SEEKING NORMATIVE GUIDELINES FOR NOVEL FUTURE FORMS OF CONSCIOUSNESS

BRANDON OTO[†]

University of California, Santa Cruz [*alumnus*]
1156 High St
Santa Cruz, CA 95064, USA
brandon@degreesofclarity.com

Received 19 June 2011
Revised 26 December 2011

The potential for the near-future development of two technologies—artificial forms of intelligence, as well as the ability to "upload" human minds into artificial forms—raises several ethical questions regarding the proper treatment and understanding of these artificial minds. The crux of the dilemma is whether or not such creations should be accorded the same rights we currently grant humans, and this question seems to hinge upon whether they will exhibit their own "subjectivity," or internal viewpoints. Recognizing this as the essential factor yields some ethical guidance, but these issues need further exploration before such technologies become available.

Keywords: uploading, transfer, teleportation, teletransporter, AI

1. Consciousness in the Singularity

Ray Kurzweil [2005] has proposed a timeline for human technological development that projects radical advances in computation, artificial intelligence, and medical augmentation within the next fifty years. Dubbing this period of radically accelerating change the "technological Singularity," he describes a litany of breakthroughs that will, among other things, allow for the creation of wholly artificial forms of consciousness, as well as the scientific manipulation of our existing minds. If we accept his predictions, even as remote possibilities, then several known but unresolved philosophical questions shift from "a centuries-old philosophical dialogue to a pressing practical matter."

Two distinct scenarios are described by Kurzweil. The first is the construction of a non-human, artificial brain.

> . . . we will have completed the reverse engineering of the human brain, which will enable us to create nonbiological systems that match and exceed the complexity and subtlety of humans, including our emotional intelligence. . . . nonbiological entities will claim to have emotional and spiritual experiences, just as we do today. [They] will claim to be human . . .

[†] 23 City View Rd, Brookline, MA 02446

Realizing this old dream of strong artificial intelligence, or AI, would have many profound implications on society and science. Just how should we view these unprecedented new creations?

The second scenario is the reconstruction or "uploading" of our existing minds into other forms.

> Uploading a human brain means scanning all of its salient details and then reinstantiating those details into a suitably powerful computational substrate. This process would capture a person's entire personality, memory, skills, and history.

Again, such a process would be wholly novel, and unclear in its use and implications. What should the relationship between oneself and one's uploaded form amount to?

Many complex questions surround these issues, and some are wholly in the domain of the philosopher. Concerning AI, we might ask: Can they truly think?[a] Can they have experiences?[b] Are they conscious?[c] And concerning mind uploading, most of the same questions come to bear, plus a vexing problem of personal identity—for we will wish to know if an uploaded copy of our mind amounts to a copy of us, or actually *is* us.

Ample debate has surrounded all of these questions, and room remains for many years and pages more. However, if we accept Kurzweil's view that these events will come to pass within the foreseeable future, we might lay aside many of the deeper ontological questions in favor of a more practical and immediate problem, which is: What is the normative status of these new technologies? How *ought* we behave with regard to them?

I propose that although much of the depth and breadth of these topics remains to be explored, if we approach them with the modest, pragmatic goal of informing our behavior, we might steal from the heart of the matter some answers, while leaving the larger concerns as fodder for the ongoing philosophical debate. We can look for "lean" answers, those that correctly prescribe action for the exact dilemmas posed, without attempting to resolve the grand philosophical questions within which they sit.

2. Exploring the Dilemmas

Our interest, then, lies with the normative questions of Kurzweil's world. What are those questions?

The fundamental ethical concern with an AI lies in its status as an individual. To simplify matters, let us consider the case of an artificially-created mind with the same overall computational power as a human brain, as well as similar overall structure and behavior. Call it Cybo. Clearly, Cybo is not a human being; he may resemble one (if we

[a] For two of the classic cases explicitly dealing with AI, see Turing [1950] and Searle [1980].
[b] Jackson [1982] for one typical approach; Dennett [1993] for the usual manner of rebuttal.
[c] The debate surrounding the philosophical zombie (promulgated not originally, but certainly most elaborately, by Chalmers [1996]) treads this terrain exhaustively, even when not explicitly discussing AI.

build him into a humanoid robotic form, for instance), or he may not (perhaps he dwells in the instruments of an aircraft), but he was not produced by the biological mechanism that defines *homo sapiens*. However, we can ask the question: Should Cybo be *treated* like a human being?[d]

This is a vague question that begs for clarification. After all, humans are treated in a wide range of ways, depending largely on their own character and behavior, and this might apply equally to an AI. However, there is a common thread between prince and pauper, martyr and murderer, and that is our sense that they all have claim to certain *rights*.

