Journal of EVOLUTION & Technology

A peer-reviewed electronic journal
published by the Institute for Ethics
and
Emerging Technologies

ISSN 1541-0099

22(1) – November 2011

Ray Kurzweil and Uploading: Just Say No!

Nicholas Agar
School of History Philosophy Political Science and International Relations
Victoria University of Wellington
nicholas.agar@vuw.ac.nz

Journal of Evolution and Technology - Vol. 22 Issue 1 – November 2011 - pgs 23-36

Abstract

There is a debate about the possibility of mind-uploading – a process that purportedly transfers human minds and therefore human identities into computers. This paper bypasses the debate about the metaphysics of mind-uploading to address the rationality of submitting yourself to it. I argue that an ineliminable risk that mind-uploading will fail makes it prudentially irrational for humans to undergo it.

For Ray Kurzweil, artificial intelligence (AI) is not just about making artificial things intelligent; it's also about making humans artificially super-intelligent.[1] In his version of our future we enhance our mental powers by means of increasingly powerful electronic neuroprostheses. The recognition that any function performed by neurons and synapses can be done better by electronic chips will lead to an ongoing conversion of biological brain into machine mind. We will *upload*. Once the transfer of our identities into machines is complete, we will be free to follow the trajectory of accelerating improvement currently tracked by wireless Internet routers and portable DVD players. We will quickly become millions and billions of times more intelligent than we currently are.

This paper challenges Kurzweil's predictions about the destiny of the human mind. I argue that it is unlikely ever to be rational for human beings to completely upload their minds onto computers – a fact that is almost certain to be understood by those presented with the option of doing so. Although we're likely to find it desirable to replace peripheral parts of our minds – parts dedicated to the processing of visual information, for example – we'll want to stop well before going all the way. A justified fear of uploading will make it irrational to accept offers to replace the parts of our brains responsible for thought processes that we consider essential to our conscious experience, even if the replacements manifestly outperform neurons. This rational biological conservatism will set limits on how intelligent we can become.

Uploading and the debate about strong AI

For the purposes of the discussion that follows, I will use the term "uploading" to describe two processes. Most straightforwardly, it describes the one-off event when a fully biological being presses a button and instantaneously and completely copies her entire psychology into a computer. But it also describes the decisive event in a series of replacements of neurons by electronic chips. By "decisive" I mean the event that makes electronic circuits rather than the biological brain the primary vehicle for a person's psychology. Once this event has occurred, neurons will be properly viewed as adjuncts of electronic circuits rather than the other way around. Furthermore, if Kurzweil is right about the pace of technological change, they will be rapidly obsolescing adjuncts. The precise timing of the uploading event is more easily recognized in the first scenario than it is in the second. It's possible that there will some vagueness about at which point electronic circuits, rather than the biological brain, become the primary vehicle of a person's psychology. The uploading event may therefore be spread out over a series of modifications rather than confined to a single one.

One reason Kurzweil is enthusiastic about uploading is that he's a believer in *strong AI*, the view that it may someday be possible to build a computer that is capable of genuine thought. Computers already outperform human thinkers at a variety of tasks. The chess program on my PC easily checkmates me, and my guesstimates of the time are almost always wider of the mark than is the reading on my PC's clock. But the computer accomplishes these feats by means of entirely noncognitive and nonconscious algorithms. Kurzweil's commitment to strong AI and his belief in the accelerating rate to technological improvement lead him to forecast computers that genuinely think instead of just performing some of the tasks currently done poorly by human thinkers. He has set 2029 as the year in which computers are likely to match and surpass human powers of thought (Kurzweil 2005, 200).

There's an alternative view about the proper goal of artificial intelligence. Advocates of *weak* AI think that computers may be able to simulate thought, and that these simulations may tell us a great deal about how humans think. They maintain, however, that there is an unbridgeable gap between the genuine thinking done by humans and the simulated thinking performed by computers. Saying that computers can actually think makes the same kind of mistake as saying that a computer programmed to simulate events inside of a volcano may actually erupt. Technological progress will lead to better and better computer models of thought. But it will never lead to a thinking computer.

Kurzweil needs strong AI to be the correct view because what we say about computers in general we will also have to say about the electronic "minds" into which our psychologies are uploaded. If weak AI is the correct view, then the decision to upload will exchange our conscious minds for entirely nonconscious, mindless symbol manipulators. The alternatives are especially stark from the perspective of someone considering uploading. If strong AI is mistaken, then uploading is experientially like death. It turns out the light of conscious experience just as surely as does a gunshot to the head. If strong AI is the correct view, then uploading may be experientially like undergoing surgery under general anesthetic. There may be a disruption to your conscious experience, but then the light of consciousness comes back on and you're ready to try out your new cognitive powers.

