Killed by bad philosophy

Why brain preservation followed by mind uploading is a cure for death

Kenneth Hayworth
(January 2010)

A tragic miscalculation:

You wake up with a splitting headache and weakness on one side of your body. Your wife rushes you to the hospital where a CT scan reveals what is wrong. You have a massive aneurysm that is about to burst in your brain and will surely kill you when it does. Furthermore, the position of the aneurysm is deep within the brain making a traditional surgical approach impossible. You are given some drugs to ease the pain and the doctor explains your options to you and your wife.

The doctor suggests the use of an untested surgical procedure that may give the surgeons enough time to clip the aneurysm and save your life. The procedure will involve lowering your body and brain temperature down to 10^{o} Celsius and then stopping your heart and blood flow for a full hour. This stopping of blood flow will allow the surgeons to complete the complicated surgery without the risk of catastrophic blood loss, and the low temperature will protect your brain from metabolic damage during the hour it will be without oxygenated blood flow.

Perplexed, you ask your doctor "Will my brain still be active during the surgery?" "No", your doctor says, "At 10^{o} Celsius all communications between neurons is halted. In fact this will be one of the tests we will use to make sure we have your brain's temperature low enough to begin the procedure." Incredulous you ask, "Then you are saying I will be *dead* for a full hour, and then you will attempt to bring me back to life?!?" The doctor attempts to reassure you, "Well technically you will meet most of the legal requirements of death during that hour, but our research on animals suggests that once we rewarm your brain and restart your heart you will simply 'reboot'. You should wake up just like you would from anesthesia following a normal surgery."

The doctor leaves the room to let you and your wife discuss. She is religious and says she believes that your soul will be called up to heaven as soon as your brain stops functioning. Being an agnostic with a scientific educational background, you are offended by such an obviously illogical juxtaposition between the metaphysical and biological. At a different time you would have sarcastically responded to your wife by asking her "Exactly which passage in the bible makes this connection between an immaterial soul and patterned neural firing in the brain?" But at this moment, with your life at stake, you are only deeply saddened that you cannot rely on your wife's counsel. However, your own scientific background can do little better. You recall a spate of popular science articles you have read about the functioning of

the brain, all ending with vague statements about the remaining mystery of human consciousness. You also recall several philosophical articles suggesting that our conscious self is an emergent phenomenon of the complex neural activity within the brain, possibly having something to do with its quantum state or some such thing. These all suggest to you that being brain dead for a full hour is not reversible. You start to imagine your doctor happily presenting you as a drooling Frankenstein-like zombie to your wife after the surgery and egotistically declaring that the surgery was a complete success.

In the end you decide not to undergo the surgery, reasoning that it has essentially zero chance of success but would cost your grieving family several thousand dollars even after insurance. You comfort yourself that perhaps the aneurysm will not burst and in time may disappear. A week later the aneurysm does burst and you die of a massive cerebral hemorrhage. Your wife and young son mourn the loss of the father he will never know.

A few years later your wife reads a news paper article about how a new surgical procedure is proving phenomenally successful at treating previously intractable aneurysms and how a dozen patients are now living normal lives after undergoing the procedure. With a sick feeling she realizes that the doctor described in the article is the same one that had suggested the surgical procedure for her husband. She is sad and disturbed but rationalizes by thinking "it was God's decision". However many years later when relating the story to her, now adult, son, he has a different interpretation of this chain of events: "So what you're telling me mom is that dad was not killed by a burst aneurysm. He was actually killed by bad philosophy!"

Profound Hypothermia and Circulatory Arrest:

How realistic is the above scenario? Very. The procedure describe above is called Profound Hypothermia and Circulatory Arrest (PHCA) and it is a real surgical technique used for treating, amongst other things, hard to reach brain aneurysms (Sullivan, Sekhar, Duong, Mergner & Alyano, 1999). It has been in limited use since the late 1950's and it does indeed lower the temperature of the brain to such a degree that all communication and patterned activity between neurons is halted (Lomber, Payne & Horel, 1999) for up to an hour. The only part of this scenario that is unrealistic is the doctor letting his patient commit suicide over such a flimsy philosophical argument. A doctor today would simply point to the hundreds of reports of patients leading high-functioning lives after undergoing the procedure. I do not know if the first patients to undergo a PHCA procedure had such philosophical misgivings.

I hope you agree with me that it would be particularly pathetic for a son to realize that his father died because of his misplaced faith in "bad philosophy". In retrospect, i.e. now that we know that people do survive the PHCA procedure, we can see how foolish it would be for the man in the story above to simply assume that stopping brain activity would result in his irreversible death. If the man had consulted a neuroscientist the sarcastic response obtained might have been "Exactly which passage in your [philosophical] bible makes this connection between an immaterial soul and patterned neural firing in the brain?"

