On the Mystery of the Self and the Selection Problem: A Mathematical Approach

Daniel Caputi

SUNY Stony Brook, UC Davis

Abstract

The self, which presents to us as an irreducible entity in which our subjective experience is directed onto, is often thought of as an unremarkable phenomenon or illusion. However, the mere perception of it cannot be neglected for our scientific endeavor to explain consciousness, and this point is illustrated through a multitude of thought experiments. These thought experiments also show the importance of differentiating selves between distinct conscious organisms, regardless of their individual phenomenological content. A distinction is made between an active subject (a self that is conscious) and a potential subject (a self that is unconscious). Potential subjects refer to selves that would otherwise be present in organisms that are currently unconscious or post-mortem. They can also refer to an infinite amount of imaginary selves that will never be born into existence. This infinite reference space shows that there is an explanatory gap between our knowledge that conscious organisms have selves and our knowledge that specific selves are mapped into specific organisms. This explanatory gap needs to be closed in order to design effective uploading technology to extend the life of our minds beyond the life of our body.

1. Background and Introduction

Despite tremendous progress in cognitive science, there remains a clear explanatory gap between understanding physical processes in the body and understanding how inner subjective states of consciousness (known as qualia) take place. This is the "hard problem" put forth by David Chalmers in 1995[1], which has since held center stage in the field of consciousness studies. While this problem is effective for capturing the mystery of qualia, there is a second piece to the puzzle. It seems that qualia doesn't just happen, but happens *to someone*. In other words, subjective experience as we know it must happen to an *experiencer*, also known as a *conscious entity* or *self*. What exactly is this self, and how does it fit in with consciousness? I will attempt to show that these questions are just as important as the questions often posed about qualia, and this analysis will reveal a surprising second explanatory gap in the nature of consciousness. This explanatory gap does not disappear even if one takes the position that the self does not exist in the way we may intuitively think of it. My hope is that this new perspective will allow more efficient progress toward understanding the full picture of consciousness and the physical place[2] it has in the universe.

This full picture is monumentally important for our future existence. It can be used to determine if we possess a natural kind of immortality, and if not, how we may be able to create it for ourselves. With recent advances in artificial intelligence, ideas have emerged about the possibility of "uploading" one's brain computational parts onto an alternate substrate, such as a computer system or robot[3]. A critical philosophical question has emerged about whether an "uploaded" organism, even if conscious, would have the same identity as the original organism[4]. This paper will explore this question in the context of a new model on the self I will propose. With investigations of consciousness confronting both the problem of qualia and the problem of the self, we may be able to not only cure death for ourselves, but erase it from those who have gone before us.

2. The Problem of the Self

Consider the following thought experiment. The simple objective is to imagine slowly disintegrating your brain. This can be accomplished by either a melting process, or slowly removing neural components or brain cells one piece at a time. If one were to take your brain and remove or demolish just one cell, it would be highly unlikely to have any notable effect on your conscious experience. We know that small quantities of brain cells can be damaged in everyday life, yet we don't seem to feel any different. But surely, if we continue destroying brain matter down to its last bits, a remaining microscopic sample of brain tissue consisting of a tiny collection of cells would not be *you*, right? So, if we slowly cut off more and more usable volume of your brain, at what point are you not yourself anymore?

You may consider a quite simple solution: we do in fact become less of ourselves when a tiny tissue of our brain is removed, but the damage is too miniscule to be recognized. Noticeable changes may begin after large chunks have been removed, and we would officially not be "ourselves" when there are no personality traits remaining. This view considers our "self" as a bunch of content. This type of self comprises memory, all inner sensory experience, personality traits, and really anything else we would consider important for identifying who a person is. In a sense, this is considering a self from an external perspective, but it also includes all the content of inner experience.

The problem with this view is that it only considerers the first piece of the puzzle, the *experience*. At least from an intuitive perspective, phenomenological experience requires *someone* to experience it. Let's define a type-1 self as the *experiential content* of an individual (including memories, sensations,

personality, etc), and let's define a type-2 self as the entity that is *experiencing* the type-1 self. In other words: the type-1 self is the experience, and the type-2 self is the experiencer. The disintegration thought experiment we described above becomes more intriguing when we become less interested in any definition of what constitutes a person's meaningful and extended selfhood (type-1 self), and rather become more concerned with the *experiencer* of the qualia (type-2 self) as the brain is falling apart.

