second step. A perfect simulation of a physical system is not a physical duplicate of that system. A perfect simulation of a brain on a computer is not made of neurons, for example; it is made of silicon. So the zombie in question is a merely functional duplicate of a conscious being, not a physical duplicate. And of course biological materialism is quite consistent with functional duplicates.

It is true that from the point of view of beings in the simulation, the simulated being will seem to have the same physical structure that the original being seems to us to have in our world. But this does not entail that it is a physical duplicate, any more than the watery stuff on Twin Earth that looks like water really is water. (See note 7 in "The Matrix as metaphysics" for more here.) To put matters technically (nonphilosophers can skip!), if P is a physical specification of the original being in our world, the simulated being may satisfy the primary intension of P (relative to an inhabitant of the simulated world), but it will not satisfy the secondary intension of P. For zombies to be possible in the sense relevant to materialism, a being satisfying the secondary intension of P is required. At best, we can say that zombies are (primarily) conceivable and (primarily) possible—but this possibility mere reflects the (secondary) possibility of a microfunctional duplicate of a conscious being without consciousness, and not a full physical duplicate. In effect, on a biological view the intrinsic basis of the microphysical functions will make a difference to consciousness. To that extent the view might be seen as a variant of what is sometimes known as Russellian monism,

on which the intrinsic nature of physical processes is what is key to consciousness (though unlike other versions of Russellian monism, this version need not be committed to an a priori entailment from the underlying processes to consciousness).

5 Uploading and personal identity

The last part of the target article focuses on uploading: transferring brain processes to a computer, either destructively, nondestructively, gradually, or reconstructively. Two of the commentaries (Greenfield and Plotnitsky) raise doubts about uploading, and three (Blackmore, Dainton, and Schneider and Corabi) focus on connected issues about personal identity.

Greenfield argues that uploading will be very difficult because of the dynamic plasticity of neurons: they are not fixed components, but adapt to new situations. However, there appears to be no objection in principle to simulation processes that simulate all the relevant complexity: not just the static behaviour but the dynamic adaptation of neurons. As Greenfield herself notes, this will require a simulation of all the chemical and biochemical machinery that makes its plasticity and sensitivity possible, but extra detail required here is a difference in degree rather than a difference in kind. Perhaps this sort of simulation is not realistic in the near term, but there is little reason to think that it will not be possible within a time frame of centuries.[2]

Plotnitsky suggests that to create an exact copy of a human being we may need to repeat the history of the universe from the beginning. But of course uploading does not require an exact copy. The most important thing is to create simulations of the essential components of the cognitive system that at least approximates their patterns of functioning and their relations to other components, to within something like the range of background noise or other ordinary fluctuations. Then an upload can be expected to produce behavior that we *might* have produced in a similar environment, even if it does not produce exactly the behavior that we *would* have produced. That is a much easier task. Again it is certainly a nontrivial task and one that may not be accomplished within decades, but it is hard to see that there is an obstacle of principle here.

Corabi, Schneider, and Dainton discuss uploading in the context of personal identity. Their discussions presuppose a good amount of philosophical backgrouns, and consequently the remain-

[2]Greenfield says that I do not define "singularity" in the target article, but it is defined in the first sentence of the article in very much the way that I define it in the second paragraph of this reply. A difference is that in the current article I speak of "systems" rather than "machines" in order to accommodate the possibility that the intelligence explosion takes place in humans by a process of enhancement.

der of this section is philosophically technical.

Corabi and Schneider are doubtful that we can survive uploading. They initially frame their arguments in terms of background metaphysical premises such as substrate and bundle views of individuals, and of three-dimensional and four-dimensional view of identity over time. I do not have firm views on these metaphysical questions and think that they may not have determinate answers, for reasons discussed in my article "Ontological Anti-Realism". However, as Corabi and Schneider themselves note, their central arguments against uploading do not depend on these premises, so I will consider them independently.