I am belaboring this point because I am virtually certain that our grandchildren will be saying the same thing about us. They will say that we died not because of heart disease, cancer, or stroke, but instead that we died pathetically out of ignorance and superstition. They will say we were killed by our "bad philosophy". In one hundred years they will ask in disbelief, "Our grandparents had the technology to preserve the precise neural circuitry of their brains for long-term storage. The best science of our grandparent's era stated unequivocally that this unique patterning of neural circuitry was the seat of the self; in it was written all memories, skills, and personality. Our grandparents seemed to grasp the quickening pace of technology, and understood that full brain scanning and simulation was around the corner. Why then did grandpa and the rest of his generation reject brain preservation and mind uploading as a means of overcoming death?" And, after considering the evidence, our grandchildren will come to the sad conclusion that we were killed by our "bad philosophy" – no matter how clear the science was, we simply could not really accept the fact that we were physical machines.

Uploading procedure circa 2110:

Let's speculatively jump forward 100 years to the year 2110 and attempt to perceive the history of our generation through our grandchildren's eyes. In 2110 it will likely be commonplace for people to upload their minds into computers and to replace their biological bodies with far superior robotic and virtual ones. A likely scenario will be as follows: A woman 100 years from now walks into a hospital to undergo a mind uploading procedure. She is given general anesthesia and wheeled into a surgical suite where an open-heart bypass surgery is performed hooking her vascular system up to a set of external pumps. These pumps first perfuse a chemical fixative, glutaraldehyde, through her vasculature so that it quickly reaches every cell in her body and in particular every neuron in her brain. This perfusion of a poisonous chemical fixative is done to crosslink the protein machinery within every cell, fixing these proteins in place and preventing decay. Next, another poisonous chemical fixative is perfused, this time osmium tetroxide, which fixes lipid molecules in place. These two steps (fixation of proteins and fixation of lipids) are crucial in that they use chemical bonding to in effect "glue" all the molecular machinery within her cells together.

Next our patient's brain and spinal cord is prepared to allow for nano-resolution scanning. The first step in this process is to perfuse her vasculature with heavy metal staining solutions, like uranyl acetate, that bind to cell membranes allowing them to be later visualized under an electron microscope. Next all water must be removed from within and between the cells and replaced with a plastic resin that can be hardened to rigidify the tissue allowing it to be cut into smaller pieces and imaged. This is done by perfusing her first with ethanol to gradually leach out all water, then with an organic solvent to leach out the ethanol, and finally with increasing concentrations of plastic resin dissolved in the organic solvent until every region of intracellular and extracellular space is filled with pure plastic resin. To assist in this process of removing and replacing water with plastic resin, several holes are drilled in our patient's skull and tubes inserted through these to reach into the ventricular and subdural spaces of her brain. Normally, our brain and spinal cord float in a liquid called cerebrospinal fluid in a "sack" of tough material called the dura mater. The inserted tubes allow the ethanol, solvent, and plastic resin to be

directly circulated within this dura mater sack while these solutions are simultaneously being perfused through the patient's blood vessels. At the end of this process the patient's brain and spinal cord are floating in pure plastic resin and every nook and cranny of intracellular and extracellular space is also filled with this plastic resin.

At this point our intrepid patient is wheeled into a 60^{o} Celsius oven which hardens the plastic resin in her brain and spinal cord into a solid block. Skin, muscle, vertebra, and skull bones are removed revealing the dura mater sack. This tough material is then peeled back to reveal a perfectly preserved brain and spinal cord (including initial segments of cranial and spinal nerves) encased in an amber-colored transparent plastic sheath. Every neuron, every synapse, every delicate neuronal process in the woman's central nervous system is now perfectly preserved down to the nanometer level – the most perfectly preserved fossil imaginable. Key molecular components like ion channel and receptor proteins are also preserved - chemically fixed in position by the glutaraldehyde crosslinks and embedded in the plastic matrix.

The patient's plastic embedded brain and spinal cord is then taken to an automated sectioning machine which uses incredibly sharp diamond knives to slice her brain into long strips just 100 microns thick, and these strips are collected on long spools of tape. A slight heating of the diamond knives is used to soften the plastic during this thick sectioning procedure, in this way no material is lost or damaged during this thick sectioning. The thousands of spools of tape, each containing many brain strips, are then loaded in parallel into thousands of electron microscope scanning machines. Each machine scans the surface of a strip with thousands of parallel electron beams, each just five nanometers wide, quickly producing a high-resolution two dimensional picture. Then a focused ion beam is used to ablate away the just-scanned top 5 nanometers of the strip, and the newly revealed surface is again imaged with the scanning electron beams. It takes 20,000 repeats of this ablation and imaging process to go through the entire 100 micron thick strip, but in the end a full three dimensional map of the tissue is produced with 5x5x5nm resolution. Once all the brain strips have been imaged via this destructive process our patient's original physical brain and spinal cord will have been completely destroyed, but a 5x5x5nm resolution volume image of her entire brain and spinal cord will have been acquired in the process.

This volume image is then processed in a computer to map out the connectivity of every neuron in her original brain, and to estimate the strength and type of each synaptic connection. This map of neuronal connectivity is then interpreted in light of the last 100 years of neuroscience knowledge (i.e. the years 2010 to 2110) to yield a map of the functional properties of all neurons and connections suitable for simulation in a computer. Recall that the plastic preservation procedure preserved the cranial and spinal nerves that connected the original biological brain with the body's sense receptors and muscles. Using these simulated nerve roots, the computer simulation is now "hooked up" to a robot body –tying signals to and from the robot's actuators and sensors into the simulated brain's cranial and spinal nerve roots.