The reason for this is simple. Most likely, as the disintegration process is ongoing, you are gradually losing consciousness. Now imagine yourself very late in the process, with only a small amount of the cerebral cortex intact. At this point, you may be "aware" on a very basic level, but severely lacking any form of organized informational processing. The type-1 self may constitute only perception of basic shapes or colors, devoid of any personality or rich interpretations of these perceptions. But as *you (the experiencer* of this very low level qualia) imagine yourself in this state of mind, it is impossible to eliminate this "you" from your imagined state. There is still a type-2 self anchored in your perception, a "self" entity on which your very low level perception is attached to. V.S. Ramachandran, in his book *A Brief Tour of Human Consciousness*, said "self and qualia are two sides of the same coin. You can't have free-floating sensations or qualia with no one to experience them, and you can't have a self completely devoid of sensory experiences, memories or emotions" (96). The type-2 self does not need to be anything complex, such as reflective awareness that one exists as a self, or an inner language that uses the terms "I" or "me". It only needs to be some central experiencer that qualia is linked to. While the type-1 self can devolve in the brain disintegration process, we can only imagine the perceived type-2 self being *either present or not*. The type-2 self, in the way we perceive it, is irreducible.

What follows from this is a startling implication. If a whole brain possesses this type-2 self, but a very tiny collection of brain cells does not, the logical conclusion is that at some critical and very finite point during the disintegration, the type-2 self (conscious entity) must suddenly disappear. This clearly goes against any intuition of the type-2 self being a larger emergent epiphenomenon of brain activity. Despite its mysterious nature, I would argue that the type-2 self is the *important* part of the self. This is because we may be able to easily upload our type-1 self from our brain into an alternate substrate, and the substrate may have conscious experience, but it would be quite meaningless if it's not *us* that is the experiencer inside.

That being said, these are my basic claims about the nature of type-2, irreducible selves:

A) They exist. Even if they are not physical entities, the "image" of them is real.

B) They are *differentiable*. Numerous conscious entities exist that are different from one another, and since each is irreducible, they do not overlap. In laymen's terms, this is simply saying that you are fundamentally a different conscious subject than I am, even if we are experiencing the same thing at the same time.

(Note: the term "type-2 self" will be used interchangeably with "irreducible self", "conscious entity", "subject", and "experiencer". Different terms tend to fit different contexts, but they all refer to the same thing.)

If you agree with these premises, you can probably skip the next section. But as with anything in philosophy, nothing can go unchallenged.

3. It's Perception that Matters

Talking about this type-2 self is a risky move because many doubt its existence. A trending thought in consciousness studies has been that the self is merely an evolutionary trick of our brains to unify everything that is represented in our minds. The idea of a self has been attacked from scientific, philosophical, and even spiritual grounds. In this section, we will tour some common thought patterns of counter-arguments given to the existence of selves, and show that with each of them, what matters is our mere *perception* of the type-2 self.

3A. Pathology and Altered States

The quote from Ramachandran above is one that many would disagree with. It is often argued that it *is* possible to experience qualia with a completely distorted (or even non-existent) sense of self in abnormal states of consciousness, and this is often used to support the idea that our awareness is fundamentally selfless. Many types of conditions can be explored, including (but not limited to) the following:

Split-brain: An individual with a split corpus callosum, and communication between brain hemispheres cannot occur. This condition is typically induced by a surgical procedure in order to treat violent seizures. Patients often behave as if they have two selves, with each self possessing distinct characteristics.

Schizophrenia: An individual with difficulty distinguishing real and imaginary input. Their consciousness is often depersonalized, with a lack of a sense that *their* qualia belongs to *them*. Agency and unity that binds their experience is also distorted.

Cotard delusion: An individual who claims that they are dead or do not exist. There appears to be selfless consciousness in this case, as affected people often do not use the "I" pronoun to describe anything pertaining to them (Metzinger 63-64).

Even outside of pathology, consciousness without the robust sense of self we experience in ordinary life may not be impossible. Individuals achieving transcendental states of consciousness (through meditation or other means) often report a clear message: the self is an illusion, we are all one and the same. Additionally, many researchers argue that this "self" in our consciousness is only present when we call upon it to be, and it is impossible to catch ourselves not having it. This quote is taken from Susan Blackmore's "The Grand Illusion: Why consciousness exists only when you look for it":

> […] perhaps there is only something there when you ask. Maybe each time you probe, a retrospective story is concocted about what was in the stream of consciousness a moment before, together with a "self" who was apparently experiencing it. Of course there was neither a conscious self nor a stream, but it now seems as though there was.

___

The first question to ask is what these findings in pathology and altered states actually tell us. While many find that these claims are consistent with the existence of selfless consciousness (or a type of consciousness other than one with a single anchored self), others are more skeptical. It may be that the type-2 self is just interpreted and expressed differently in our unique language systems, rather than certain individuals actually experiencing some inconceivable form of consciousness. Ramachandran himself is famous for studying these phenomena, so it is interesting to hear this skeptical position coming from him, noting that "even in the extreme case of a split-brain patient whose two hemispheres have been surgically disconnected, the patient doesn't experience doubling subjectivity, each

hemisphere's 'self' is aware of only itself – although it may intellectually deduce the presence of the other" (105).