Chief among these is that any human being has the right not be killed. In other words, the act of a human dying is a moral negative; other things being equal, it would be better for him to have survived. This need not be considered indefeasible; we can still judge other considerations as important enough to sometimes outweigh this right (for instance, the necessary execution of a criminal, or compassionately ending a life of chronic suffering). Although the validity of this "right" might be debated, it is enough for our purposes to accept that intuitively, nearly everyone today finds it to be at least generally true. Where disagreement occurs, it most often finds traction in the ambiguity of just how widely or narrowly the right should obtain. For instance, do non-human animals warrant this right? Do unborn fetuses? What about individuals with incurable illnesses that have degraded them physically and mentally? Even acts of atrocity, such as murder or genocide, typically occur through the individual negation or minimization of this right (the viewing of one's victim as "not human," for instance) rather than a lack of recognition for its existence in the first place.

We might expand this category of basic human rights to include, say, the right not to endure pain, or the right to act freely without undue constraint. Such rights might all arise from the same source, or perhaps some can obtain without others. For the sake of our discussion, we will consider only the most basic of these: the right not to be killed. Call this a *right to life*.

Does Cybo have a right to life? Like the question of animal rights, this is fundamentally the question of whether he possesses whatever essential quality or virtue confers this same right to humans. So to answer this question, we will have to identify that virtue. It might be as unique and irreducible as "being human," or much more general, but in order to perceive Cybo's ethical status, we must assay whether he does or does not have this criterion.

Importantly, this is not the question of *why* such a virtue ought to confer a right to life. That is a deeper question—one that can be asked equally of humans—and rather than examine it here, we will simply assume the relationship is sound. We are not attempting to establish that humans have a right to life, simply whether, if they do, AIs do as well.

Now, what of the question of mind uploading? The principle technology here is somewhat similar to the creation of AI. Using advanced imaging technology, scientists scan my brain (or if necessary, my entire body), analyzing it down to the smallest particle

[d] Kurzweil [2005] raises but does not answer this question.

and digitizing that information as a stored blueprint. They then use that blueprint to reinstantiate my form into a new, artificial manifestation. Let this be a "wet" form, such that it comes out looking very much like a human being, with gray brains and soft skin; or let it be a "hard" form, such that it comes out looking like a robot, with metal surfaces and a microchip processor; or even let it be a "soft" form, such that it exists only in the memory banks of some kind of digital storage. Call this new creation Metoo.

Metoo is an artificial intelligence—he is artificial, in the sense that we created him by a non-human method, and he is intelligent, to whatever extent I am intelligent—and we can ask the earlier question regarding his right to life. But we can ask another question, too: is Metoo someone just like me, or is he *me?* In the parlance of identity theory, is he only *qualitatively* identical to me—resembling me exactly and possessing the same qualities, like one carbon atom resembles another? Or is he also *numerically* identical to me—being, in fact, the same entity, like one carbon atom resembles *itself*, such that "Metoo" and "me" are just two ways to refer to the same individual?

This is not an easy question to answer, nor even an easy question to ask, given the vagueness and mutability of the words and concepts involved. After all, it is far from settled what identifies "me" even before uploading, and it seems futile to try to compare two entities when we lack a full understanding of each. Yet in Kurzweil's Singularity, this uploading technology would actually be available, and whether the philosophical problems are resolved or not, we will have to make certain decisions. For instance, we might ask whether Metoo has the same legal claim to my property as I do. If I have a job, does he have one, or is he unemployed? Is he married to my wife? These questions are important, but our concern here is with a deeper one: In what cases should I regard the survival of Metoo as tantamount to my own survival?