The initial states of activity of neurons in her simulated brain are set to approximate waking up from a full shutdown surgical procedure like the PHCA procedure described above. Then the simulation

is started up. Our patient groggily opens her new robotic eyes and cautiously stretches her new robotic arms into the air. It will take many weeks of practice for her to get fully used to controlling this robotic body and some tweaking of the simulation parameters before she is satisfied with her new sensory receptors, but she already knows the operation was a success as she practices recalling memories of her previous biological life. Once fully checked out, she heads home in her new robotic body to reunite with friends and family members many of which have also undergone a similar procedure.

Inevitability:

By the year 2110 such mind uploading will probably be as common place as laser eye surgery is today. No one will be seriously bothered by the philosophical questions that mind uploading provokes today. No one will ask "Sure it will have my memories, it will act like me, and it will even think it is me, but will it really be me?" Once the procedure has been performed a few times this question will be as silly as us asking today if a person having undergone a PHCA procedure is still the same person, or for that matter if a person who receives a heart transplant is really the same person.

The steps in the speculative uploading procedure above are all straightforward extrapolations from current technology. The perfusion of chemicals to preserve and prepare brain tissue for electron microscopic imaging is straight from textbook protocols (Hayat, 2000) except that these are usually not applied to whole brains. The scanning technology exists today to image brain tissue at 5x5x5nm resolution (Knott, Marchman, Wall & Lich, 2008) although currently far too slowly to image a whole brain in a reasonable time. Large scale simulations of brain circuits are becoming increasingly sophisticated (Markram, 2006) and humanoid robotic bodies are reaching higher and higher levels of dexterity all the time (e.g. the Honda Asimo robot). The writing is on the wall – not only do current cognitive and neuroscience models of the brain support the feasibility of mind uploading, much of the technology necessary for mind uploading is already in existence in primitive form today. From a historical perspective, mind uploading is just around the corner and with it a cure for all aging and disease. More importantly, the initial step of mind uploading – a surgical procedure to preserve the brain and spinal cord in plastic – could be demonstrated today given relative small amounts of monetary and intellectual resources. Since a brain that is chemically fixed and embedded in plastic will not change for thousands of years, this means that anyone today that wanted to could be preserved so as to be uploaded when the requisite technology arrives. Their plastic-embedded brain, and the brains of their friends and loved ones, could simply sit on the shelf for a hundred years to be reawoken and reunited perhaps as soon as 2100.

Rejection of reason:

But despite the scientific obviousness that a cure for death is at hand and despite the fact that the main argument for this cure was published over 20 years ago (Olson, 1988), the vast majority of people, scientists and laypeople alike, reject the idea of mind uploading out of hand. Perhaps even you,

gentle reader, find yourself vigorously rejecting these arguments. Why? There are many reasons. Most people have sadly been brainwashed by the world's major religious cults (Christianity, Islam, Judaism, etc.) to reject the very fundamentals of scientific materialism that have produced modern society. Others simply lack up to date knowledge in cognitive and neuroscience research, or lack understanding of technological developments in brain preservation, electron microscopic imaging, neural simulation, and artificial intelligence. These reasons can certainly explain why most laypeople are oblivious to the potential promise of brain preservation, however this still leaves millions of scientifically well educated free-thinking people that have also rejected the possibility – and their rejection is what has stalled research into brain preservation and prevented it from becoming available as a legal medical procedure.

It is notoriously difficult to get people to clearly articulate the reasoning behind their rejection of mind uploading – it is often stated as simply an intuition that it will not work. However, it is important to clearly articulate the reasoning behind this intuition so that it can be evaluated in light of the available scientific facts. After all, the history of science and technology is filled with overturned intuitions. To this end, I will attempt to clearly articulate the main philosophical intuition people express for rejecting mind uploading, and then show why this intuition is wrong.

From my many discussions with scientifically well educated, free-thinking individuals who nonetheless reject mind uploading, I have perceived a common theme underlying their rejection. They, like the man with the aneurysm in the story above, reject mind uploading as a personal cure for death based on a flawed philosophical intuition about the nature of the self. This flawed intuition is that there is something fundamentally unique about our instantaneous (moment to moment) conscious self that sets us apart from all other people. They associate their "true self" not with their unique set of memories, skills, and proclivities which they have built up over their lifetime of experience, but instead with this vaguely-defined instantaneous self which in fact does not exist at all. From this fundamental contradiction flows a host of incorrect inferences such as the idea that a self cannot be copied from its biological substrate into a computer, and that a self cannot exist as multiple copies simultaneously. They assume that this unique instantaneous quality of self will be lost in the uploading procedure, being replaced with a new self that is not their own.

I will attempt to show why this unique instantaneous self does not really exist, but I will also show why we believe so much that it does – in fact, it is a useful cognitive illusion built into us by natural selection similar to the many useful perceptual illusions that natural selection has endowed us with. Grounded with this better understanding of the self we can then understand why an uploaded mind will provide us with the same instantaneous sense of a unique self that we have now – that is, we will understand why an uploaded copy of us will be *us*.