But even if it were true that selfless consciousness were possible, it would be a mistake to take these altered state revelations to support the idea that the self is fundamentally illusory just because it can be dissolved under certain conditions. While the disintegration thought experiment is aimed to show that there *exists* a type-2 self that is irreducible, it is separate from the question of whether or not it is *possible* to have experience without an irreducible conscious entity. While the idea of "free floating qualia" without an attached experiencer may seem bizarre, I am not arguing that it is impossible. But the existence of selfless experience does not negate the ordinary sense of self.

3B. Phenomenal Self Model

Thomas Metzinger proposes a phenomenal self model (PSM) in which the content of our consciousness is held and unified. This model, he argues, was a useful adaptation in our evolutionary history because it allowed an organism to interact with both its internal and external environment intelligently. Metzinger describes the rubber hand illusion, in which subjects place one hand behind an optical barrier while a rubber hand is placed in front of it. Both the rubber hand and the actual hand are stroked repeatedly, and after a few minutes, many subjects feel a sense of ownership to the rubber hand. A "whole body analog" was created for this experiment, where subjects had their backs repeatedly stroked as they watched a virtual reality projection of their back (and the stroking) a few feet in front of them. Many subjects reported a feeling that their body was displaced in front of their vision, and the stroking sensation occurring at the location of their virtual back. In Metzinger's view, these experiments demonstrate the ability to manipulate the integrated sense of self in carefully designed experiments. His central claim is that the sense of self feels so real because we are unable to recognize our PSM as a model, as the model itself is transparent.

In my view, the PSM proposal is an excellent attempt at explaining why we feel we have selves, and it may carry truth. However, I would caution against using it to conclude that the irreducible type-2 self doesn't exist, because *this model does not negate our own experience*. Our experience alone is enough to prove the type-2 self as an irreducible entity, much like the fact that the existence of consciousness is proven automatically by our experience of it. Many people are familiar with the optical illusion of the mirage. On hot days, turbulent mixing of air near the ground can create a false image of water on roads. Is the water real? No, but the *image* of it most certainly is. Even if the irreducible type-2 self is just an image and not "real" in the way we may intuitively think of it, that does not negate the ontology of it within our own conscious minds. We can simply define the type-2 self as an image that we perceive without it losing any explanatory power, importance, or mysteriousness. This *sense* or *image* of an irreducible entity is important, because this is what we want to preserve in mind uploading. It is the *sense of irreducibility* that fundamentally produces additional questions about the nature of consciousness that will be discussed in subsequent sections. Even if the type-2 self is only *perceived* to be irreducible, the disintegration thought experiment still works, because at some critical point in the disintegration process, this *perception* of irreducibility must suddenly change. No matter how one looks at it, there is *something* in the nature of consciousness that is irreducible.

3C. Overlapping Qualia

This point is mainly to examine the differentiable property of type-2 selves. Let's examine an essay by Kenneth Hayworth, "Killed by Bad Philosophy". Hayworth writes this essay as the director of the Brain Preservation Foundation in order to make the case that mind uploading will preserve identity. I want to

make it clear that I do not intend to attack his motives to preserve brains. In fact, I believe that his work may be critically important for curing death, as he claims. My only point of this analysis is to show why the type-2 self should not be rejected as, at the very minimum, an important construct. Hayworth states:

> 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).

To paraphrase the authors viewpoint: my intuition tells me that my current experience can be described as the qualia I am experiencing now plus an additional quale, a "further fact"[5], of "being [author's name]". The remainder of the article goes on to argue that there is a Point of View (POV) self that comprises the moment-to-moment experience we have of the world, and a memory (MEM) self that comprises our "set of declarative, procedural, and perceptual memories". The author points out that from a qualitative perspective, there is more similarity between the conscious states of Ken being happy and John being happy than there is between Ken being happy and Ken being sad. An additional point is made about how the POVself would not consciously notice any dramatic sudden change in the brain wiring unless it was actively engaging in a process that involved it at the time. An example is given where if one were to suffer a stroke to the Wernike's (language) area of the brain while hiking in the woods, one may not notice anything has happened until one tried to speak. The author concludes that since the POVself is only used for real-time informational processing and is essentially oblivious to the MEMself, it carries virtually no specific facts about a person's identity on its own. Since our MEMself is what determines our uniqueness as people and truly sets us apart from one another, it is only this MEMself that we really need to be concerned with for preserving identity. The MEMself can be preserved simply by making a functional copy of the brain wiring.