Kurzweil explains:

> . . . we will ultimately be able to upload this pattern to replicate my body and brain to a sufficiently high degree of accuracy that the copy is indistinguishable from the original. (That is, the copy could pass a "Ray Kurzweil" Turing test.) . . .
>
> Although the copy shares my pattern, it would be hard to say that the copy is me because I would—or could—still be here. You could even scan and copy me while I was sleeping. If you come to me in the morning and say, "Good news, Ray, we've successfully reinstantiated you into a more durable substrate, so we won't be needing your old body and brain anymore," I may beg to differ.
>
> If you do the thought experiment, it's clear that the copy may look and act just like me, but it's nonetheless *not* me. I may not even know that he was created. Although he would have all my memories and recall having been me, from the point in time of his creation Ray 2 would have his own unique experiences, and his reality would begin to diverge from mine.

If we thus are presented with the technology to upload our minds, we may find there to be many advantages to this—for instance, we might enjoy a newfound ability to easily expand and upgrade our mental capacity, just like we upgrade our personal computers. And one truly unique benefit stands out: an uploaded mind could escape, or at least

mitigate, the possibility of biological death. An individual with a mortal illness might transfer his mind into a healthy body; the victim of a freak accident might find solace in reverting to a "saved" version of himself held in backup. Or by scanning oneself in Boston, destroying the original, and then reinstantiating the pattern in Paris, we might achieve easy and near-instantaneous travel.[e] However, all of these supposed advantages hinge upon one key point: Metoo must actually be numerically identical to me. If he is not, then there is no sense in talking about my surviving a mortal illness or traveling to Paris, because I am merely dying, and someone qualitatively identical to me is taking my place. We can ask whether Metoo has a right to life, but personally, what I will want to know before being uploaded—especially if we plan to destroy the original—is whether I, the original, am about to die. Hopefully, the answer is "no," because then we can reap the full benefits of these new technologies; but most important is simply that we know the answer beforehand.

If we are classical Cartesian-type dualists about the mind, then we would likely answer that my essence is not transferred when I am uploaded. If we hold that what is essential about "me" is something like a non-physical soul, then it is doubtful my soul would be conveyed in the process of uploading; uploading, after all, is simply a process of remotely recreating my *physical* pattern. This is one answer to our dilemma, but for the purposes of this discussion we will avoid it, instead remaining generally in the realm of the physicalists; it is better suited to the context of the Singularity, and dualism tends to end rather than encourage further discussion.

Even physicalist approaches to the mind, though, are too many to count, and it is not clear which we should apply. Can we answer the normative question—how should I view the uploading of my mind?—without requiring a model of consciousness with which to examine it? Let us make the attempt, by boiling down the question to a much leaner one indeed.

3. The Essential Virtue

The question of artificial intelligence and the question of mind uploading both hinge upon the same point. Why do we value our human lives? When we observe a right to life, by what criterion do we confer that right to humans, yet not to potatoes? And when I consider uploading my mind, what is it that I seek to preserve and "carry over"?

Is it my physical body? Certainly we do value the bodies we have, but seemingly only as a tool. I would object to my left leg being needlessly amputated, but only because its loss would make my life more difficult, much the same as if my car broke down. The same goes for my various organs, my sensory apparatus, even essential pieces like my heart. Cut away everything and I suppose that I will still be here (although not for long), so long as you leave my brain; but there is no bargain by which you could compel me to

[e] This standard thought experiment is usually called the "teletransporter case," and although making many appearances in science fiction, it has also become a standby in discussions of personal identity. Parfit [1985] gives one of the better descriptions. The teletransporter itself is not one of Kurzweil's predictions, but all of the necessary principles are.

cut away my brain.[f] So what we value most must be associated with the brain, either the thing itself or some quality thereof.

Could it be something wholly unphysical, a disembodied soul of some kind? Possibly, but dualists have their own set of problems, and few now overtly carry their flag. As mentioned, we will skip over this view here.

Perhaps what we value is the human *mind*. But what do we mean by that? Is it intelligence—a certain capacity for computation, for manipulating information? This makes some sense, as it is our intelligence that seems to set us aside from other animals, and certainly from potatoes—but it is at least partially false, because we do not yet ascribe moral privilege to our personal computers, and in certain functions those already radically exceed human computational capacity. Moreover, we recognize that some humans are endowed with greater intelligence than others, and although this is considered admirable, they do not earn any more points *qua* human beings; they have no greater right to life, unless we add extrinsic considerations like their value to society. It remains to be seen just how closely artificially-constructed intelligences will resemble human ones, but at least it seems clear that the ability to dumbly process data or perform equations should not, on its own, be sufficient to grant them status beyond that of useful devices.