Corabi and Schneider argue against destructive uploading on the grounds that it can yield strong spatiotemporal discontinuity in the path of an individual. I might be in Australia at one moment (in a biological body) and then in the US the next moment (in uploaded form) without traveling through the points in between. They suggest that objects do not behave this way (at least if they are construed as substances). However, it is not hard to find objects that exhibit this sort of discontinuous behavior.

In 1713 Yale University moved from Wethersfield to New Haven. I do not know the exact circumstances, but it is not hard to imagine that the move happened with the issuing of a decree. At that moment, the university moved from one place to another without passing through the places in between. One could also imagine versions where it exists for a brief period at both locations, or in which there is a temporal gap during which it is located nowhere. I take it that universities are objects, so there is no general objection to objects behaving this way. There are also objects such as electronic databases that can quite clearly be destructively uploaded from one location to another. Whether we construe objects as substrates or as bundles, a plausible theory of objects should be able to accommodate phenomena such as this. So I do not think that a plausible theory of objects will rule out discontinuous motion of this sort.

Perhaps Corabi and Schneider hold that there is a disanalogy between universities and people. Perhaps people are fundamental entities where universities (and databases) are nonfundamental entities, for example, and perhaps the continuity constraint is more plausible where fundamental entities are concerned. They say explicitly that they intend their arguments to apply on a materialist view (on which people are not fundamental), however, so this cannot be what is going on. And if we assume a substance dualist view on which people are fundamental nonphysical entities, there is not much reason to suppose that nonphysical entities are subject to the same continuity constraints as fundamental physical entities.

Corabi and Schneider also argue against gradual (destructive) uploading. They say that it is

subject to the same issues concerning spatiotemporal discontinuity, at the "dramatic moment at which the data is assembled by the computer host and the isomorph is born". Here I suspect that they are conceiving of gradual uploading in the wrong way. As I conceive of gradual uploading, there is no such dramatic moment. A functional isomorph of the original is present throughout. Its neurons are replaced one at a time by uploaded copies, leading from a 100% biological system to a 99%-1% system (biological-silicon, say), a 98%-2% system, and so on until there is a 100isomorph of the original. Insofar as the person changes location it will be a gradual change, one neuron at a time.

The same misconception may be at play in Corabi and Schneider's formal argument against gradual uploading, where they appeal to a scenario in which a single person is gradually uploaded to two locations. Their premise (A) says "If [gradual] uploading preserves the continuity of consciousness, then the continuity of consciousness can be duplicated in multiple locations". As I conceive of gradual uploading, this premise is much less obvious than Corabi and Schneider suggest. Gradually uploaded systems are not built up from nothing; rather they are connected to the original system throughout. At the initial stages, the 99uploaded version of the remaining 1could be causally integrated with two such systems. Perhaps one could be a backup copy that does not causally affect the brain, but then it will not count as causally integrated. Perhaps two copies could both be integrated with the brain by a sort of causal overdetermination, but if so the combined system is most naturally treated as a single system.

Perhaps one might run a version of the latter scenario that eventually leads to two different independent systems, both of which will have a sort of continuity of consciousness with the original. This sort of case is best treated as a sort of fission case, perhaps akin to a case where one gradually splits my brain while I am still conscious. It is not easy to know what to say about those cases. Perhaps the most natural view of these cases holds that they involve a sort of survival that falls short of numerical identity but that nevertheless yields much of what we care about in numerical identity. In any case, the possibility of uploading does not seem to add any difficulties that do not already arise from the possibility of fission.

Dainton suggests that my claim that continuity of consciousness is our best guide to personal identity is in tension with the further-fact views and deflationary views of personal identity (both of which I have some sympathy for): if continuity secures identity, then there is no room for further facts and no room for deflation.

Considering first the further-fact view: here I meant only to be saying (as Dainton suggests) that physical facts and synchronic mental facts (facts about the states of a subject at specific times)

can leave open questions about identity. In the article I suggested that once we specify continuity of consciousness over time in addition, there is no such open question. Still, on reflection I do not think the matter is cut and dried. One can consistently hold that continuity is the best guide that we have to consciousness without holding that it is an indefeasible guide. So the "best guide" view is compatible with there being an epistemological further fact about identity over and above facts about continuity.