This argument may sound convincing by choice of wording, but this ultimately fails to disprove a fundamental idea: that qualitative states of consciousness *happen to subjects*. Essentially, I do not feel that "being [author's name]" is a quale at all. Rather, I feel that "[author's name]" is an entity on which all of my qualia is being directed. This misclassification is important, because one will not find inherent uniqueness between conscious organisms by only considering experiential content. Again, we cannot rule out the existence of qualia without an attached subject, but I know that I perceive myself as a subject, and I want *this subject* to survive. I, [author's name], am having experiences right now that you, the reader, are not, and we have two distinct subjects (type-2 selves). While Hayworth seems to be considering the POV-self as an analogue to the type-2 self I defined, both POV and MEM self content can be thought of as part of the type-1 self, because they both consider experience (rather than an experiencer). If our POVself content happens to overlap at any given time, it does not mean that our instantaneous selves are not unique. Rather, it would simply mean that the same experience is happening to *two subjects*. This is the simple further fact about our identity that some have gone to great lengths to deny: you and I are *distinct* subjects of consciousness, and this difference exists regardless of the content of our POV or MEM selves (hence the "further fact").

4. Why Differentiable Subjects can Annoy Philosophers

This idea of *differentiable subjects* is understandably disturbing because the boundaries of subjects can get quite messy in philosophical thought world. Personal survival is no longer a matter of opinion in

what one considers to be a person, but an objective fact with a binary yes/no solution, and it is not clear in certain circumstances if survival occurs or not.

The uploading problem is one such example. At first glance, it may seem obvious that your personal identity would survive an upload. If everything that matters about you is the result of the exact structure of your brain, it would make sense that you (as in your type-2 self) survive the upload because your brain structure would essentially be preserved, even when your brain itself is destroyed. But here's where things get dicey. Imagine that instead of directly *replacing* your brain with an uploaded equivalent of your mind on an alternate substrate, we utilize the upload as a *copy* of your mind *while preserving your original body*. The process of uploading is exactly the same otherwise, except that from your perspective inside your original body, nothing should have happened because the scanning and copying is non-invasive. So intuitively, whether or not your type-2 self is transferred into the alternate substrate depends on whether or not your original body is preserved, but yet this seems logically absurd because nature should not care about this detail. Since intuition leads us to believe several conflicting ideas, it may be appealing to believe that the idea of differentiable subjects is somehow flawed.

Let's consider a similar but slightly different scenario. Imagine that you are about to undergo some type of surgery to modify (but not necessarily damage) brain structure. The surgery will require general anesthesia, rendering complete unconsciousness. Thomas Clark, director of the Center for Naturalism, considers what would happen during this kind of surgery in his essay "Death, Nothingness, and Subjectivity":

> […] How much of a change between [me] and [modified me] is necessary to destroy personal subjective continuity? At what point, that is, would we start to say "Well, [Tom] 'died' and a stranger now inhabits his body; experience ended for [Tom] and now occurs for someone else"? It is not at all obvious where to draw the line.

It seems logical to believe that a very small change in brain structure under surgery, say, on the scale of a few neurons, would not change the conscious entity inside his body. If we accept this, it is also logical to believe that making radical changes under surgery would also preserve his conscious entity. This may sound like a slippery slope argument, but the alternative, given the irreducibility of the perceived type-2 self, may be even less plausible: at some highly specific threshold of brain alteration, the conscious entity would change, and any less degree than that threshold would mean the original entity survives. In other words, we would have to accept that the difference between subjective experience continuing for Tom and subjective experience ending for Tom (while beginning for another subject) would come down to a single brain cell. But if we accept that Tom's subject is preserved after making radical modifications to brain structure in surgery, we may as well also accept that in the death of one arbitrary person followed by the birth of another arbitrary person, the new individual born is the same subject as the individual that died. This is because both death to birth and extreme brain modification under surgery involve radical changes to brains between streams of continuous conscious experience, and it is hard to see how these situations would be viewed differently in the eyes of nature. So again, we are confronted with conflicting intuitive ideas when accepting the notion of differentiable subjects.

To resolve this dissonance, there are two positions one may take. One position may be similar to Hayworth's. On this account, we would be denying that Tom is some unique subject of experience. Despite that Tom perceives his conscious subject of experience as an irreducible entity, and that it makes sense that *his* experiences are *only his*, there is somehow an ontological overlap between his core self and another person's core self if the content of their POV or MEM selves are similar. There is no ontologically objective way to answer whether or not Tom (as an experiencer of consciousness) died based on the amount of brain changes that occurred during surgery. Rather, it is simply a matter of what one considers to be "Tom" (as an experiencer of consciousness). One who takes this position may

not worry about death at all if they have a twin who is very much like them. In their view, since the conscious entity embodying the twin is hardly different from their own, they can die without noticing much change. This position should theoretically be held for one who disputes the idea that entities are differentiable regardless of phenomenological content.