What about *consciousness*? When we invoke this word, we seem to mean a certain sort of self-awareness or sentience, and although we will have to decide exactly what we mean by this, it seem much closer to the desired virtue. Both fools and geniuses are conscious, or so we believe, but not a pocket calculator, and perhaps not even a supercomputer. As a result, we ascribe a certain moral status to the former that is withheld from the latter.

Consciousness alone does not satisfy our interest altogether, however, because although I am conscious, and you are conscious, we are two distinct consciousnesses, and this separation is essential to the thing itself. No matter how similar we two might be, fundamental to the concept of consciousness is the fact that we each have our own; that is, we are numerically distinct. So although consciousness is a necessary virtue, it is not enough to group together all consciousnesses; they are not fungible. My consciousness is not interchangeable with yours, despite the fact that we both may otherwise have the same status (we both claim the same right to life, for instance).

So the key matter is that we place a certain moral value on the existence of individual consciousnesses. Let us call each such individual a *person.* Why do we value persons? Again, it is not because of any extrinsic worth they have—or it may sometimes be, but there is a more common denominator. A useful man has many things that an inept one lacks, but their basic moral rights are equal, we hope.

To help unwind the quality that distinguishes and gives value to persons, consider Kurzweil's example of the non-destructive duplicate. Upon awakening from a deep sleep, scientists inform me that overnight they scanned my brain in all its particulars, and in the early hours of the morning their laboratory executed a rush production and built a Metoo,

[f] Most people from laymen to physicians would find this intuitively true, but see the work of Eric Olson on animalism for a differing perspective.

housing an exact copy of my brain. A scientist hands me a gun—"No point in having two of you around, so you'll just be wanting to do away with this old substrate, I suppose." I cry foul; I want no such thing. Why not?

The reason I demur is because I still find myself inhabiting the original form. There is nothing *qualitative* that is lacking about the new Metoo; he is, natch, exactly as good as me in body and mind. If I have intelligence, so does he; whatever memories I possess, he also has; my personality is his personality. Granted, as time passes he will begin to lead a divergent life, but at least at the moment of my scanning we can be said to be qualitatively identical. I cannot say that he is not as good, or even that he is different; he is the same. Yet we *are* distinct, and that is why I hesitate to obliterate my old consciousness. The one, singular difference we have is that our viewpoints differ.

I am lying in bed; I see gray walls around me, photos of my family, I feel the linen sheets and I hear the soothing words of the scientists. I am the endpoint of all my various external senses, along with an internal "sense," the introspective ability to know my own thoughts.[g] As for Metoo, he is elsewhere, experiencing other things; he could call me on the phone any moment and we could discuss what was transpiring in each of our lives, but they are nonetheless distinct perspectives. If he were struck dead, I would find it unfortunate, but my life would march on. In short, the essential difference that distinguishes our two instances of consciousness—the virtue that makes us two persons rather than one—is that we manifest different loci of subjective experience.[h]

The things I care about are the things that concern "me," that conscious person who houses my personal point-of-view. I would frown on being shot in the head because those negatives will be experienced *by* me. In Thomas Nagel's words [1974], there is something it is "like" to be me; there is some sort of viewpoint housed among my hardware and software, capable of experiencing my unique input.[i] Call this *subjectivity.* This is what we chiefly mean by "consciousness"; when we discuss different conscious persons, these viewpoints are the tokens we are counting; and this is what we strive to safeguard when we uphold a right to life. This is also what we are assaying for in so-called intelligent machines; if a calculator seemed to possess subjectivity, seemed to be home to a consciousness with its own reality, then that would be the essential feature we should recognize as conferring personal rights and recognition.

The claim I am making is extremely limited. Just what one's "subjectivity" amounts to is a topic of wide debate. Indeed, the dualist camp we discarded is one of the only theories that lets us call subjectivity any sort of further fact in the mind; most contemporary accounts are happy to call this phenomenon an illusion, or a construction,

---

[g] Ned Block [1995] suggested the useful distinction of the phenomenal-consciousness versus the access-consciousness. What I am describing here is an amalgam of p-consciousness with some, but not all, of Block's idea of a-consciousness. Simply put, insofar as internal thoughts and narration are available to the "sense" of introspection, I consider them as analogous to external phenomena for the purposes of creating subjective perspective.