In what follows I outline a debate between Kurzweil and an opponent of strong AI, philosopher John Searle (who first presents his argument in Searle 1980). I present them as asking humans who are considering making the decisive break with biology to place a bet. Kurzweil proposes that we bet that our capacities for conscious thought will survive the uploading process. Searle thinks we should bet that they will not. I will argue that even if we have a high degree of confidence that computers can think, you should follow Searle. Only the irrational among us will freely upload.

We can see how this bet works by comparing it with the most famous of all philosophical bets – Blaise Pascal's Wager for the prudential rationality of belief in the existence of God. The Wager is designed for those who are not absolutely certain on the matter of God's existence, that is, almost all of us. It leads to the conclusion that it is rational to try as hard as you can to make yourself believe that God exists. Pascal recommends that people who doubt God's existence adopt a variety of religious practices to trick themselves into belief.

Pascal sets up his Wager by proposing that we have two options – belief or disbelief. Our choice should be based on the possible costs and benefits of belief and disbelief. The benefits of belief in God when God turns out to exist are great. You get to spend an eternity in paradise, something denied to those who lack belief. The cost is some time spent in religious observances and having to kowtow to priests, rabbis, imams, or other religious authorities. If you believe in God when there is no God, you miss out on paradise; but then so does everyone else. True disbelief brings the comparatively trifling benefits of not having to defer to false prophets or to waste time in religious observances. The Wager is supposed to give those people who currently think that God's existence is exceedingly unlikely a reason to try as hard as they possibly can to make themselves believe in God. The reward for correctly believing is infinite, meaning that it's so great that even the smallest chance of receiving it should direct you to bet that way. Doubters could look upon faith in the same way as those who bet on horses might view a rank outsider that happens to be paying one billion dollars for the win. But God is a better bet than any race horse – there's no amount of money that matches an eternity in paradise. This means that only those who are justifiably certain of God's nonexistence should bet the other way.

My purpose in presenting Pascal's Wager is to establish an analogy between the issues of the prudential rationality of belief in God and the prudential rationality of uploading your mind onto a computer. What I refer to as "Searle's Wager" moves from the possibility that strong AI is false to the prudential irrationality of uploading. While I will be defending Searle's Wager I certainly do not mean to endorse the Pascal's Wager argument. The two Wagers are similar in their appeals to prudential rationality. Both seek to establish the motivational relevance of a proposition that many will judge likely to be false. But they differ in other respects. As we will see Searle's Wager does not share some of the salient flaws of Pascal's Wager.

Uploading is an option that is not yet available to anyone. Kurzweil thinks that we'll have computers with human intelligence by 2029 and uploading will presumably be technologically feasible only some time after that. This means that we're speculating about the decisions of people possibly several decades hence. But our best guesses about whether people of the future will deem uploading a bet worth making has consequences for decisions we make now about the development of artificial intelligence. If we're confident that uploading will be a bad bet we should encourage AI researchers to direct their energies in certain directions, avoiding more dangerous paths.

Kurzweil versus Searle on whether computers can think

To understand why we should bet Searle's way rather than Kurzweil's we need to see why Searle believes that computers will be forever incapable of thought.

Searle's argument against strong AI involves one of the most famous of all philosophical thought experiments – the Chinese Room. Searle imagines that he is locked in a room. A piece of paper with some "squiggles" drawn on it is passed into him. Searle has no idea what the squiggles might mean, or indeed if they mean anything. But he has a rule book, conveniently written in English, which tells him that certain combinations of squiggles should prompt him to write down specific different squiggles, which are then presented to the people on the outside. This is a very big book indeed – it describes appropriate responses to an extremely wide range of combinations of squiggles that might be passed into the room. Entirely unbeknownst to Searle, the squiggles are Chinese characters and he is providing intelligent answers to questions in Chinese. In fact, the room's pattern of responses is indistinguishable from that of a native speaker of the language. A Chinese person who knew nothing about the inner workings of the room would unhesitatingly credit it with an understanding of her language. But, says Searle, it is clear that neither he nor the room understands any Chinese. All that is happening is the manipulation of symbols that, from his perspective, are entirely without meaning.

What Searle says about the Chinese Room he thinks we should also say about computers. Computers, like the room, manipulate symbols that for them are entirely meaningless. These manipulations are directed, not by rule books written in English, but instead by programs. We shouldn't be fooled by the computer's programming into thinking it has genuine understanding – the computer carries out its tasks without ever having to understand anything, without ever entertaining a single thought. Searle's conclusions apply with equal force to early twenty-first-century laptop computers and to the purportedly super-intelligent computers of the future. Neither is capable of thought.