Alternatively, one may hold that everyone is entirely affected by their own death as far as subjective experience *for them* is concerned. It makes sense to talk about a "me" experiencing the world, because "I" perceive it to be there. My body's death would affect me, and only me, directly. Having a twin would not mean that I survive my own death any more than someone without a twin would. Some of my phenomenal content would be preserved, but I wouldn't be there to experience it. It is true that if we accept this position, answers to questions such as "at what point does my conscious entity get replaced with another one in brain modification surgery?" or "at what point in brain disintegration does my perceived type-2 self just disappear?" become less clear. But this is no reason to deny the reality of what our consciousness fundamentally comprises. These questions present epistemological uncertainty as opposed to ontological uncertainty, and there is no reason to believe that the tools of science will not be able to get us to an answer. I argue that despite all of these attempts to explain away the self as a non-problem, there is no evidence against what our intuition actually tells us about you and I being fundamentally different subjects. Additionally, negating the importance of our perception would be very difficult. I know from my experience that "I" am here in *this* body around me, and not in some other body such as my mother's. Could I be wrong about this? It is very hard to see how.

To summarize: The view of a self that stands independent from the content of subjective experience has been discredited by numerous philosophers, but without clear good reason. Existence of pathology, models demonstrating the degree to which the sense of self can be manipulated, and self-boundary thought experiments do not refute the existence of this self. While some may argue that the existence of this self would overturn numerous findings in psychology and neuroscience, I submit that it would be far easier to accept that our picture of consciousness is simply incomplete than to deny the foundation of my very existence.

One further subject to touch on before moving forward is the idea of a deflationary identity. Under this view, the type-2 self is unstable and does not survive throughout an organism's lifespan (Chalmers 2010). This is because occasionally, one's conscious entity is being replaced by another conscious entity, and the new entity then captures the memories of previous entities as if it were its own. Some possible reasons to hypothesize this replacement will be discussed in the next section. This view does not dispute the existence of the type-2 self as defined in this paper, but holds that any *particular* type-2 self in an organism is only maintained for a short amount of time, as opposed to its entire lifespan as we might think. It is necessary to accept that type-2 selves are differentiable to hold this view, because this view specifically states that "selves" are replaced in spite of a (mostly) unchanging MEMself.

5. The Power of Potential

If we accept that we have something that we can call irreducible and differentiable type-2 selves, the need to solve the problem of how to preserve this type of self in an upload becomes clearer. However, I would argue that by only asking "how do I know that an upload will be me?", we are not confronting the root of the problem. Ultimately, our perspective should shift from the question of how to preserve identity when changing substrates to the question of how to create a conscious system with our identity from scratch. This new perspective has more power because in theory, it would not only allow effective uploads of us, but it opens up a possibility to resurrect those who have already passed.

In other words, the true nature of the mystery lies in the question of why the body you find yourself in possesses *your* conscious entity *as opposed to someone else's*. Just as that there is nothing special about a living brain from an objective standpoint that would lead one to believe that it is conscious[6], there is nothing special about the molecular arrangement of your body that would lead an objective observer to look at it and believe it is you (as opposed to someone else) in there. What I would like to propose is a new framework that can make more sense of this problem.

In order to grasp the new framework, let us review the concept of *potential* in science, using *potential energy* as an example. Potential energy is a useful construct because it allows us to make predictions about the future state of a system. For example, a roller coaster about to take a 100 foot plunge would have more potential energy than a roller coaster about to take a 30 foot plunge (relative to the bottom of each respective track), even though the physical state of the coasters would be identical at the top of each hill. The word "potential" in this context simply refers to energy that is not currently active, but may become active. This is an example of a practical reason to conceptualize a property or entity that is imaginary in the eyes of nature. The interest to science is to learn the mechanisms behind how seemingly imaginary properties become very visible and real. We have a fairly established science that can explain how potential energy translates into kinetic energy.

With consciousness, the problem is almost a perfect analogue to potential energy. The main difference is a lack of science behind it, but understanding it in this framework should help lead us toward one. Our new framework will involve conceptualizing the "existence" of non-existent conscious entities. We will define a "potential subject" as a conscious entity, an experiencer of consciousness, *that does not exist*. An "active subject", on the other hand, will be defined as an experiencer of consciousness that *does exist*.