[h] Dennett's adroit 1978 story "Where Am I?" gives an interesting examination of a scenario where sensory input is less proprietary. Although intriguing, I do not think that his example contradicts my argument.

[i] I disagree with much of Nagel's commentary on reductionism, but the language he introduced remains especially useful for discussions about experience.

or a useful fiction, or at best an emergent sort of pattern.[j] Any or all of these views might be accurate, or a combination thereof; we do not need to invoke a literal Cartesian theater. But I do *seem* to have subjectivity; intuitively I believe, and cannot help but believe, that I have a unique perspective which is host to my specific experiences. What does this actually amount to? I do not know. But whatever it is, that phenomenon is what we elevate in the search for consciousness. For the purposes of our investigation, let it remain a black box; let it be real, or imaginary, or whatever you will, but *whatever it is we currently have, that is what we wish to preserve.*

One might object that there is at least one other good candidate for our desired virtue—one other way in which I differ from the duplicate Metoo. This is in our agency, our individual originating points for action; we each have our own unique ability to make decisions and execute them. We each have our own free will, so to speak, and although those actors may be similar in quality, they are essentially distinct.

However, although this is a very valued trait of the human mind, and perhaps essential for what we know of the human condition, I think it is not our true area of concern here. Imagine the case of a man who suffers from *total locked-in syndrome*, a neurological disorder in which the body is utterly paralyzed from head to toe. No control over any voluntary muscles is possible, including, in the most extreme cases, the muscles that manipulate the eyes. However, such patients are often otherwise cognitively intact, and enjoy normal functioning of their sensory apparatus. They feel, hear, smell, and so forth, and they see through eyes that they cannot direct; their thoughts are still their own, although they cannot express them.[k] This is a tragic situation, and we might consider it barely living. But we do not consider it like being dead. Since these patients still do have subjective *experience*—there is still something it is like to be them, although it is something rather unfortunate—we do not find it acceptable to abuse or murder them. They are still valid persons, and indeed, they are still the persons they were prior to their illness. Their loss of agency is not tantamount to loss of consciousness, so long as they continue to demonstrate subjectivity.[l]

This point is further demonstrated when we try to imagine a counter-case—an individual with agency, but no experiential input whatsoever. This is almost an incoherent proposal. A hypothetical individual could be blind, and deaf, and anosmic (unable to smell), and ageusic (unable to taste), and wholly unable to feel touch, temperature, and the various more subtle senses such as balance and proprioception. Yet he nevertheless would be able to experience his own thoughts; his internal narration would be his to receive. To propose that he would be deaf to this too is essentially to propose brain-death. A person who sensed nothing externally and had no access to his

[j] For just a few examples of the myriad ways we can deflate or reduce away the subjectivity of consciousness, see variously Kant [1787], Hume [1739], Searle [1992], and Dennett [1993].

[k] Typical sufferers of locked-in syndrome retain some function, most often of the ocular muscles. However, the most severe reported cases do demonstrate total loss of all voluntary movement. There is some current work being done with direct brain interface using electrodes that may undermine this example, though.

[l] The interruption of experience involved in sleeping, or even deep coma, is not a counter-example to this model. Even if we propose dreamlessness, when I sleep I will eventually awake, at which point my subjectivity resumes. There is no requirement here for an *uninterrupted* stream of consciousness. And if I never will awake, then it does seem reasonable to treat me as though I were dead.

thoughts would, in a real sense, be no longer conscious, even if he could still actually perform actions. If he moved, or formed thoughts, he would not know he had done so. He would be a true philosophical zombie; it would not be like anything to be him.

One might argue, "Well, the locked-in patient may be unable to act externally, but he still does have agency; he can still act *internally,* or form decisions, even if he is prevented from executing them. If he could not do this, we would not consider him properly conscious." This is valid, and to weigh it we must consider the case of a person with experience, but no agency. The locked-in patient has no bodily agency, but how can we do away with his mental agency? This would be to have the reins of your thinking hijacked by an external force; you could bear witness, but someone else would be initiating your thoughts.