What is the self?

Our intuition tells us that being me (Ken) right now staring at these words on my laptop screen is fundamentally different from being another person, say my friend John, staring at these words on his laptop screen. Of course there is truth to this, John and I will understand these words in a somewhat

different way and will react somewhat differently to them. But our intuition also tells us that being Ken right now staring at these words is somehow fundamentally similar to being Ken driving in his car to work. There is a "being Ken" quale (singular of qualia) that is similar even in these two very different experiences (reading and driving) that is utterly missing in John's conscious experience (and is replaced with the "being John" quale).

I submit that this intuition is wrong, and that it is fundamentally incompatible with our computational view of the brain's functioning. I will try to demonstrate this in the paragraphs that follow. I take the time to do this because I believe that this is the central incorrect intuition that causes most people to reject the idea of mind uploading out of hand and that leads others to demand unrealistically complicated "uploading" procedures (e.g. Cryonics – must be the same physical molecules, Gradual Transfer – awareness must be unbroken throughout the transfer procedure, Almost Exact Coping – e.g. molecular resolution simulation, and even Quantum Teleportation – must have the same quantum state). If we can see past this incorrect intuition we will be able to better evaluate exactly what is necessary to preserve and upload someone – and we will see that the technologies necessary (at least to preserve our brain with sufficient fidelity for a successful upload) are within our grasp today.

The debate over mind uploading revolves around a central question, "What do you consider to be you?" Mind uploading is useless if this personal definition of "you" is not successfully transferred. I would like to argue that this "you" is not, and should not be confused with, our moment-to-moment ongoing sense of self. It should instead be associated with our unique set of declarative memories (semantic and episodic), our unique set of procedural "memories" (memories for how to react both physically and mentally under particular circumstances), and our somewhat unique set of perceptual "memories" (circuits for how, for example, we recognize that this particular arrangement of spots on our retina is a straight line). Note that I am using the term "memory" for these in the technical sense of (Fuster, 1995), as those changes to the cortical circuitry that we have learned over a lifetime and that guide cognition and behavior. Our particular set of these memories is unique because we have built them up over the course of our lives. These declarative, procedural, and perceptual memories are stored mostly in the unique synaptic connectivity of the cerebral cortex and basal ganglia (Anderson, 2004).

For expositional purposes, let's call the self that we experience on a moment-by-moment basis our "point of view" self (POVself), and let's call the self that comprises our set of declarative, procedural, and perceptual memories our "memory" self (MEMself). Of course, to make a whole person you need both – the POVself experiencing the world on a moment-by-moment basis, and the MEMself guiding our internal and external actions based in part on the current contents of the POVself and in turn modifying the state of this POVself. You can roughly think of the POVself as the currently active representational state of the brain, and the MEMself as the hardwired circuitry of the brain. However, it is important to think of the POVself and MEMself in information terms only. It does not matter which particular neurons are firing to represent that a bus just pulled out in front of you or which particular axonal connections cause you to swerve the steering wheel – what matters is the informational content of these neural representations and circuits. In this way a comparison of Ken's, John's, and simulated Ken's POVselves

and MEMselves can be made even when there is no one-to-one mapping between neurons in different brains.

I argue that there is in fact surprisingly little that is unique (from person to person) about our POVself – its informational content is tiny compared to the MEMself, the POVself varies all over the map within a single individual, and crucially this varying *overlaps* with all other humans' moment-to-moment POVself experiences. For example, Ken's immediate conscious experience of feeling happy is much more similar to John's immediate conscious experience of feeling happy than it is to Ken's immediate conscious experience of feeling sad. If Ken stands in the northwest corner of a room looking southeast, the immediate conscious experience of this point-of-view is almost identical to that which would occur if John were to stand in the same corner looking in the same direction. Of course there are differences in what would attract our next visual saccade. Ken might saccade to the sexy brunette to the right whereas John might saccade to the sexy blonde to the left. Since both girls were in our peripheral vision, our disparate saccades could not be tied to a difference in our POVselves (we both experienced a brown blur to the right and a yellow blur to the left) but instead to a difference in our perceptual and procedural memories. If the blonde approached, John's POVself would experience her as more attractive, but again this should really be attributed to a difference in our perceptual memories. I doubt that the quale (the conscious feeling) of attractiveness is truly different between Ken and John, it is just a difference between what external stimuli will give rise to the activation of our "attractiveness neurons" and what further neural and behavioral consequences an activation of those attractiveness neurons will engender.

If the brunette were to start talking to Ken in Spanish, he would hear only a string of gibberish since he does not understand Spanish; however, if she were to talk to John, who is fluent in Spanish, he would experience discrete words and meaning. Is this a difference between Ken and John's POVselves? Of course it is, but this difference is only manifested when someone is speaking Spanish to us. To see this we can ask "Would John know if somehow one was to destroy all the connections in his brain that underlie his ability to understand Spanish?" He would know this eventually, but only the next time someone starts to speak to him in Spanish. I would argue that his POVself would not blink to such a destruction unless it was in the middle of a conversation in Spanish.