The reason to posit a construct of potential subjects is the same reason to posit potential energy. We can begin to understand this by considering temporary disruptions to consciousness – that is - a period of unconsciousness between two periods of consciousness for a particular subject. Some things that may cause this include dreamless sleep, being put under general anesthesia, or suffering severe head trauma[7]. In any of these cases, during the time that *you* are unconscious, *you* would be referred to as a *potential* subject of consciousness at that time. The justification for ascribing a term to a presently non-existent feature is much the same reason we would say a roller coaster on the top of a hill has potential energy; we are referring to the future state of the system. In the case of the roller coaster, we are referring to the energy that will exist (mostly in the form of speed) when the coaster gets to the bottom of the hill. In a living but unconscious system, we are referring to the conscious entity that will exist when the subject wakes up. You, as a subject (experiencer) of consciousness, would be restored to the active state.

While it may take a bit more imagination, a particular conscious organism can be thought to have a potential subject before its conception and after its death. Assuming consciousness is restricted to my living brain, the conscious entity that is me is *active* now, but *potential* before my conception and after my death, just like it is potential during general anesthesia. While this may sound like a dualist position, it is not. The potential subject is merely an entity we are constructing, much the same way we construct the idea of potential energy, because it allows us to make predictions about future consciousness. The potential subject is not a "spooky" thing.

So to clarify a central point: Every conscious organism alive today has an *active subject* of experience, as well as a corresponding *potential subject*. You can imagine this as a giant board of on-off (or in this case, active-potential) switches, with one switch for each organism with a type-2 self (which is either present

or absent, because it is irreducible). At any given time, each switch is either on or off. Additionally, when an organism dies, the switch is permanently shut off (to simplify the problem – we'll momentarily assume that there is no consciousness after death). The switch does not disappear however, because *in principle* it could be switched on again[8]. So in addition to the switches for organisms alive, which can be either on or off depending on the organism's current activity, there is a whole set of permanently off switches for organisms that have died. There is also a whole set of switches for all conscious entities that will come into existence in the future, but for now those switches are off.

While the amount of conscious organisms that will ever be born is probably a humongous number, even this does not represent all of the possible subjects that *could* come into existence. Consider the following quote from the beginning of Dawkin's 1998 book Unweaving the Rainbow:

> We are going to die, and that makes us the lucky ones. Most people are never going to die because they are never going to be born. The potential people who could have been here in my place but who will in fact never see the light of day outnumber the sand grains of Arabia. Certainly those unborn ghosts include greater poets than Keats, scientists greater than Newton. We know this because the set of possible people allowed by our DNA so massively exceeds the set of actual people. In the teeth of these stupefying odds it is you and I, in our ordinariness, that are here (Dawkins 1).

The "potential people" in the above quote can refer to potential subjects that will never become active. Even though they will never see the light of day, their potential conscious entities (that will never become activated) are something we can make reference to. In addition to the fact that "the set of possible people allowed by our DNA" is huge, we illustrated earlier that two identical bodies may have different conscious subjects, so even this set does not represent an upper limit to the amount of potential subjects. There really is no conceivable limit to the amount of conscious organisms that *could theoretically* come into existence, and there is no upper limit to the amount of conscious subjects we could *imagine*. I would therefore argue that the number of *potential subjects* is unlimited or infinity. Even assuming that the number of conscious organisms that will ever exist in the multiverse is finite, there are an infinite amount of subjects that could become conscious but never will become conscious. The potential subject concept still works – because *in thought world* we can imagine that the multiverse will output infinite life given an infinite amount of time (even though infinite time is unlikely). These infinite subjects will just always stay at the "potential" side of the switch. So why pretend they even have potential? Even if the laws of physics dictate that these infinite subjects have no potential, the laws of philosophy do not. It's the laws of philosophy, in this case, that are going to get us to the answers we need and tell us about what we can achieve with them.

So your body, the biological body that you find yourself in right now, did not have to contain your conscious entity. From an objective perspective, not only could it just as easily contain my conscious entity, or anyone else's, but it in fact had an infinite amount of options! Even if it wasn't really an "option", like a God choosing a conscious entity to put inside your body, something in nature must have determined it. This is our big second explanatory gap in consciousness studies. How do we go from knowledge that living organisms have conscious entities to the idea that *specific* conscious entities are mapped to *specific* bodies? What distinguishes a lucky potential subject that will eventually find itself in the light of the world via an organism and one that will not? When subjects are extinguished by death, are they naturally placed back into the lucky bin?

These questions can ultimately be collapsed into this one: "How do specific potential subjects become active subjects?" The answer is not clear at all, and this is our ultimate missing science, which I call "the selection problem". But when we examine this issue head-on, we see that the options are surprisingly quite limited. Though the possibilities are broad, we can at least begin to break down the issue to design good experiments.