There is probably no sensible real-world example of this, but we can imagine a future example. Scientists insert electrodes within your brain, allowing them to instantiate mental events without your volition. They type on a keyboard, "*These are my thoughts,*" and you hear yourself think those words; they click an icon of an orange cat and into your mind springs the image of a furry pet.[m] You are mentally paralyzed; you can not think of cats on your own, and indeed could not even form the intention to do so, but you do still experience some thoughts. Are you nevertheless conscious? It seems so. The very idea of bearing witness to these mental phenomena is to say that there is *someone* bearing that witness; the scientists would otherwise be playing for an empty theater.

So we have identified the exact virtue upon which both of our future dilemmas hinge: it is the capacity for subjectivity. In what circumstances should an AI merit "personhood" status, with all the ethical trappings thereof, including a right to life? Only if it possessed its own subjectivity. And in what circumstances should my own mind, uploaded into a new substrate, be considered "me"? Only if the newly-manifested mind demonstrated subjectivity—specifically, my *own* subjectivity, that same viewpoint I originally manifested prior to the procedure.

4. Epistemic Criteria

We must next ask what sort of evidence we should require in order to determine the presence, or absence, of subjectivity. How are we to know if an AI possesses subjectivity? Or after uploading my mind, how are we to know if Metoo is continuing my own subjectivity?

The most obvious method is to ask them. Recalling our AI Cybo, why can we not simply inquire of him whether he finds himself to be an experiential center, a locus for his various inputs—so long as he has adequate faculties to communicate in this sort of way? One wonders whether the question would even make sense if Cybo's answer were in the negative. Could a being with no consciousness have any grasp of what it means to

[m] This thought experiment, and certainly my language in depicting it, is rather facile; I do not mean to rouse ire by painting such an outrageously Cartesian picture of the mental space. It should be considered as an illustrative example only, meant to show the intuitive nature of our idea of consciousness, rather than as any sort of serious possibility.

be conscious? Perhaps; this may be consciousness-chauvinism; it may be quite possible for an AI to have enough intelligence to understand such a question while still avowing that it lacks that quality. "This system does have a variety of inputs, and those inputs are used as important factors in the determination of appropriate outputs using internal guidelines," Cybo might tell us, "but there is no overall unity to this process, or a feeling that it is all being experienced from a single vantage-point." [n]

Could Cybo be mistaken? It is perhaps tempting to believe him if he tells us that he has no subjectivity; but if he tells us that he *does,* what then? Do we have any reason to disbelieve him? One good reason is that if he has any desire for self-preservation, he might be cognizant of the fact that answering "no" is a good way to abandon his right to life, and possibly get himself scrapped when a newer model comes around. This decision need not be made in the interest of preserving his consciousness; even the dumbest machines can have basic mechanisms for survival. A car does not flash the Check Engine light because it fears for its safety.

What about my uploaded duplicate, Metoo? We can ask him the same question about his consciousness, but it seems inevitable that he will answer in the affirmative. If I judge myself conscious now, he should judge himself equally conscious. What if we ask him whether he considers himself *me*? Again, there seems little doubt that he will affirm. I consider myself me, and he has everything that I have, mental and physical. He may not even know that he is the duplicate rather than the original unless he is told, and it seems unlikely that this knowledge will appreciably change his sense of identity.

We must remember our purpose here. Our goal is to establish normative guidelines for the treatment of Cybo and Metoo, practical guard-rails preventing us from accidentally violating our own moral values. We are not attempting to untangle the underlying metaphysical questions of consciousness. Two thousand years of philosophy have been unable to objectively demonstrate that even our fellow humans are conscious; it is futile to think that we will discover a method now. Rather, we treat our fellow humans as conscious as a matter of pragmatic belief based on their outward behavior, and the same procedure ought to apply here. For these loose requirements, we need only ask what manner of error would cause the most harm.

For Cybo, the dilemma is easier. There is less to be lost in erring on the side of personhood. If an AI is developed that claims subjectivity, then it probably ought to be respected with a right to life. We may find ourselves wrongly treating a certain number of "zombie AIs" as persons under this policy, but the only loss is that we will no longer be able to use and discard them like dumb tools. Indeed, for such purposes, we might even go out of our way to develop AI that either lacks sufficient complexity or lacks certain traits needed to support subjective consciousness—a cautious effort to segregate our AIs into persons and non-persons.