Defenders of strong AI have mustered a variety of responses to Searle.[2] I will not present these here. Instead I note that Kurzweil himself allows that we cannot be absolutely certain that computers are capable of all aspects of human thought. He allows that the law of accelerating returns may not bring conscious thoughts to computers. According to Kurzweil, the fact that "we cannot resolve issues of consciousness entirely through objective measurement and analysis (science)" leaves a role for philosophy (2005, 380). Saying that there is a role for philosophy in a debate is effectively a way of saying that there is room for reasonable disagreement. This concession leaves Kurzweil vulnerable to Searle's Wager.

One reason we may be unable to arrive at a decisive resolution of the debate between Kurzweil and Searle is that we aren't smart enough. In the final stages of Kurzweil's future history we (or our descendants) will become unimaginably more intelligent. It's possible that no philosophical problems will resist resolution by a mind that exploits all of the universe's computing power. But the important thing is that we will be asked to make the decision about uploading well before this stage in our intellectual evolution. Though we may then be significantly smarter than we are today, our intelligence will fall well short of what it could be if uploading delivers all that Kurzweil expects of it. I think there's a good chance that this lesser degree of cognitive enhancement will preserve many of the mysteries about thought and consciousness. There's some inductive support for this. Ancient Greek philosophers were pondering questions about conscious experience over two millennia ago. Twenty-first-century philosophers may not be any more intelligent than their Greek counterparts, but they do have access to tools for inspecting the physical bases of thought that are vastly more powerful than those available to Plato. In spite of this, philosophers do not find ancient Greek responses to questions about thought and consciousness the mere historical curiosities that modern scientists find ancient Greek physics and biology. Many of the conundrums of consciousness seem connected to its essentially subjective nature. There is something about the way our thoughts and experiences appear to us that seems difficult to reconcile with what science tells us about them. It doesn't matter whether the science in question is Aristotle's or modern neuroscience.

Searle's Wager

Searle's Wager treats the exchanges between Searle and his many critics in much the same way that Pascal's Wager treats the debate over God's existence. The availability of uploading will present us with a choice. We can choose to upload or we can refuse to. There are two possible consequences of uploading. The advocates of strong AI think that the computers we are uploaded into are capable of conscious thought. If Kurzweil is right, you will not only survive, but your powers of thought will be radically enhanced. If the doubters are right, then uploading is nothing more than a novel way to commit suicide.

Suppose we lack certainty on the question of who is right. We must place a bet. We can bet that Kurzweil is right and upload, or we can bet that Searle is and refuse to. For simplicity's sake let's suppose that if you accept the offer to upload yourself into the computationally superior electronic medium then you must consent to the destruction of the now obsolete biological original. It's not hard to imagine candidates for uploading consenting to this. Keeping the original around will seem to machine super-intelligences a bit like leaving an australopithecine former self hanging around to occasionally bump into and be embarrassed by. In the paper's final section I briefly explore the implications of uploading while leaving the biological original intact.

I propose that if you are not certain that Kurzweil is correct it is irrational to upload. My reasoning is summarized by the following table.

	Kurzweil is right. The Upload is both you and capable of conscious thought.	Searle is right. The Upload is incapable of conscious thought.
Choose not to upload	**[A] You live** You benefit from enhancements that leave your biological brain intact. You miss out on other more significant enhancements.	**[B] You live** You benefit from enhancements that leave your biological brain intact. You are spared death and replacement by a non-conscious Upload.
Choose to upload and destroy your biological brain	**[C] You live** You benefit from enhancements available only to electronic minds. Your life is extended. Your intellect is enhanced. You are free of disease.	**[D] You're dead** You are replaced by a machine incapable of conscious thought.

You may be about as confident that Kurzweil is right as you can be about the truth of any philosophical view, but if there is room for rational disagreement you should not treat the probability of Searle being correct as zero. Accepting, as Kurzweil does, that there is reasonable disagreement on the issue of whether computers can think leaves open the possibility that they cannot. This is all that the Wager requires. One difference between Searle's Wager and Pascal's Wager become apparent. An oft-made objection to Pascal's Wager is that it illicitly restricts alternatives. It fails to account for the possibility of self-effacing Gods who punish belief, to list just one possibility overlooked by the Wager. The alternatives for Searle's Wager are more clearly binary – one either survives uploading or one doesn't.