6. Possible Solutions and Associated Problems

6A. Parameters that may identify the self

To think about the selection problem, we can consider some parameters that could potentially preserve our type-2 self in an upload. It is natural for many deep thinkers to hold a position on what causes a type-2 self, and whether or not this criteria will be met in an upload will determine whether or not they believe an upload will be effective in preserving identity for a subject. Let's explore the possible parameters of physical properties that could be the answer to the selection problem, based on the principle that an organism's brain as a whole is what "selects" one entity out of the infinite options. These parameters essentially represent three mutually exclusive positions one could take on the selection problem.

- **A:** Numerical molecular arrangement: the numerical identity[9] of the molecules arranged a particular way in your body is what determines everything about your personal identity, including the conscious entity your body will contain. Therefore, uploads can never contain the entity that was in the original person.
- **B:** Qualitative molecular arrangement: the qualitative identity[9] of the molecules arranged a particular way in your body is what determines everything about your personal identity, including the conscious entity your body will contain. Since this would require an upload to be a carbon-based biological substrate, and an exact copy down to the molecular level would inherently include damage from aging along with the natural aging process, there may be little point unless these features could somehow be removed without changing the conscious entity.
- **C:** Qualitative functional arrangement: A Body's conscious entity is determined by its functional parts. Therefore, uploads will preserve identity if equivalent functional parts of the original brain are preserved in a substrate independent mind, even if the substrate is non-biological.

Upon philosophical investigation, these possibilities face a number of challenges. The more obvious one is that all of these parameters are changing continuously over time. If one were to accept that the conscious entity inside a conscious system is *entirely* determined by the *current state* of any of these parameters, they would be accepting quite an extreme version of the deflationary view. We would literally be dying (by conscious entity replacement) probably hundreds of times per second without knowing it, because the chemistry of our brains is always changing[10].

This particular problem does have a conceivable solution, however. It is necessary to separate the questions "how do we create a specific conscious entity?" and "how do we preserve a specific conscious entity?" When a living organism is developing and reaches sufficient complexity for consciousness (and genesis of a conscious entity), one of the above parameters (A-C) may be called upon to make the "selection" of which subject to activate out the infinite options. But once the entity is generated, the organism may be able to hold on to it in spite of a physically evolving brain. Perhaps only when consciousness is regained after a temporary period of loss (such as sleep), the brain would generate a new conscious entity (which would be "selected" by whichever parameter was called upon in the first place, but the new entity would be selected based on the current state of the body). This "stream stabilization" view is also a deflationary view, but a less extreme one because a continuous stream[11] of consciousness presumably lasts a bit longer than an instantaneous conscious moment. To resist a deflationary view altogether, one could postulate a "life stabilization" view, holding that a single type-2 self will "jump" from one conscious stream to the next throughout the lifespan of the organism. Or, perhaps consciousness is never truly and fully lost for an organism as long as it is alive.

If any version of the deflationary hypothesis were true, some may see little point in crafting an uploading system to preserve a person's specific conscious entity, because the entity is not even preserved in everyday life. Any form of uploading would still be better than nothing though, as the type-1 self would be preserved. But still, the mystery of how the selection occurs each time a "new self" is generated would remain. It still may be worth solving this puzzle in order to make survival through uploading a more meaningful thing.

If the "life stabilization" idea held any merit, we may have two options for uploading. One option would involve somehow tracing back the history of the organism to the original point at which it became conscious, and re-create that structure to activate the same entity in the new substrate. This would be nearly impossible. Another method would be to simply *preserve* the current entity, perhaps by *gradual destructive uploading*. This process refers to replacing each brain part, one at a time, until the entire system is uploaded (Chalmers 2010).

6B. Why these Parameters are Probably Wrong

Ultimately, these parameters (A-C) are more questionable than we may make them out to be. I find it extremely difficult to imagine A. We would effectively need to accept the following: starting with a conscious system, we could destroy it, and then rebuild it using the same numerical molecules. At the same time, we build an exact copy of the system using qualitatively identical molecules. The numerically identical body would have the same conscious entitiy as the original and the qualitatively identical body would have a different entity. Since both bodies have identical function in the laws of nature, why would this be? It is hard to imagine why nature would care which of the two, if any, gets the original conscious entity.

Should B or C be true, imagine the following scenario. Two twin bodies are generated in identical environments, so both bodies contain the exact same qualitative configuration of molecules. Under this hypothesis, the two twins are not separate subjects – they are the same conscious entity in two places at once. This is because parameters B and C do not allow for the numerical identity of molecules to have a role in the selection problem. Perhaps this is easy to conceive if the two bodies are experiencing identical qualia, but what if, following the initial time that these bodies were generated, one was then led into a different environment? If stream or life stabilization were true, we would have the same entity perceiving two separate worlds simultaneously. The only alternative under the B and C parameter view is to reject the stabilization views while accepting the extreme deflationary position.