For uploaded minds, things are not so simple. The case of Metoo seems relatively easy: although we will have to find ways to (for instance) properly allocate my property

[n] Dennett has described a reductionist view of the self as a "center of narrative gravity" useful for organizing the various functions of the brain. One wonders whether or not an intelligent entity like Cybo could go about his many functions, including survival, *without* ever needing such an organizing construct. It seems unlikely, but is certainly not impossible.

between him and I, we each have our own subjectivity, and there is no conflict so long as we are not forced to fight for sole custody of my identity. But there are many cases where the decisions are murkier.

Recall the version of uploading wherein I scan myself in Boston, reinstantiate myself in Paris, and destroy my original body. There is no overt conflict here; only one person remains, and it is a person who asserts my identity. But does the lack of conflict make any difference? If my subjectivity differs from Metoo's in the case where I (the original) am not destroyed, surely it differs just as much in a case where I *am* destroyed. The thing I seek to preserve, my subjective experience, will have been lost. The only difference is that I will no longer be here to assert my claim; that is, the person whose rights have been abused will be unavailable to render objections.

It is important to understand the implications of this possibility. In such a scenario, we will be left with a single person, who claims that he is me, has all external evidence of being me, and probably has all internal evidence of being me as well. In a sense, he *is* me, in his eyes and in the eyes of third parties. There is only a single perspective by which he is not me, and that is my own, that of the original person. And since I will not be here to complain, there is *nobody* who will complain; by all appearances, I will have simply traveled to Paris. Yet if we accept our intuitive understanding of these events, the fact remains that the person who started in Boston did not survive the journey. We are therefore faced with the possibility of an undesired outcome—the destruction of my person—that will, after taking place, leave no evidence that it has occurred. It would quite literally be a perfect crime.

There are some commentators, chief among them the verificationists, who would be inclined to say that if we cannot detect this loss, no loss has occurred. But I must beg for an exception to their principles in this one case. The harm we are discussing is a negative harm; it is the loss of a subjective person, just the same as if he were killed in a plane crash. By placing the complainant beyond our access, such lethal events by their very nature tend to hide the evidence. We acknowledge death as undesirable now, at least to the extent that we do not wish to die, even though it is clear that after the event, the interests of the dead will no longer be relevant. If we grant that it is unacceptable to kill a man—without his consent—even if nobody would note his passing, then we must agree that it is equally unacceptable for him to be uploaded to Paris. Even if he *does* consent, then it would at least be tantamount to suicide.

In short, under at least one plausible interpretation of uploading, there is a real possibility for unevidenced harm. And if this is the interpretation we eventually accept, then we must be willing to assert the danger of uploading, *despite* the lack of any voices to lead the victims. It is a quiet crime; we should have to adopt this position on the strength of belief in the logical, intuitive, and philosophical reasoning that led us there, rather than on any subjective or objective evidence.

One more wrinkle warrants discussion, and that is the case of gradual change. A Metoo duplicate is physically and mentally just like me, yet seems not to *be* me, largely because he is manifested as a physically distinct entity. Yet consider this: the physical particles that make up my body are not permanent. Every cell and atom in me is in a constant process of destruction and renewal; even the molecules of my brain itself will all

"turn over" within the course of a month's time.[o] So me-next-month will have the same blueprint as current-me, but will be composed of entirely different physical particles. Is me-next-month therefore a Metoo? The only difference is that the changeover from me to me-next-month occurs gradually, while the changeover from me to Metoo occurs instantly. Is this another example of the perfect crime? If so, it is one that is committed continually throughout the course of human history, and we seem to accept it.[p] It is far from clear how best to understand these opposing examples.

5. The Road Ahead

Where does this leave us in our search for normative "guard-rails" in the coming Singularity?

There is a certain simplicity yielded in distilling our concern to the focused matter of preservation of subjectivity. Some problems seem to give up relatively easy answers when viewed in this manner. The development of AI, for example, does not appear to present an impossible ethical dilemma, even if it continues to remain an open metaphysical matter.

