Jesse Prinz

Singularity and Inevitable Doom

Chalmers (2010) has articulated a compellingly simple argument for inevitability of the singularity — an explosion of increasingly intelligent machines, eventuating in super forms of intelligence. Chalmers then goes on to explore the implications of this outcome, and suggests ways in which we might prepare for the eventuality. I think Chalmers' argument proves both too much and too little. If the reasoning were right, it would follow inductively that the singularity already exists, in which case Chalmers would have proven more than he set out to. Moreover, I will suggest that, if the singularity already exists, we are doomed. Fortunately, Chalmers' reasoning is problematic. I will consider several objections. Unfortunately, the most serious problem is that human life may end long before the singularity is created. In that case, we are doomed either way. Should we care? Perhaps not.

1. Chalmers on the Singularity

Chalmers' argument can be briefly summarized as follows. The first premise says: There will be artificial intelligence (AI) before long. This premise is based on the assumption that the human brain — an obvious source of intelligence — is a machine, and we will be able to emulate this machine eventually. The second premise says: If there is AI, there will be AI+ (a superior form of intelligence). Chalmers reasons that the method by which we create AI will be extendable, meaning we can use this method to create forms of intelligence that exceed what we find in ordinary human brains. For example, we could make brain simulations that work faster than human brains, or we could simulate evolutionary pressures to evolve increasingly intelligent

Correspondence:
Email: jesse@subcortex.com

Journal of Consciousness Studies, **19**, No. 7–8, 2012, pp. 77–86

machines, or we could tweak learning algorithms to make them more and more powerful. Chalmers' third and final premise says: If there is AI+, there will be AI++ (i.e. super-intelligence). The argument for this is a simple induction. If there is artificial intelligence of some degree *n*, then we can expect there to eventually be intelligence of degree *n*+1 (for reasons given in defence of the second premise); this seems true for any *n*.

These three premises lead to the conclusion that there will be AI++, the kind of intelligence postulated by the singularity hypothesis. Chalmers clarifies and qualifies this in a couple of important ways. First, the time frame. The first premise, which says that we will eventually create AI, is estimated to come true in a matter of centuries, with AI+ following within decades after that, and other incremental increases also taking decades. So AI++ may be likely, but it's not coming within our lifetimes or the lifetimes of anyone alive today, assuming we don't find ways to dramatically extend life. Second, AI++ is likely, but not quite inevitable. There could be defeaters, such as natural disasters or a lack of motivation. The conclusion that there will be AI++ is more accurately stated, 'There will be AI++ absent defeaters'. This is a substantive claim, of course, but the qualification is important, and we will come back to it. The likelihood of AI++ depends on the likelihood of defeaters.

After presenting his argument, Chalmers goes on to express some anxiety about the singularity. What if super-intelligent machines are unfriendly? What if they have no need for us? Chalmers argues that we might try to bypass such risks by keeping our intelligent machines virtual, but he notes that brilliant virtual machines will be clever enough to escape such containment. He suggests that we proceed slowly and develop ways to integrate human minds with the ever-increasing forms of intelligence, so we don't end up subordinate to it. To do so, we might be best off becoming virtual ourselves — converting human minds into digital bits that can be uploaded to computers and steadily enhanced. Such uploading might appear to threaten consciousness or identity, but Chalmers argues that such concerns can be allayed. He argues that digital analogues of ourselves are conscious by appeal to his old fading qualia argument: it seems unlikely that replacing one neuron by a digital chip would disrupt consciousness, and a person who underwent gradual replacement would deny that consciousness had faded; so there is little reason to think that digitalization entails zombification. As for selfhood, the trick would be incremental change. If Parfit is right, survival depends on continuity between past and future selves, and gradual digitalization and digital

enhancement would ensure survival in this sense. Chalmers does not think the Parfitian move is decisive, but it does give some reason for optimism.

In summary, the singularity is inevitable barring defeaters, and we can increase our prospect of survival by creating virtual copies of ourselves and integrating them into the simulations that will eventuate in super-intelligence.