Epistemic modesty and the task of assigning probabilities to Searle's Wager

A second difference between Pascal's Wager and Searle's Wager now becomes apparent. The infinite reward for correctly believing in God means that we can bypass the philosophical debate about God's existence. We are absolved from attaching probabilities to God's existence or indeed of assessing the philosophical quality of any argument for the existence of God. So long as there's a nonzero probability of God existing then belief is supposed to bring an infinite expected return. For reasons fully explored in the following sections, neither option in Searle's Wager promises an infinite return. This means that we must make some attempt to attach probabilities to the propositions "Kurzweil is right" and "Searle is right." If the probability of Searle's conclusion being true is nonzero but tiny then even a very large but finite return for correctly not uploading may translate into a small expected return. Advocates of Searle's Wager, unlike advocates of Pascal's, must therefore pay attention to the quality of the arguments supporting its conclusion.

Poor arguments may fail to sufficiently raise the probability of the proposition that computers are incapable of thought.

How exactly should we assess the probability of Searle's conclusion being true? I suspect that a poll of working philosophers of mind, those who might be considered the relevant experts, would reveal some hostility to Searle's argument. The notion that computation can, in principle, fully capture the workings of human brains is sufficiently widespread among philosophers of mind to have the status of a dominant ideology. Those who subscribe to this dominant ideology might be tempted to assign a very low probability to the conclusion of Searle's argument. To them I recommend an appropriate epistemic modesty. When intelligent people have arrived at a conclusion that differs from your own you are rationally required to take seriously the possibility that they might be right. This is especially so when your opponents seem to be convinced by the quality of reasoning supporting their view rather than being motivated by some ideology that makes a certain conclusion morally mandatory or by a financial inducement. Although you may believe that Chinese Room argument to be unsound, you shouldn't reject its conclusion in the way that you might dismiss anti-Semites' arguments against the existence of the Holocaust or Big Tobacco's arguments in favor of health benefits from cigarettes.

Appropriate epistemic modesty means that even those who feel sufficiently confident about the falsehood of Searle's conclusion to assert this with gusto in philosophical debate should assign it a non-negligible probability of being true. While this may be less than fifty percent, it should be significantly higher than zero. The relatively high probability derives from the method by which philosophers arrive at their conclusions. Philosophers are typically not tasked with questions that have straightforward "yes" or "no" answers. A philosophical answer arrives only after a fraught process of evaluating and weighing reasons supporting alternative views. The option that receives endorsement is the one perceived to be supported by weightier reasons. Philosophers' expertise in representing this process by means of valid arguments, arguments whose premises logically guarantee conclusions, may obscure this process but it doesn't change it. Consider how this might apply to Searle's conclusion. We know of many examples of entirely unintelligent computation – that performed by calculators and cell phones are two examples. Searle presents the computation performed by future artificial intelligences as just more of the same. Kurzweil, on the other hand, thinks that additional power and sophisticated programming can make all the difference. Early twenty first century observers of this debate find themselves having to decide which of these alternative perspectives is more plausible. But neither is so absurd as to be rejected out of hand.

Suppose you decide that there's an eighty percent probability that Kurzweil is right and computers can think. You'd feel justified in vigorously supporting the claim in a philosophical exchange. Recast in a wager form, with its focus shifted from theoretical rationality to prudential rationality, this degree of credence is compatible with a, perhaps, twenty percent credence in the truth of the conclusion that you reject. I contend that a twenty percent (or lower) degree of credence in the proposition that Searle is right suffices to make uploading prudentially irrational.

Why death could be so much worse for those considering uploading than it is for us now (or why D is so much worse than B)

Pascal envisages our betting behavior being influenced by the fear of missing out on heaven. In Searle's Wager, the fear of death is operative. Perhaps there's a difference between these penalties that affects their power to motivate a bettor's choice. Death is a very bad thing for most of us, but it's not as significant a loss as missing out on an eternity in paradise. Indeed, for some people it may not be much of a loss at all. For example, a person about to expire from cancer can choose between certain death from disease and a merely possible death by uploading.

We shouldn't mistake our present circumstances for those of people presented with the option of uploading. Candidates for uploading are unlikely to find themselves stricken with terminal cancer and prepared to give the procedure a go. If the gerontologist Aubrey de Grey is right about the near future of our species, we could soon become ageless, immunized against cancer, heart failure or any of the other diseases that might incline us to disregard caution about uploading (de Grey and Rey 2007). He claims that there's a good chance that people alive today will achieve millennial life spans. They'll do this by systematically fixing up their brains and bodies, i.e. without recourse to uploading.