If we put things in terms of conceivability, the key issue is whether one can conceive of cases in which continuity takes a certain pattern and certain identity facts obtain, and cases with the same pattern in which other identity facts obtain. The case of gradual fission while conscious might be an example: perhaps I can conceive myself surviving as the left or the right hemisphere? It may even be that ordinary continuity of consciousness is epistemically compatible with identity and its absence. For example, it is not obviously inconceivable that a single stream of consciousness could involve different subjects at the start and at the finish (perhaps Cartesian egos could swap in and out of a stream of consciousness?). There are tricky issues here about how to characterize continuity without presupposing a single subject throughout, but I think one can appeal to a notion of q-continuity (analogous to Parfit's q-memory, and needed in any case to handle fission scenarios) that makes no such presupposition. Then there may at least be an epistemic gap between q-continuity and identity.

Another worry is provided by cases where continuity is absent, such as the gap between sleep and waking. Here it is arguable that one can conceive of both surviving this gap and failing to survive it. Dainton tries to remove the gap by extending his notion of continuity to C-continuity that obtains across this gap: roughly, the idea is that the pre-sleep and post-waking states are C-continuous iff, had the intermediate systems been conscious throughout, these states would have been part of a single stream of consciousness. Dainton has thought about this matter more than me, but my initial reaction is that even if continuity without identity is inconceivable, C-continuity without identity may be conceivable. The counterfactuals involved in C-continuity may be such as to change the facts about identity: for example, it seems quite consistent to say that a stream involves a single subject over time, but that if it had been interrupted then there would have been two subjects over time. But I may be wrong about this.

Insofar as I take a further-fact view I take the view that there is Edenic survival, with deep further facts about the identity of a self as there might have been in Eden. (Here, as in my "Perception and the Fall from Eden", we can think of Eden as the world as presented by our experience and intuitions, and of Edenic survival as analogous to Edenic colors, the primitive colors that obtain

in Eden.) I think Edenic survival involves primitive identity facts: facts about the identity of subjects over time that are not reducible to any other facts. If so, there will always be an epistemic gap between non-identity facts and identity facts. If continuity of consciousness is specified in a way that builds in identity (the same subject has a certain stream over time), then there need be no gap between continuity and identity, but if it is specified in a way that does not (as mere q-continuity, for example), then there will be such a gap. Some non-identity-involving condition such as q-continuity may serve as a sort of *criterion* for identity over time, but it will be a criterion is merely contingently associated with identity (perhaps as a matter of natural necessity). From this perspective, a view (like Dainton's?) on which there is no epistemic gap between q-continuity and identity is already somewhat deflationary about identity and survival.

On Dainton's view, gradual uploading preserves identity (there is continuity of a stream of consciousness or the potential for it), while destructive uploading does not (there is no such potential). Insofar as I hold an Edenic view, I think it is at least conceivable that someone survives destructive uploading; this once again brings out the epistemic gap between facts about C-continuity and facts about survival. I am far from sure that this is naturally possible, though: it is not out of the question that the sort of Parfit-style psychological relation found in cases of destructive uploading provides a contingent criterion for identity. On the Edenic view this is an issue for speculation that philosophy and science may have a hard time resolving. Still, insofar as continuity of consciousness provides a (contingent) sufficient condition for survival, then gradual uploading will serve at least as well to ensure survival as destructive uploading, and quite possibly better.

That said, I am a little skeptical about Edenic survival and somewhat more sympathetic to deflationary views. I think that once up gives up on Edenic survival, one should hold that there are no deep facts about survival. There are many relations connecting subjects over time, but none carry absolute weight. On a moderate version of the view (like Parfit's), there is some reason to care about each of these relations. On an extreme version of the view, there is no special reason to care about one rather than another (or at least no reason for the distinctive sort of caring that we typically associate with identity relations); the only thing that distinguishes them is that we do care about some rather than others. Either way, it is quite consistent to hold that continuity is crucial to the relation that we care about most or that we should care about most.