2. Are We Living in a Simulation?

On reflection Chalmers' elegant argument seems to prove too much. Suppose that he is right to AI++ is inevitable. And suppose that in pursuing AI++ we learn to make digital simulations of ourselves that preserve our thoughts and qualia. If we can create such virtual copies of ourselves, then we might, in fact, be such copies now. We wouldn't be able to tell. In fact, I think the conclusion that the singularity is inevitable entails that we probably are living inside a simulation. Here is why.

A super-intelligent being would be capable of creating simulations of possible worlds. It is likely that it would do so. First, if Chalmers is right, creating a simulated reality would be a step we ourselves would initiate en route to creating AI++. Second, a super-intelligent being could optimize its choices by simulating multiple worlds and figuring out which is best, in much the way that a chess master anticipates possible games in order to decide how to move. Given the complexity (in the technical sense) of the world, such simulations might be the only way to make optimal decisions.

Both of these two points have a further implication. If simulations of the world, including simulated versions of ourselves, are possible, there are likely to be many of them. If we create simulations while preparing for the singularity, we might multiply them to explore different possible outcomes to increase our prospects for survival. We might also store simulated versions of the past, in order restore a past state if disaster were to strike. If there were a super-intelligence, it too would want to multiply simulations to make optimal choices. In fact, like a Leibnizian God, it might create every possible simulation and select only the one that is best. It might also simulate the past and many variations of the past to have complete knowledge of history, and a sense of the modal space that surrounds history: the ways in which we could have acted differently in our steps towards super-intelligence. Super-intelligent beings might just proliferate virtual worlds for fun in much the way we proliferate stories in fiction.

Thus, if there were a super-intelligent being, there would probably be many many simulations of the world we live in. Now recall that we have no way of knowing whether we are in one of these simulated worlds or the actual world. Which is more likely? Well, if AI++ is inevitable, and if the inevitability of AI++ entails that there will be multiple simulations, then it's more likely, statistically speaking, that we are in a simulation. That probability increases as the number of simulations increases. It is hard to say how many simulations there would be, but assuming that there is more than one, say two, then it's that many times more likely that our 'reality' is merely virtual. So, if the singularity is inevitable, we are more likely to be virtual than real. It also follows that, if the singularity is inevitable, it is probably actual, since it's inevitability would entail that we are probably living in one of its simulations (assuming it creates more numerous and accurate simulations than we do, en route to the creation of the singularity).

In some way, that would be a very happy conclusion, given the anxieties expressed in the second part of Chalmers' paper. If we worry that a super-intelligence would be unfriendly and threaten our way of life, then we need only reflect on the fact that we may be living alongside (or within?) the singularity already, in which case, life with the singularity is, trivially, no worse than it is now. Of course, the singularity may also be simulating some less pleasant worlds, but that need not concern us because we are lucky enough to be living in a fairly benign simulation — at least for those of us who are simulated to be living healthy lives in affluent nations.

3. Bleak Implications of Virtual Reality

Unfortunately, there is trouble in paradise. Let's assume that we are living in a simulation created by the singularity. In this simulated world we continue to pursue all of our ends, including the goal of creating intelligent machines. And, in the stimulated world, Chalmers' argument may lead to the conclusion that the singularity is inevitable. Indeed, if his reasoning is right, that conclusion follows. But the singularity that created this simulated world of ours knows that. It knows that we will take steps that would lead to the creation of super-intelligence. Moreover, it knows that a super-intelligence would be clever enough to escape the simulation. Were this simulated super-intelligence to escape the simulation, it might seek to destroy the super-intelligent being that created the simulation. If it had a desire to exist, the simulated super-intelligence would want autonomy and conquest

lest its fate depend on some other intelligent being. A war between super-intelligent beings might ensue.