There's a keenly argued debate over the viability of de Grey's vision of the future.[3] My argument here does not depend on the veracity of this and others of de Grey's claims. Rather, it relies on the achievability of his plan *relative* to that of uploading. De Grey's immediate goal is something he calls Longevity Escape Velocity (LEV). LEV doesn't require full and final fixes for heart disease, Alzheimer's, cancer, and the other conditions that currently shorten human lives. What's essential is that we make appreciable and consistent progress against them. LEV will arrive when new therapies reliably add more years onto our lives than the time it takes to research them. According to de Grey, anyone who is alive at this time and has access to the full range of therapies should expect a millennial life span. New therapies will keep on coming, granting additional years faster than living consumes them. My point here requires only that LEV is likely to arrive sooner than uploading. Uploading requires not only a completed neuroscience, total understanding of what is currently the least well-understood part of the human body, but also perfect knowledge of how to convert every relevant aspect of the brain's functioning into electronic computation. It's therefore likely to be harder to achieve than LEV.

This guess about future technological development could be wrong. Suppose we discover how to upload humans well in advance of achieving longevity escape velocity. Then it's possible that candidates for uploading may be facing slow deaths from untreatable diseases. They could be diagnosed with terminal cancer and properly view themselves as having nothing or very little to lose from uploading.[4] In these circumstances uploading could be prudentially rational for some people. For those who aren't terminally ill it will still make sense to direct their hopes and expectations toward the relatively risk-free life extension and quality of life enhancement brought by LEV. Moreover, the issue of what is rational for those who know how to upload but have no expectation of achieving LEV is different from that which we face now. Given what we know about the relative technological challenges of uploading and achieving LEV we should expect the former to postdate the latter. Any recommendations about research in artificial intelligence should place more weight on this more likely pattern.

I conclude that people presented with the option of uploading are unlikely to find that they've got little to lose should the procedure fail to transfer their minds into machines. They'll be loath to renounce the variety of enhancements compatible with the survival of their biological brains.

Why uploading and surviving is not so much better, and possible worse than refusing to upload (or why A is not so much better than, and possibly worse than C)

Perhaps you'll miss out on a great deal if you choose to upload and Searle happens to be right about the procedure's consequences. But the potential gains could be truly massive. Uploading opens up enhancements much more dramatic than those made possible by the comparatively few nanobots and neuroprostheses properly deemed compatible with the biological brain's survival. If Kurzweil is right, freed of biological limitations we can become massively more intelligent. The gap between enhancements compatible with the survival of the brain and those enabled by uploading and incompatible with its survival is likely to be very large.

To begin with, the uploaded mind will be more an upgrade than a copy. The electrochemical signals that brains use to achieve thought travel at one hundred meters per second; this sounds impressive until you hear that electronic signals in computers are sent at three hundred million meters per second. This means that an electronic counterpart of a human biological brain will think "thousands to millions of times faster than our naturally evolved systems" (Kurzweil 2005, 127). But there's a more theoretical reason why Kurzweil believes that "[o]nce a computer achieves a human level of intelligence, it will necessarily soar past it" (2005, 145). Computers are technology. Improvements to them are governed by the law of accelerating returns. Although biological brains may improve over time, they're subject to the dramatically slower, intergenerational schedule of biological evolution. Kurzweil thinks that machine minds will learn how to fully exploit the computational potential of matter and energy (2005, 29). They will cannibalize ever-increasing quantities of the previously inanimate universe, reconfiguring it to enhance their powers of thought. According to Kurzweil, "[u]ltimately, the entire universe will become saturated with our intelligence. This is the destiny of the universe. We will determine our own fate rather than having it determined by the current 'dumb' simple, machinelike forces that rule celestial mechanics" (2005, 29).

Suppose that we accept that the enhancements compatible with the brain's survival are likely to be significantly more modest than those enabled by uploading. What should those who are considering uploading make of this gap?

Consider the measures taken by economists to convert the objective values of various monetary sums into the subjective benefits experienced by individuals. For most of us, a prize of $100,000,000 is not 100 times better than one of $1,000,000. We would not trade a ticket in a lottery offering a one-in-ten chance of winning $1,000,000 for one that offers a one-in-a-thousand chance of winning $100,000,000, even when informed that both tickets yield an expected return of $100,000 – $1,000,000 divided by 10 and $100,000,000 divided by 1,000. The $1,000,000 prize enables you to buy many of the things that you want but that are currently beyond you – a Porsche, a new iPod, a modest retirement cottage for your parents, a trip to the pyramids, and so on. Many of us also have desires that only the higher reward will satisfy – a mere million won't buy a luxury Paris apartment or a trip to the International Space Station aboard a Soyuz rocket, to give just two examples. So we have no difficulty in recognizing the bigger prize as better than the smaller one. But we don't prefer it to the extent that it's objectively better – it's not one hundred times better. The conversion of objective monetary values into subjective benefits reveals the one-in-ten chance at $1,000,000 to be significantly better than the one-in-a-thousand chance at $100,000,000.