This is not really a thought experiment since people have strokes in Wernicke's cortical area and lose the ability to understand spoken language. If, while hiking alone in the woods, I had a stroke in my Wernicke's area I may very well not realize that I had lost anything until I got home and my wife started talking to me. This is crucial since it is saying that relying on the continuity of my POVself is a very unreliable way to judge if I have lost something that I think everyone would consider to be a crucial part of themselves - their ability to understand their native language.

Even more convincing are hemineglect patients that have suffered a stroke to a large region of their right parietal cortex. These patients have lost the whole left side of their body and visual world but often do not appreciate this loss. If we were talking to a woman during a brain surgery procedure asking her continually if she feels she has lost anything, and then we accidentally produced such a lesion, it is quite possible that she would respond that she feels as if nothing just happened. Think about this - we could disconnect the entire left side of her phenomenal world (through a lesion in the parietal cortex)

and she would not realize that she has lost anything. This should raise serious questions about the nature of the POVself. Our intuition assumes a kind of omniscience to our POVself (our moment-to-moment consciousness) that somehow incorporates all the various memories and abilities of the MEMself. In truth, the POVself is actually so myopic that it is oblivious to all aspects of our vast MEMself that are not actively engaged at the current moment.

The famous Clive Wearing case displays this minuteness of the POVself even more starkly. He suffered brain damage so extensive that he has lost all his previous episodic memories and has lost the ability to form any new episodic memories. He lives in a perpetual present. His notebook is filled with statements like "It is now 5:05pm, This is the first moment I have been alive." A few minutes later he will cross this out and write "It is now 5:09pm, THIS is REALLY the first moment I have been alive." And this pattern continues for page after page, notebook after notebook. He literally does not judge himself as being conscious (at least not before the present moment). He has lost the sense of continuity that is central to the very concept of a self. Yet at other times, when he is not attempting to access episodic memories, he can achieve a POVself state that appears perfectly normal. For example, if you can get him in the middle of a task (e.g. playing the piano or making small talk) he will act completely normal and seems to not realize that he is missing anything. It seems that during such an engaging task (not involving episodic memory) his POVself is essentially identical to what it was before his brain damage occurred. Think about this. If you, gentle reader, were engaged in playing the piano and someone destroyed all the neural circuitry encoding the episodic memories of your life, then you (just like Clive Wearing) would not know that something was missing until you stopped playing the piano, at which point you would attempt to access your episodic memories, find none, and declare that this is the first moment you have been conscious!

What about such self-defining qualities like temperament? I may get furiously angry when someone cuts me off on the freeway where as John may casually dismiss the event. However, this is not due to a fundamental long-term difference in our POVselves. Temperament is not a fundamental unchanging quality within the POVself - we all have the capacity to get furiously angry, and we all have times when we are at ease. Instead, it is due to a difference in our perceptual and procedural memories activating the "anger neurons" in this particular situation in me but not in John. Theoretically, one could modify the perceptual and procedural memories that activate a person's "anger neurons" thus changing their temperament from calm and laid-back to violent and temperamental. Would the person immediately know the difference? No, not until something happened, e.g. a passing motorist cuts them off, that triggered them to utter a violent outburst of profanity. At such a time they may say to themselves, "Wow, why am I so angry today? I am usually more laid-back about these things." In summary, your temperament is not "always present" in your conscious self; an event is needed that will either trigger anger or not, and this response will then be compared to your store of declarative facts about yourself (e.g. "I am usually laid-back.").

What I am trying to illustrate with these examples is just how miniscule the content of our POVself is and how wildly variable it is. I am certainly not saying that we can exist without our POVself. In such a case there would be no ongoing activity and surely no one home. What I am saying is that we should not be looking for our *uniqueness* in this POVself. Our wide range of POVself states overlaps

almost completely with those of other people except in those few instances that we call up an episodic memory that is unique to our personal history.

The self that we and our loved ones know is a product of our tiny, variable, non-unique POVself being steered down a particular path by the enormous number of unique perceptual, declarative, and procedural memories in our massive MEMself. Our intuition that there is a unique part of our POVself that is not shared by other people and that defines "being me" is simply incompatible with our scientific understanding of the brain as the above examples show.

In the above terminology, the technique of mind uploading seeks to copy the neural circuitry encoding a person's unique MEMself and then simulate that person's brain generating a new POVself. This simulation will "feel from the inside" just like the original because the source of our feeling of continuity of self stems from our ability to call up our unique episodic and self-oriented declarative memories.

The source of our mistaken belief:

Of course this explanation begs the question as to why we feel so strongly that there is a uniqueness to our instantaneous POVself, and why we tend to identify our "true self" with it and not with our truly unique MEMself. I believe the answer to this riddle is that our sense of self is based on a type of cognitive illusion quite similar in nature to well-studied perceptual illusions. Case in point is the well known perceptual illusion involving our peripheral vision. The retina in our eye is designed such that only a tiny region in the center, the fovea, has high acuity. This means that we can only distinguish fine detail in the tiny central region of our visual field, about twice the width of our thumbnail at arm's length. Yet until we are told this fact and test it for ourselves it seems for all the world that we see the entire visual field in great detail. Demonstration of this fact is often accomplished by looking straight ahead while bringing a playing card in your outstretched hand from the periphery toward central vision and seeing how long it takes to identify the card. It is shocking how close to central vision the card must be brought in order to identify it. The incredibly strong intuition that we see our entire visual field in uniform detail is thus merely a perceptual illusion.