Of course, the super-intelligent being who created the simulation we are in knows about this threat. It knows that we might take steps towards creating a being that could compete with it. To prevent that from happening, it would almost certainly build in mechanisms that guarantee that there won't be an intelligence explosion. What guarantee might exist?

Recalling Chalmers' argument, there are two possible defeaters: lack of motivation and catastrophe. Clearly the first of these is unlikely. We are motivated to create AI and AI+. That motivation might suddenly wane by some contrivance in the simulation we are living, but that seems unlikely. It would involve a sudden change in our motivational structure. We could be simulated in such a way that our ambitions change radically, but given the human quest for knowledge and technology, this alteration would involve a radical shift in goal structure tantamount to a biological reconfiguration or a world event that caused a radical break from our historical trajectory. Such interventions might not guarantee permanent apathy. In time, we might begin our quest for super-intelligence anew. It seems likely then that the super-intelligent being would stop the intelligence explosion by means of catastrophic interference. It would guarantee that we are doomed. It might rig a world-crushing natural disaster to occur before the intelligence explosion advances far enough to escape such outcomes. If Chalmers is right that we are a few centuries from the advance to superior intelligence, this disaster would have to come safely before progress is made. Perhaps doom is around the corner.

Thus, if we are living in a simulation, we are probably doomed. All our efforts to avoid this fate, including the strategies suggested by Chalmers, are for naught.

One might think that a super-intelligent being wouldn't simply exterminate the human race, because doing so would be immoral. But this assumes that intelligence entails ethical regard to sentient beings. There is no reason to think this is so. Despite two thousand years of trying, philosophers have never been able to establish that intelligence alone has moral implications. There is nothing irrational (contrary to fact and logic) about killing. Indeed, given the minimal goal of self-preservation, which may be a precondition for the intelligence explosion, it may be rational to destroy anything that poses a threat. Super-intelligent beings would likely see us as a threat (whether real or virtual), and they would work to neutralize us. Of course, such beings might have goals that transcend cool reason. Chalmers raises

the possibility that we might work to ensure that super-intelligent beings have concern for us, but he realizes that there is little we can do to make that outcome likely. Smart beings can adapt concerns to serve their interests, and it is in their interest to destroy us.

Chalmers thinks we can escape this outcome by incorporating ourselves into artificially intelligent beings. But this strategy won't work if we are already in a simulation. The super-intelligent being that created us won't let us get that far. The beings we hope to create and with which we hope to be integrated pose a threat. Catastrophe will be engineered to impede progress and destroy human live.

4. Problems with Chalmers' Argument

If the arguments so far are right, then Chalmers has unwittingly proven that we are probably living in a simulation created by a super-intelligent being. And, if that is the case, we are probably doomed. That's an unsettling thought.

One could try to block this conclusion by quibbling with Chalmers' argument. Against the premise that AI is inevitable, one might argue brain simulation is not feasible. Brains are the most complex structures we know and creating one would be prohibitive. More plausibly, one might reject the cascade from AI to AI++ by noting limits on extendibility. Even if we could create brain simulations that work more efficiently than human brains, we couldn't necessarily keep improving brain-power without limit. Storage might require increasing size, and size might require increasing energy, but both size and energy are finite resources. Moreover, smarter brains may never become super-intelligent because the skill-set implemented by brains may have limited potential. For example, would a person with perfect memory be vastly more intelligent than the rest of us? She might be slower at making decisions and prone to repeating past ideas rather than inventing new ones. Would better learning algorithms lead to super-intelligence? Presumably not, because the application of intelligence requires skills for using information that has been learned. Can we create increasingly smart machines by simulating evolution? Unlikely, because evolution requires just enough intelligence to survive, and we have no idea how to create environments whose challenges demand more and more intelligence. Chalmers' induction premise requires some method for moving up to arbitrary levels of intelligence. But it's not clear that known methods of engineering guarantee such an outcome.