I think that the subjective significance of the gap between enhancements compatible with the brain's survival and those incompatible with it is unlikely to match its objective magnitude. In fact, there may not be too much of a gap to those considering uploading between the appeal of the objectively lesser enhancements compatible with the survival of our brains on the one hand, and the objectively greater enhancements enabled by uploading on the other. The more modest enhancements will satisfy many of the desires of people for whom uploading is an option. They will live significantly longer, be freed of disease, play better bridge, learn foreign languages with ease, and so on. Uploading may enable feats well beyond those whose biological brains are supplemented with only those electronic chips deemed compatible with the brain's survival, but we have comparatively few desires that correspond specifically with them. There's a reason for this. Desires are a practical mental state – they motivate people to act in certain ways. Those who want to get fit do exercise; those who want to lose weight go on diets; and so on. Desiring radical enhancement is a matter of placing your faith in the law of accelerating returns and waiting for something quite miraculous to be done to you. There's little you can do about it now beyond reading and rereading Kurzweil's book *The Singularity Is Near* and enrolling in your local chapter of the World Transhumanist Association.

One way to adjust the subjective values of $1,000,000 and $100,000,000 is to make the choice from a standpoint of considerable wealth. Donald Trump is likely to look on the smaller sum as barely enough to achieve anything worthwhile. The larger sum, on the other hand, may suffice to acquire some significant piece of Manhattan real estate. This could lead him to prefer the one in a thousand chance to acquire $100,000,000 to the one in ten chance to receive the mere million. Is there an analogous move that can be made in respect of enhancement? It's true that those who've already achieved super-intelligence are likely to be more impressed by the objectively greater enhancements possible after uploading than we are. The problem is that it's impossible for us to adopt this standpoint. We're necessarily deciding about uploading in advance. Compare: We might be able to imagine Trump's contempt for a mere million dollars; but in advance of actually acquiring his wealth we're unlikely to be motivated by this imaginary contempt.

I suspect that the problem may actually be more serious than that signaled by the previous paragraphs. The manner of radical cognitive enhancement permitted by uploading may be worse than the more moderate variety compatible with the survival of our brains in the light of some of our more significant desires. Many of the things that we desire may be contingent on our current level of cognitive powers. We want to protect our relationships with our loved ones. We want to promote and honor our strongest moral and political ideals. Radical enhancement may not remove our capacity to protect, promote, and honor these commitments. But it may remove our desire to do so. I suspect that most parents find equally chilling the prospect that, at some point in the near future, they won't be able to protect their children and the prospect that, at some point in the near future, they won't *care* about their children's welfare. Concern about doing the things we currently most want to do may, therefore, lead us to place a low value on radical cognitive enhancement.

What does all this mean for those presented with the option of uploading? I suspect candidates will prefer more modest, safe enhancements to those whose potential magnitude is greater but come with a risk of death.

Can we test the hypothesis that uploading preserves the capacity for conscious thought?

Might it be possible to test the hypothesis that someone could survive uploading? We don't yet have such a test. But if a test becomes available soon enough to inform decisions about whether or not to upload, then, contingent on its results, candidates for uploading may be more confident that they'll survive the process.

Suppose that some time between now and the advances required to upload, neuroscientists identify all the parts of the brain responsible for mathematical calculation. They might carefully excise them from your brain, perfectly preserving them so that they could be reintroduced after the experiment. The removed parts would be replaced by an electronic implant programmed to perform calculations using techniques that are similar to those used by humans. You are then requested to perform some calculations. The experimenters would ask you whether or not you are conscious of multiplying 5 by 5 to arrive at 25. Have you noticed any differences between your current and your former experience of multiplication?

It's not clear that you would notice any such difference even if Searle is right about how this implant works. It is unclear that there is any distinctive variety of conscious thought associated with multiplying. There is, in contrast, a distinctive kind of awareness associated with arriving at the answer 25 and preparing to say "5 times 5 is 25" but this awareness would be shared in the case in which the answer arrives by neuronal computation and the case in which the neuroprosthesis presents it to consciousness. Perhaps there are other opportunities for differences in awareness. For example, you may be aware of stages on the way to arriving at the answer. Suppose you consciously decide to arrive at the answer by the cumbersome route of starting with 5 and adding four lots of 5 to it. You may become conscious of a series of intermediate stages in the calculation. For example, having resolved to add four lots of 5 you become aware that the first addition gives the answer 10, that the second gives 15, and so on. But it would be hard to distinguish the hypothesis that the implant's calculation is a form of thinking from the hypothesis that it is noncognitively number-crunching and presenting bulletins of its workings to the conscious part of your mind.[5]