Although some examples of mind uploading appear answerable in this way—mainly the non-destructive cases—there is still a great deal of ground that may only be approachable using traditional philosophical methods. We still do not know for certain whether uploading my mind is a sensible method of travel, or of "backing up," and certainly it is unclear what we could expect from a transformation to a purely digital existence.[q] And these questions seem to hinge entirely on the details of how we understand the identity of persons. Good work has been done on this front, but it will have to be developed further in order to answer our concerns.[r]

At the moment, let us simply state the following:

(i) The essential quality of human consciousness is personal subjectivity.

---

[o] Some commentators invoking this example seem to feel that there are biological elements of the brain that remain relatively permanent, but Kurzweil cites John McCrone [2004] in denying this.

[p] This general case of natural turnover vs. uploading is directly from Kurzweil [2005]. He finds it particularly germane because in his predictions, gradual artificial augmentation of our biological selves is the most probable means by which our transformation to artificial substrates will begin.

[q] Kurzweil does propose a shift to wholly "soft" consciousness, without any fixed physical substrate. This complicates the question of identity still further, as the exact locus of a specific person becomes harder and harder to isolate. An anonymous reviewer of this piece noted the logical next step: the recent trend towards decentralized or "cloud" type network storage might mean that a given consciousness could be distributed widely across the internet of tomorrow. One wonders if at some point we would simply have to accept that the nature of such persons would be wholly different from those we currently know, and their subjectivity too alien to understand.

[r] Parfit [1985], section 3 gives a particularly useful round-up of these questions of identity, as he examines a number of directly applicable cases including the teletransporter and various important examples of fission. Although Parfit is arguing for a particular resolution to the dilemma that may not be widely accepted, his overall analysis of the different options and concerns within identity theory is relatively non-partisan.

(ii) AIs that demonstrate subjectivity should be viewed as being like humans, at least as far as a right to life; and other rights may also obtain.

(iii) Methods of uploading human minds that successfully transfer a person's existing subjectivity should be viewed as transferring the original person. Methods that do not should be viewed as the creation of a new person.

(iv) How to determine transfer of subjectivity remains an open question. However, it should be approached cautiously, because it is possible that such transfer will constitute a moral harm that will not be empirically obvious.

Further exploration of this topic should be a priority for those who consider an impending Singularity as a serious possibility. To countenance this possibility is to invite these vexing philosophical questions into the realm of immediate and essential political, legal, and ethical dilemmas.

6. References

Block, N. [1995] On a Confusion About a Function of Consciousness, *Behavioral and Brain Sciences* **18**, pp. 227–247.

Chalmers, D.J. [1996] *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, New York and Oxford).

Dennett, D.C. [1993] *Consciousness explained* (Penguin Books, London), p. 398.

——— [1978] *Brainstorms* (Bradford Books, Montgomery, Vermont), pp. 310–323.

Hume, D. [1739/1888] *A Treatise of Human Nature,* ed. L Selby-Bigge (Oxford University Press, Oxford).

Jackson, F. [1982] Epiphenomenal Qualia, *Philosophical Quarterly* **32**, pp. 127–136.

Kant, I. [1787/1929] *Critique of Pure Reason,* translated by N. Kemp Smith (MacMillan, New York).

Kurzweil, R. [2005] *The Singularity is near: When humans transcend biology* (Viking Press, New York), pp. 383, 377, 189–199, 380, 377, 378, 383–384, 299–317, 320–326, 383–385, 324–325.

McCrone, J. [2004] How Do You Persist When Your Molecules Don't?, *Science and Consciousness Review,* **1.1**. Available via Wayback Machine. http://web.archive.org/web/20050901012905/http://www.sci-con.org/articles/20040601.html Accessed 26 Aug 2010.

Nagel, T. [1974] What is it Like to Be a Bat? *Philosophical Review* **83**, pp. 435–450.

Parfit, D. [1985] *Reasons and Persons* (Oxford University Press, New York and Oxford), pp. 199–201, 199–347.

Searle, J. [1980] Minds, Brains and Programs, *Behavioral and Brain Sciences* **3**, pp. 417–57.

——— [1992] *The Rediscovery of the Mind* (MIT Press, Cambridge).

Turing, A. [1950] Computing Machinery and Intelligence, *Mind* **59**, pp. 433–460.