Indeed, without an account of what intelligence is, it's not even clear that it makes sense to talk about incremental increases. Often what we call 'intelligence' is really ingenuity, and that involves putting information to new uses. But ingenuity may not be a scalar resource that can keep expanding. If there were beings with optimal problem solving skills, that would not guarantee a continuous advance in ingenuity. Ingenuity comes only when the skilled individual is confronted with new problems, and there is no guarantee that new problems will continue to appear, much less that we can think of such problems as being ordered in a way that can be incrementally quantified.

For these reasons, I am not yet persuaded of Chalmers' conclusion. But there is also a more basic problem. Recall that the conclusion is implicitly conditional. AI++ is inevitable if there are no defeaters. As noted above, this means the likelihood of AI++ depends on the comparative likelihood of events that would reverse the advance of technology over the next few centuries. Sadly, such events are not unlikely. Formal versions of doomsday arguments are controversial, but there are several well-known threats that could have a major impact on human life.

Putting aside cosmic crises of the kind that caused mass extinctions in the past, we face several very pressing threats. One threat is a global pandemic. Less than a century ago, the Spanish flu killed between 50 and 100 million people in the space of 2 years. That was only 3% of the world population, but now with more global travel, and more resilient viruses, some fear that we could be done in by disease. Another threat is environmental destruction. More than a third of the natural world has been destroyed over the last 30 years, and major loss of the icecaps is expected within the next 40 years, though total loss may be 1,000 years away. Energy sources are dwindling and species are dying. At the moment, these trends do not threaten to kill off our species, but the rapid and radical destruction of the environment could lead to a dramatic change is lifestyle over the coming centuries. Loss of energy could spell trouble for technology. The biggest threat, however, may be weapons of mass destruction. Extant nuclear devices could destroy the planet, and as technology improves, nuclear and chemical weapons will become easier and easier to make. Such options may prove attractive for rogue governments and crazed individuals. In addition, we are likely to see the invention of new weapons that pose serious threats. Most relevant here is the creation and military of use self-replicating nanobots, which could potentially spread out of control like a virus.

This point about military technology can be pushed a bit further. Right now, wealthy governments are in a particularly good position to

fund the quest for intelligent machines, and they are motivated to do so, because such machines have attractive military applications. Autonomous fighting robots could invade a country without deploying single soldier. These machines would not need to have super-intelligence. They could be expert systems singly focused on conquest. But such robotic invaders could potentially pose a threat if they gained enough autonomy to compete with us for energy and other resources. Before we get to AI++, a belligerent breed of AI+ machines might work to destroy us, and to block the advance towards the singularity.

In other words, Chalmers' argument may actually point towards its own defeat. If there is an incremental increase in intelligence, we may end up creating machines that have the power and motivation to destroy us, and to stop technological progress. Such machines would prevent the singularity from coming into existence.

Is there any reason to think the singularity is more likely that these defeaters? Perhaps, but more argument would be needed to see why. Within the last hundred years, we have developed at least two technologies that could exterminate our species (nuclear and chemical weapons) and we have done irreversible damage to the plant. What will the next hundred years bring? Probably both good things and bad, and the bad could be so dangerous as to outweigh the good. Along these lines, the most powerful response to Chalmers may be that the very advances advertised by his argument pose a threat. As machines get smarter, they may become more hostile, and this may spell doom well before the singularity.

5. Living With Doom

I have been arguing that human beings are doomed. If the singularity already exists, then we are living in a simulation and that simulation is likely to have a catastrophe pre-programmed into it to prevent an intelligence explosion from within. If the singularity doesn't exist, the technological advances that would take us toward it are likely to eventuate in machines that are smart enough to destroy us, but not yet super-smart. Either way, the future looks grim.

Should we worry about this? I think not. Perhaps the main shortcoming with Chalmers' paper is not his argument for the inevitability of the singularity, but anxieties he expresses about it. Chalmers encourages us to devise ways to survive the rise of super-intelligent machines. But why should we do that? What's so bad about doom?