I am inclined to think the extreme view is more consistent, although it is also more counterintuitive. Moderate deflationary views do not seem to be able to tell a good story about *why* causal or continuity connections give us reason to care in an identity-like way. Our ordinary reasons for caring about these relations seem to stem from the fact that we take them to provide good criteria

for something like Edenic survival. Once we have discarded Edenic survival, these reasons seem to have little residual force. Something similar may well apply to Dainton's view that reduces survival to a sort of continuity.

So insofar as I endorse a deflationary view, I incline toward an extreme deflationary view on which identity-based concern for the future is irrational or at best arational (after philosophical reflection). Blackmore articulates this view nicely (Prinz and McDermott also appear to hold versions of it). As she notes, there may be other reasons to care about future selves. We can reasonably care that our current projects are fulfilled, for example, and it may be that caring about future selves (like caring about others in our community) makes things better for everyone. Speaking for myself, whether or not there are *reasons* for identity-based concern about the future, I find this sort of concern impossible to abandon. I think it will probably always function in our lives the way that other arational desires function. Still, all this means that if we come to care about our future uploaded selves in much the same way that we care about our future biological selves, then uploading will be on a par with ordinary biological survival.

Dainton goes on to consider the hypothesis that we are inhabiting a simulated universe. As well as considering Bostrom's simulation argument (which can be used to suggest that this hypothesis is quite likely) and some ethical and practical issues, he also engages my argument (in "The Matrix as Metaphysics") that this hypothesis is not a skeptical hypothesis. That is: it is not correct to say that if we are inhabiting a simulated universe, tables and chairs (and so on) do not exist. Rather, ordinary external objects exist, and we should just revise our metaphysical views about their ultimate constitution. Here Dainton objects to my argument on the ground that the computational structure of a simulated universe does not suffice for it to yield a properly spatial universe. To yield a true space, the underlying processes in a simulation would need to be laid out in the right sort of spatial arrangement. Without that arrangement, the simulation scenario will indeed be a skeptical scenario.

I think this is the best way to resist the argument of "The Matrix as Metaphysics", as I noted in that article. As with the case of survival and of color, I think we may have a grip on a primitive concept of space—call it a concept of Edenic space—that requires certain special primitive relations to obtain, relations that may not obtain in a simulation. But as with Edenic color and Edenic survival, I think there are serious grounds for doubt about whether there is Edenic space in our world. Perhaps there might be Edenic space in a Newtonian world. But relativity theory and quantum mechanics both give grounds for doubt about whether space can be Edenic. Still, we are not inclined to say that there is no space in our world. In the case of color, after we drop

Edenic color we tend to identify colors with whatever play the relevant roles, for example in producing our color experience. Likewise, after dropping Edenic space, I think we identity space with whatever plays the relevant roles, both in scientific theory and in producing our experience. But this functionalist conception of space then opens the door for there to be space in a Matrix scenario: we identify space with whatever plays the relevant role in that scenario, just as we do in quantum mechanics and relativity (and even more so in more speculative theories that postulate more fundamental levels underneath the level of space). Of course there is much more to say here, for example about the choice between functionalist and primitivist conceptions of space; I try to say some of it in my forthcoming book *Constructing the World*.

6 Conclusion

The commentaries have reinforced my sense that the topic of the singularity is one that cannot be easily dismissed. The crucial question of whether there will be a singularity has produced many interesting thoughts, and most of the arguments for a negative answer seem to have straightforward replies. The question of negotiating the singularity has produced some rich and ingenious proposals. The issues about uploading, consciousness, and personal identity have produced some very interesting philosophy. The overall effect is to reinforce my sense that there is an area where fascinating philosophical questions and vital practical questions intersect. I hope and expect that these issues will continue to attract serious attention in the years to come.[3]

Bibliography

Aleksander, I. 2012. Design and the singularity: The philosopher's stome of AI? *Journal of Consciousness Studies* 19.