What is often not asked is why we have such an incorrect intuition in the first place. The deep answer to this question is that it is built into us by natural selection for a practical purpose. Normally we can access details in any part of our visual field, including the periphery, within 300 milliseconds by simply moving our eyes. Our phenomenal self model (Metzinger, 2003, Metzinger, 2009) encodes this fact (i.e. the fact that we have quick access through eye movements) through the phenomenal experience of clarity of the entire visual field. In contrast, if part of our visual field is truly blurred or obstructed (say by a bad pair of glasses) this is encoded by our phenomenal self model through the experience of blurriness.

Unfortunately the present essay is too short to go into the details of this phenomenal self model (PSM), but this concept is crucial nonetheless. The PSM is the brain's representational content

underlying our sense of conscious awareness (see Metzinger 2009 for a more detailed exposition). The PSM is a continually updated simulation of the self (encoded in the firings of particular neurons in our brain) which describes exactly what it is like to be us. The PSM describes a single unitary self with a particular point of view experiencing perceptual qualia and emotional feelings, having desires and goals, reasoning, making decisions, and taking actions in the world. The brain as a whole uses the contents of the PSM to intelligently guide behavior in an integrated fashion; as such it is one of the best "inventions" that evolution has built into the human animal. The conscious self we experience is the one described in our brain's PSM.

The illusion of clarity across the entire visual field can ultimately be traced back to a simple statement in the PSM that says "The entire visual field looks clear and detailed". This statement is useful to have in the PSM since it tells the rest of the brain that it has quick and easy access to the entire visual field through eye movements. However, we do, and should, find this explanation disturbing. When you stand on top of a mountain looking out at a forest full of hundreds of trees each with thousands of leaves and experience a rush of visual clarity and complexity, where is this feeling of clarity and visual complexity coming from? The implication is that this instantaneous feeling of visual clarity and complexity is not based on the brain encoding the millions of visual details of the forest scene; it is instead based on a very tiny statement in the PSM saying "The entire visual field looks clear and detailed". You may find this fact difficult to swallow, but it has been verified through countless experiments in visual psychology.

Now back to our sense of self. I stated above that our sense of self is based on a type of cognitive illusion quite similar in nature to the perceptual illusion that we see clearly all the way out to our periphery. What I mean by this is that it seems to us that the episodic memories of our life experiences and self-oriented declarative memories like our name, our likes and dislikes, our temperament, etc. are somehow always present in the here and now. However, given what science has discovered about the attention limitations of the brain we know that this cannot be true – only one or a few declarative and episodic memories can be active in the brain at any one time. Like the perceptual illusion described above, this illusory inflation of self may be based on a simple statement in the PSM that says "All the unique episodic and declarative memories that set me apart from other people are present in the here and now".

To summarize, we incorrectly associate our "true self" with our instantaneous POVself based on the intuition that there is a unique "being Me" quale present at all times in our consciousness. Our intuition for this unique "being Me" quale is based on the fact that we can quickly access any part of our episodic memory, and our PSM is built to reflect this fact giving us the illusory feeling of continual access to vast amounts of self knowledge. This illusory feeling *is* our "being Me" quale, but it is no way unique to a particular person. It is a hollow placeholder pointing to our truly unique and vast quantity of *inactive* information that is our MEMself.

Good philosophy and the philosophy of good:

Science gives us only facts not value judgments. It is up to each of us to decide what we should identify as our "true self", and this will in turn guide our opinion of whether mind uploading is personally desirable. I argue that we are left with two alternative philosophies which are compatible with the above facts: 1.) We can identify our "true self" with our truly unique set of millions of perceptual, declarative, and procedural memories comprising our MEMself. In this case, brain preservation and mind uploading represent a real cure for personal death. 2.) We can identify our "true self" with our non-unique POVself, changing from circumstance to circumstance like a leaf in the wind. The core unchanging aspects of this POVself (the fact that there is a conscious point of view) we share with all other human beings. From such a perspective personal death becomes meaningless – we truly are each other, it is just an illusion that we are fundamentally different. This thought, derived from basic neuroscience facts, obviously has parallels in Eastern religious philosophies. For example, the Advaita Vedanta school of Hindu philosophy states that enlightenment is reached when one realizes that one's Atman (self) is in fact identical to all other selves and is Brahman (God). The moral implications of this philosophy should be clear, treat your neighbor as you would yourself, not because God tells you too but because they *really are* you.

There is a third alternative to the above two philosophies - an appreciation that both are correct and it is just personal preference how we weight the two. I personally subscribe to this philosophy because it allows one to appreciate the specialness of every person while simultaneously feeling a strong spiritual bond with all of humanity. We can appreciate that each person is absolutely unique – the product of distinct experiences throughout their life that will insure that the path their thoughts and actions take in any given circumstance will be fundamentally different than another's in that same situation. At the same time this philosophy provides a scientific justification for our fundamental moral intuitions – I know it is wrong to hurt another because I know it is wrong to hurt me.