It's possible that the experience of gradually replacing parts of your brain with electronic prostheses would be a little like neurodegeneration. Alzheimer's disease progressively destroys parts of its sufferers' brains. People tend to become aware of the consequent loss of cognitive skills only when they fail to perform tasks that were once well within their grasp. Suppose Searle is right about the impossibility of machine thought. The gradual replacement of systems in your brain with electronic equivalents would be like Alzheimer's in one significant respect. It would involve the progressive erosion of your powers of thought. The difference would be that you would lack an external check on this loss. You would be like a person with Alzheimer's who has an electronic assistant that informs her of the names of people whom she has forgotten. The Alzheimer's patient is probably aware of the assistant. You, in contrast, may be entirely oblivious of the operation of your implanted assistant. The fact that you would perform all the tasks you used to be able to perform would deprive you of any awareness of the ongoing shrinkage of your mind.

Perhaps there's another way to test the hypothesis that uploads can think. We could ask them. There are likely to be plenty of people to ask. Just because it's irrational to upload doesn't mean people won't do it. People do many things that are irrational.

"Uploads" will do a great many things strongly indicative of conscious thought. They will insist that they are conscious. In his response to Searle, Kurzweil presents these claims as a strong pragmatic justification for accepting the consciousness of computers we will become, or will attempt to become:

> these nonbiological entities will be extremely intelligent, so they'll be able to convince other humans (biological, nonbiological, or somewhere in between) that they are conscious. They'll have all the delicate emotional cues that convince us today that humans are conscious. They will be able to make other humans laugh and cry. And they'll get mad if others don't accept their claims. (Kurzweil 2005, 378-79)

Purged of its question-begging language, this is exactly what Searle expects. In his thought experiment, something that works like a computer exactly matches the linguistic behavior of native speakers of Chinese without understanding a single word of Chinese. He allows that a computer installed in a robot might perfectly replicate human conscious behavior without ever entertaining a single conscious thought. The machines' skill at convincing people is no more a hallmark of truth than is a subprime lender's capacity to convince home purchasers that they'll cover the mortgage repayments with ease.

The computers that have replaced those who have chosen to upload may be sincere in their protestations of consciousness – either that or they could perfectly simulate sincerity. Human beings who are being insincere are typically aware that they're acting deceptively. There would be no such awareness in your computer replacement. Suppose someone were to challenge your claim that you are conscious. This insult might prompt you to introspect on the workings of your mind, an activity that leads you to vigorously assert that you really are conscious. Now consider the uploaded version of you. It will go through the processes that are the exact machine equivalents of this introspective gathering of evidence. But if Searle is right, this conviction will be the consequence of entirely nonconscious, nonrational manipulations of symbols.

Is it really so unlikely that computational efficiencies mandated by the law of accelerating returns will have detrimental consequences for consciousness? Consider one of the chief benefits of the switch from electrochemical to electronic processing. Kurzweil presents the latter as dramatically faster than the former. An electronic copy of your mind that made no improvements to the ways in which you approached problems would think a million thoughts for your one. It's possible that this gain in speed imperils your conscious experiences. There is much that our brains do that we are unaware of. They compute the angles and velocities of tennis balls enabling us to just reach out and catch them. They make adjustments to the positioning of our bodies that prevent us from falling over. They translate a diverse collection of perceptual inputs into the impression of a baby crying. The thing that these otherwise very different kinds of mental processing have in common is that they are very swift relative to conscious thought processes. Could there be an absolute upper limit in processing speed above which consciousness cannot be sustained? The mystery of consciousness precludes a definitive answer to this question. If the answer is yes, then Kurzweil's proposal to speed all of our thinking up may completely banish consciousness from our minds. None of the kinds of thoughts that we currently entertain will last long enough to be conscious. Of course one might obviate this particular concern by slowing down the electronic circuits. But this would eliminate much of the motivation for the conversion from electrochemical to electronic computation.

Why we shouldn't make electronic copies of ourselves

So far I've argued for a relatively restricted claim. It is unlikely to be rational to make an electronic copy of yourself and destroy your original biological brain and body. This leaves open the possibility of making the copy and *not* destroying the original, an option that might seem to offer the best of both worlds. Humans do not need to risk death. They can copy themselves into machines that will then be free to follow the law of accelerating returns all the way to the Singularity.

This raises questions about the relations between what Kurzweil refers to as mostly original substrate humans – beings that are substantially biological – and beings that are either super-intelligent, or who act as if they are. In what follows I use the prefix "quasi" to summarize Searle's view about the intelligences that will emerge from the law of accelerating returns. Accordingly, if Searle's right then our future could be shared with quasi-super-intelligent beings. Elsewhere I defend a pessimistic conclusion about the likely relations between humans and super-intelligent or quasi-super-intelligent beings (Agar 2010, chapters 4 and 8). Here I confine myself to a single warning about these relations.