One answer to this question is that doom is undesirable psychologically. Many of us dread the thought of dying, and that dread may make

us uncomfortable with the idea that our species is doomed. But this source of discomfort may be irrelevant in the present context. First of all, it may be irrational to fear death. As the Epicurus famously argued, death is not a loss to the dead, because death is nothing. More to the point, doom may come some centuries from now, after we are long gone.

Chalmers right reply is that we still have two reasons to be worried about inevitable doom: a concern for future generations, and the lost prospect of extending life through cryogenics and future uploading. Let me consider these in turn.

We do have special concern for our offspring, which is probably biologically based, culturally reinforced, and highly biased (some would kill a village to save a daughter or son). But what about our concern for unborn generations? Such concern cannot be *de re* in the sense that we cannot have concern for specific people that do not exist. So concern for future generations is more likely to be concern that the human species continue, but why should we care about that? Concern for the species is no more rational than concern about one's own death. The loss of the human race would only be bad if there were creatures who survived to mourn that loss, but that wouldn't be the case in the scenarios I've described. The artificial agents who outlived us would not regret our fate. We might even take comfort in knowing that we'd been superseded by more intelligent beings.

But what about the cryogenic scenario? Some people have been preserving their brains with the hope that future generations will bring them back to life an then upload their contents to digital media, where they can be stored indefinitely. From this perspective, all of us are potentially immortal. Even if it is irrational to fear death, might we not revel in the thought of prolonged life? The doom scenarios may extinguish these hopes for life extension, and that may be cause for concern.

I think this hope is misplaced. Consider first the scenario in which we are living in a simulation. In that scenario, our virtual existence may in fact be a consequence of the fact that our brains were uploaded at some earlier time, so we may already be beneficiaries of the digital road to longevity. But notice, if the contents of our brains were stored and entered into a simulation at some earlier point in time, they may have also been entered into numerous other simulations at that time. Some of these may be pleasant, some unpleasant. Now ask, from the point of view of your previous self, prior to cryogenics and uploading, does it matter that any of these simulations was actually created? If so, which one? From your point of view right now, the fact that there are

other virtual copies of your past self is irrelevant. They should have no more value to you than any other strangers you don't know about. If I threatened to kill you now, it would offer little reassurance to point out that there is a copy of you floating around in a virtual world. Against such reassurance, you will protest: but that's not me! Nor would it be reassuring to learn that this parallel self knew about you or had 'memories' that were drawn from your life. By the same token, however, you should have no special concern for your future selves. They are just other selves who happen to be like you and share some memories with you.

If this reasoning is right, then we should not be worried about doom in the scenario where belligerent robots exterminate the human species. True, such a scenario would prevent your stored brain from being uploaded, but you shouldn't have any concern about continuation into the future. Life is a continual recreation of new selves, the self you are now is neither helped nor harmed by its successors; the self you are now is ephemeral.

Such Parfitian thoughts should bring us some comfort when contemplating doom. Parfit himself uses such reasoning to say we should have concern for future generations, because, once we give up on the idea of personal identity, moral concern can extend, without self-serving bias, to strangers. This inference presupposes that moral concern tracks loci of utility. We should care about anyone who can experience happiness. Of course, the same utilitarian line might, following Mill, assign special value to the forms of happiness that can be experienced by intelligent beings: intellectual pleasure trumps carnal pleasure, says Mill, by our own standards. If Mill's reasoning is right, then the pleasures obtained by superior forms of intelligence may outweigh our own in the utilitarian calculus. Those inclined toward hedon counting should take comfort in the thought that human extinction may usher in artificial agents whose pleasures trump ours, and this gives us one more reason to welcome doom rather than shunning it.

The foregoing has been an exercise in augury. I am not confident in my ability to forecast the future. Perhaps the doom-casting outcomes I've described are less likely than the rosier outcomes envisioned by Chalmers. But, in reflecting on doom, we might come to realize that the future matters less than we think, and that brings attention back to the present. Rather than safeguarding against our eventual destruction, we might work to make things better here and now.