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

Bostrom, N. 2003. Are we living in a simulation? *Philosophical Quarterly*.

Bringsjord, S. 2012. Belief in the singularity is logically brittle. *Journal of Consciousness Studies* 19.

Broderick, D. 2012. Terrible angels: The singularity and science fiction. *Journal of Consciousness Studies* 19:20-41.

Brown, R. 2012. Zombies and simulation. *Journal of Consciousness Studies* 19.

[3]Thanks to Uziel Awret for all his work in putting together the special issues and for his comments on this reply.

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

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

Chalmers, D.J. 1997. Moving forward on the problem of consciousness. *Journal of Consciousness Studies* 4:3-46.

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

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

Chalmers, D.J. 2010. The singularity: A philosophical analysis. *Journal of Consciousness Studies* 17:7-65.

Corabi, J. & Schneider, S. 2012. The metaphysics of uploading. *Journal of Consciousness Studies* 19.

Dainton, B. 2012. On singularities and simulations. *Journal of Consciousness Studies* 19:42-85.

Dennett, D.C. 1978. *Where am I?* In *Brainstorms* (MIT Press).

Dennett, D.C. 1995. Facing backward on the problem of consciousness. *Journal of Consciousness Studies* 3:4-6.

Dennett, D.C. 2012. The mystery of David Chalmers. *Journal of Consciousness Studies* 19:86-95.

Goertzel, B. 2012. Should humanity build a global AI nanny to delay the singularity until its better understood? *Journal of Consciousness Studies* 19:96-111.

Greenfield, S. 2012. The singularity: Commentary on David Chalmers. *Journal of Consciousness Studies* 19:112-118.

Hanson, R. 2012. Meet the new conflict, same as the old conflict. *Journal of Consciousness Studies* 19:119-125.

Heylighen, F. 2012. Brain in a vat cannot break out. *Journal of Consciousness Studies* 19:126-142.

Hofstadter, D.R. 1981. A conversation with Einstein's brain. In D.R. Hofstadter and D.C. Dennett, eds., *The Mind's I*. Basic Books.

Hutter, M. 2012. Can intelligence explode? *Journal of Consciousness Studies* 19:143-166.

Kurzweil, R. 2012. Science versus philosophy in the singularity. *Journal of Consciousness Studies* 19.

Lampson, B.W. 1973. A note on the confinement problem. *Communications of the ACM* 16:613-615.

McCorduck, P. 2012. A response to "The Singularity," by David Chalmers. *Journal of Consciousness Studies* 19.

McDermott, D. 2012. Response to 'The Singularity' by David Chalmers. *Journal of Consciousness Studies* 19:167-172.

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

Nunn, C. 2012. More splodge than singularity? *Journal of Consciousness Studies* 19.

Plotnitsky, A. 2012. The singularity wager: A response to David Chalmers' "The Singularity: A Philosophical Analysis". *Journal of Consciousness Studies* 19.

Prinz, J. 2012. Singularity and inevitable doom. *Journal of Consciousness Studies* 19.

Sandberg, A. and 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]

Schmidhuber, J. 2012. Philosophers & futurists, catch up! *Journal of Consciousness Studies* 19:173-182.

Shanahan, M. 2012. Satori before singularity. *Journal of Consciousness Studies* 19.

Shulman, C. & Bostrom, N. 2012. How hard is artificial intelligence? Evolutionary arguments and selection effects? *Journal of Consciousness Studies* 19.

Steinhart, M. 2012. The singularity: Beyond philosophy of mind. *Journal of Consciousness Studies* 19.

Tipler, F. 2012. Inevitable existence and inevitable goodness of the singularity. *Journal of Consciousness Studies* 19:183-193.

Voorhees, B. 2012. Parsing the singularity. *Journal of Consciousness Studies* 19.

Yampolskiy, R. 2012. Leakproofing the singularity: Artificial intelligence confinement problem. *Journal of Consciousness Studies* 19:194-214.