Francis Fukuyama is pessimistic about relations between enhanced and unenhanced beings. He connects liberal democracy with an approximate empirical equality between citizens. Accordingly, he asks "what will happen to political rights once we are able to, in effect, breed some people with saddles on their backs, and others with boots and spurs?" (2002, 10). Perhaps supremely intelligent machines will see no value in liberal social arrangements and hence take advantage of those who have chosen not to upload. The transhumanist thinker James Hughes (2004) is more optimistic. He argues for a *democratic transhumanism* that would require the equal treatment of all persons. The citizens of societies organized according to the principles of democratic transhumanism will understand that the vast differentials in power between the unenhanced and the enhanced have no bearing on their personhood. Hughes thinks that democratic transhumanism can ensure a harmonious future for societies that comprise individuals who are making the transition from humanity to posthumanity at varying speeds, or not at all.

I think that there's little chance of a stable social compact between uploaded beings and what Kurzweil refers to as mostly original substrate humans or MOSHs. The powers of MOSHs are likely to be static or to be increasing only gradually. If the Law of Accelerating Returns applies to machine minds then uploaded beings will be becoming more intelligent and powerful at an ever increasing rate. It's possible that Hughes's democratic transhumanist principles may be successfully implemented in societies in which the gaps between enhanced and unenhanced are both moderate and relatively constant. Democratic transhumanism may spare the unenhanced and the first generations of enhanced from the depressing outcomes depicted in the movie, *Gattaca*, for example. But there's room for doubt about the prospects for a stable social compact between beings whose intellectual and physical powers are static and beings whose powers are increasing at an exponential rate.

I have argued that people presented with the option of uploading their minds into computers are likely to judge it prudentially irrational to do so. They will assess the expected rewards from uploading as insufficient to justify the risk of death. Death denies them futures in which technologies compatible with the survival of their biological brains rejuvenate them and enhance their powers of thought. Our best guesses about these future choices have implications for our support of contemporary research in artificial intelligence. We should proceed with caution.

Acknowledgements

I would like to express thanks to Neil Levy, Mark Walker, and Russell Blackford.

Notes

[1] Ray Kurzweil presents his vision of humanity's future in Kurzweil 1990, 2000, and 2005. Kurzweil's Web site is also essential viewing: http://www.kurzweilai.net/.

[2] For Kurzweil's responses to these criticisms, see Kurzweil 2002.

[3] For one skeptical response see Estep, et al. 2006: 84.

[4] Thanks to Mark Walker for making this point.

[5] Daniel Dennett (1991) explores a variety of puzzles such as these concerning phenomenal consciousness. He concludes that the notion of phenomenal consciousness is incoherent. His arguments may make those who reject Searle's Wager more confident about uploading. But they should not be viewed as reducing to zero either the probability that biological human brains support phenomenal consciousness or that uploading destroys this capacity.

References

Agar, Nicholas. 2010. *Humanity's end: Why we should reject radical enhancement*. Cambridge MA: MIT Press.

de Grey, Aubrey, and Michael Rae. 2007. *Ending aging: The rejuvenation breakthroughs that could reverse human aging in our lifetime*. New York: St Martin's Press.

Dennett, Daniel C. 1991. *Consciousness Explained*. Boston: Little, Brown.

Estep, Preston W., III, Matt Kaeberlein, Pankaj Kapahi, Brian K. Kennedy, Gordon J. Lithgow, George M. Martin, Simon Melov, R. Wilson Powers III, and Heidi A. Tissenbaum. 2006. Life-extension pseudoscience and the SENS plan. *MIT Technology Review* 109: 3: 80-84.

Fukuyama, Francis. 2002. *Our posthuman future: Consequences of the biotechnology revolution*. New York: Farrar, Straus & Giroux.

Hughes, James. 2004. *Citizen cyborg: Why democratic societies must respond to the redesigned human of the future*, Cambridge MA: Westview.

Kurzweil, Ray. 1990. *The age of intelligent machines*. Cambridge, MA.: MIT Press.

—. 2000. *The age of spiritual machines: When computers exceed human intelligence*. London: Penguin.

—. 2002. Locked in his Chinese Room: Response to John Searle. In *Are we spiritual machines?: Ray Kurzweil vs. the critics of strong A.I.*, ed. Jay W. Richards. Seattle, Wash.: Discovery Institute: 128-167.

—. 2005. *The Singularity is near: When humans transcend biology*. London: Penguin.

Searle, John. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3, no. 3: 417-57.