Implications for mind uploading:

When people discuss what fidelity of copying and simulation would be necessary for a successful mind upload (i.e. one that is still "me") they often don't know where to start. I think this is mainly due to the false intuition I have been discussing above – the intuition that there is a "being Me" quale present at every moment in our first-person conscious experience (no matter what act we are engaged in or what our current feelings are). They insist that this "being ME" quale is different from all other persons' and define their true self as this "being Me" quale – something that doesn't really exist in the first place.

Because of this false and misleading intuition, many people declare the entire mind uploading endeavor impossible. One often stated argument against mind uploading is that it would seem to logically imply that one could also create multiple copies of a person. In fact it does imply that multiple simultaneous copies of an individual are possible. In the above language, these copies would start out

with identical MEMselves but would have different moment-to-moment POVselves because of their experiencing the world through different bodies. Over time different experiences would result in differences between the copies' MEMselves as well. When do the two copies become really separate individuals? There is no real answer to this, it depends entirely on how one weights the two philosophies described above.

One should not underestimate the deeply counterintuitive nature of the argument I am making here. It calls into question the very logic behind our sense of self. To see this consider the following thought experiment: A researcher has just created a perfect copy machine, able to make a molecule for molecule copy of an individual. The researcher wishes to use this device to perform an experiment to see if human beings are fundamentally illogical. John volunteers to be the test subject in the experiment. The researcher first sits John down in a chair and then proceeds to poke John in the arm with a big pin. John screams in pain, pulls his arm away, gets really mad at the researcher, and tells him not to do that again. The researcher explains to John this is part of the experiment that he agreed to. The researcher then asks John why he got mad at being poked in the arm. John simply explains "It hurt."

The researcher then says "John, I am going to poke you in the arm again ten minutes from now. How do you feel about that?" John starts to get really upset and says that he feels anxious and mad that he will be poked in the arm. The researcher asks John why he is feeling anxious and mad about an event that will only happen to his future self. "John, why do you care if I poke your future self in the arm?" John responds, "Well, I feel connected to my future self. That future self is 'me', and if it feels pain then I will feel pain." Ten minutes late John gets poked again.

Next the researcher informs John that he is going to be put through the perfect copying machine and an exact replica of him, John#2 will be created. Being one of the engineers who built this perfect copy machine, John understands its workings inside and out and knows that it will indeed create an exact copy. Before this procedure is performed, the researcher informs John that after the copying procedure is complete that he is going to poke one of the two Johns in the arm again. The researcher asks, "Which John should I poke, John#1 (the original) or John#2?" John thinks hard about this. "Well, I know that there will be no physical difference between myself, err… I mean John #1, and the copy John #2. Sure they will be made out of different atoms but that fact is inconsequential since our atoms are continually being replaced by simple metabolism anyway. I guess the only logical answer is that it does not matter which John you poke after the procedure, although of course I would prefer you not to poke either of us."

John is momentarily rendered unconscious and the copying procedure is performed producing John#2. John#1 and John#2 are awoken and the researcher asks, "So, which of you thinks he is the original?" A back and forth debate ensues between the two Johns, but quickly they come to the same conclusion: "Well doc, we seem to have the same memories, personality, and reasoning skills", says one John. "That's right, we are exactly the same except for trivial things like me being on the right side of the room and him being on the left" says the other John. "So I guess we would both agree that it doesn't really matter which one is the original. Just call us John A and John B."

"Great", says the researcher who then turns to John A. "I am going to poke you in the arm ten minutes from now. How do you feel about this John A?" John A recoils back and says soberly "Well I am afraid that it will hurt me." Turning to John B the researcher asks, "And how do you feel about me poking John A in the arm John B?" Even though John B tries to hide his emotions he is visibly relieved that he will not be getting poked. "Well, I am concerned about John A the way I would be about a close brother. I certainly wish that you would not poke him." John B pauses and then altruistically offers the following "In fact, why don't you just poke me instead."

The researcher smiles, "You two have just demonstrated how humans are fundamentally illogical when it comes to the idea of the self. When I asked the original John before the copying procedure which person I should poke after the procedure, he correctly said that it makes no difference. The physical symmetry of the situation left no other conclusion. There was no physical basis for him to relate more strongly to one vs. the other. However when I just now asked him again this same question he gave a totally different response, even though virtually the same symmetry applied – only trivial differences separate the two of you. There is no physical basis for you, John A, to relate more strongly to John A ten minutes from now than to John B ten minutes from now. Yet predictably you, John A, are viscerally compelled to feel more connected to John A in the future seemingly simply because you share the same atoms with him – a fact that you previously admitted should be considered inconsequential."

We must remember that our instincts and feelings were designed by natural selection without any regard for overall logical consistency. Evolution has built into us a strong instinct for self preservation, and over the entire course of our evolution this "self" has been synonymous with the single physical body we currently occupy – protect its future and you protect your genes. In fact, it is likely that evolution has built into us a set of instincts and intuitions that greatly *exaggerate* the real differences between individuals since those genes that happen to exaggerate such selfish instincts are more likely to help competition and thus be passed on.

We have never faced a situation remotely like mind uploading in which we should logically identify as strongly with a future copy of our self as we should with our future self (although the strong bond of a parent to a child might be an approximation). Many people take their intuition about this situation, i.e. that we seem to not be able to identify as strongly with an exact copy of our self that is sitting across the room as we do with our physical self, as proof that mind uploading will not work. However this logic is flawed. It is assuming that our intuition is correct, and then reasoning from this intuition. This reasoning is faulty in the same way that it would be faulty to reason that the sun revolves around the earth because it "seems that way to us". We must accept that our intuitions about the nature of the self can be as flawed as our intuitions about the nature of the solar system. In fact the last hundred years of cognitive science research has been a history of overturned intuitions about the nature of the self.

Conclusion:

This essay has touched on some very abstract and esoteric philosophical themes, but it is important to remember that this is not merely an academic debate about the nature of the self –it is a real life and death issue. The technology to preserve one's brain is here already so anyone reading these words potentially has the opportunity to experience mind uploading firsthand and see the future hundreds of years from now. However if you, gentle reader, decide for yourself that you wish to have your brain preserved and participate in this "greatest adventure imaginable" you will find, as I have, that your ability to do so is being blocked at every turn.

Funding for brain preservation research is virtually non-existent and only a handful of researchers worldwide are seriously pursuing it. It is illegal to start a brain preservation procedure before legal death has been declared, a requirement which drastically diminishes the quality of preservation. No hospital is setup to perform a brain preservation procedure, and only a handful of unregulated cryonics organizations currently perform such procedures (requiring decades of untrustworthy low-temperature storage). The surgical techniques employed by these cryonics organizations have not been verified in open scientific journals to preserve the structure of neural circuits across a whole brain. Thus even though the technology for quality brain preservation is in principle here today, it is virtually impossible to get one's brain preserved.

I believe this lack of research interest and hostile legal environment can be traced directly back to the "bad philosophy" I have been discussing in this essay. Our faulty intuitions about the self are so deeply embedded into us by evolution that even the majority of scientists are not immune. Our currently accepted cognitive and neuroscience models of the brain unequivocally support the feasibility of mind uploading, and much of the technology necessary for mind uploading is in existence in primitive form today. Why then has the scientific community not embraced mind uploading as a cure for personal death? I believe the answer is that they simply have not carefully considered all of the arguments.

With careful consideration and when presented with overwhelming scientific evidence, we have learned in the past to disregard even our most powerful intuitions. When we look up at the sun moving across the sky we can now override our naïve intuitions and "feel" the earth rotating under our feet. Similarly, when one carefully studies and considers the body of cognitive and neuroscience data amassed over the last century, one can begin to viscerally accept that we are physical machines. We can begin to "feel" our fundamental computational nature, and "feel" the tiny capacity both of our visual field as well as our moment-to-moment self. I, for one, feel as protective of my future uploaded self as I do my future physical self. I look forward to experiencing the world 100 years from now in a robotic body, and I will fight for my right to do so just as I would fight for my right to undergo any surgical procedure that could save my life.

My hope is that many other intelligent, free-thinking individuals will carefully evaluate the evidence that mind uploading is a potential cure for death. If enough people do so then our grandchildren a hundred years from now will not be saying that we were killed by our "bad philosophy", they will instead be welcoming us into their world.

References:

Anderson, J.R. (2004). An Integrated Theory of the Mind. *Psychological Review 111* (4), 1036-1060.
Fuster, J.M. (1995). Memory in the cerebral cortex : an empirical approach to neural networks in the human and nonhuman primate. Cambridge, Mass.: MIT Press.

Hayat, M.A. (2000). Principles and techniques of electron microscopy : biological applications. Cambridge, UK ; New York: Cambridge University Press.

Knott, G., Marchman, H., Wall, D., & Lich, B. (2008). Serial Section Scanning Electron Microscopy of Adult Brain Tissue Using Focused Ion Beam Milling. *The Journal of Neuroscience, 28* (12), 2959 –2964.

Lomber, S.G., Payne, B.R., & Horel, J.A. (1999). The cryoloop: an adaptable reversible cooling deactivation method for behavioral or electrophysiological assessment of neural function. *J Neurosci Methods, 86* (2), 179-194.

Markram, H. (2006). The blue brain project. *Nat Rev Neurosci, 7* (2), 153-160.

Metzinger, T. (2003). Being no one : the self-model theory of subjectivity. Cambridge, Mass.: MIT Press.

Metzinger, T. (2009). The ego tunnel : the science of the mind and the myth of the self. New York: Basic Books.

Olson, C.B. (1988). A possible cure for death. *Med Hypotheses, 26* (1), 77-84.

Sullivan, B.J., Sekhar, L.N., Duong, D.H., Mergner, G., & Alyano, D. (1999). Profound hypothermia and circulatory arrest with skull base approaches for treatment of complex posterior circulation aneurysms. *Acta Neurochir (Wien), 141* (1), 1-11; discussion 11-12.