There is an even more fundamental issue, which trumps almost everything discussed so far. Even if we were able to understand how consciousness arises, how we develop our sense of self, and how to map potential entities into different bodies based on some physical arrangement parameters, we would still be left with a huge explanatory gap. *What is it* that makes one particular arrangement or set of molecules *me* as opposed to *someone else*? Or we could think of this from the other direction: why is *my* consciousness the equivalent of *this* particular set of molecules as opposed to any other? It seems that we may be inevitably faced with a fundamentally random process in nature. To me, this idea is just as mind-boggling as the randomness in quantum uncertainty. Intuitively, nothing in the universe should be random. Albert Einstein clang to this intuition of systematic cause and effect, ultimately rejecting central aspects of quantum mechanics. Perhaps someday, new physics will illuminate a perspective on quantum uncertainty that will allow us to grasp the nature of random output. Similarly, the apparent randomness found in the nature of the self may be within theoretical grasp.

While this is one possible endpoint, we have not exhausted all possible options. Instead of taking the "whole brain" approach which involved parameters A-C, perhaps a more plausible solution to the selection problem is that some central mechanism in the brain is responsible for selecting one potential subject (as opposed to other potential subjects) to activate. By "central mechanism", I mean an explicit large-scale brain function or reaction that reliably has the same entity activated even with conscious stream disruptions throughout the organism's lifespan. Even though this would implicitly depend on qualitative molecular arrangement (parameter B), the difference is that the mechanism for the selection process would be explicit, thus removing the explanatory gap[12]. While this may sound like an absurd idea, the only conceivable alternative is to accept some fundamental randomness. This mechanism may be a quick silver-bullet in our search for truth, similar to the discovery of DNA as the "life force" that has always been searched for.

Thus far, we have been assuming that consciousness is generated by the brain, and consciousness for a subject begins and ends in their own body. This is commonly taken as a given, since modern neuroscience can demonstrate one-to-one mapping between brain function and mental phenomena. However, my overall take on this position is that it only considers the individual parts of consciousness (such as senses and cognition) and fails to consider consciousness as a whole with the irreducible sense of self. Additionally, while the quality of evidence for consciousness survival of bodily death is debatable, it would be foolish to ignore the fact that some evidence exists, for example, see work developed by Gary Schwartz and Ian Stevenson. Efforts have been made to link quantum mechanics and consciousness, with growing success (Radin 2012, Hameroff 2014). Perhaps such mechanisms could allow theoretical room for the brain to act as a receiver, rather than a producer, of consciousness. In any case, we can take natural survival as a *theoretical possibility* and consider what implications it may have on this model of the self.

In this scenario, perhaps the infinite amount of potential subjects we postulated are really active subjects embedded in universal fabric[13]. All possible conscious entity "switches" imaginable (an infinite amount of them) would be on, at least periodically. This infinite sea of active subjects would be an unlimited supply of consciousness and in line with many eastern philosophies. Specific active subjects/entities may get drawn into specific organisms when some critical feature in their living brain is formed. Alternatively, perhaps particular entities ingrained into the universe can split off from the sea and play a role in generating a body for itself, though this is purely speculative. This may seem like an extreme violation of Occam's razor: why postulate an infinite amount of unnecessary entities? But postulating these entities may actually yield the simplest explanation, which will be discussed in the next subsection. Someone living in a 2-dimensional flatland may think the idea of an infinite number of flatlands stacked on top of each other is absurd, but in reality we know that an infinite amount of these flatlands simply creates an extra dimension. There may be a simple, *natural* answer within the laws of physics that can explain the existence of this higher dimension of consciousness, with our brains simply accessing a single infinitesimal slice of this dimension[14].

6C. Different types of infinity and why they are important

We are faced with some questions at this point about the nature of infinity in these contexts. In mathematics, there are multiple types of infinity, some of which are provably larger than others. One category of infinity is *countable* infinity, which includes a set of all real numbers that can *in principle* be listed (for example, integers and fractions). Another category is *uncountable* infinity, which includes the set of *all* real numbers. One cannot, even in principle, list all the real numbers that lie on a number line. These so-called uncountable infinities are therefore larger than countable infinities[15]. Countable infinity can be represented by an infinite amount of discrete points, while uncountable infinity would be a

continuous line; discrete values, no matter how small, are always separated by an infinite amount of numbers.

Applying these concepts to active and potential subjects is difficult at best. But it matters, because it is possible to have *both* an infinite amount of potential subjects *and* an infinite amount of active subjects. An infinite amount of active conscious entities (subjects) does not necessarily include the *entire possible set* of conscious entities, because infinity plus infinity equals infinity. It is possible that there naturally exist an infinite amount of active conscious entities, but instead of being embedded in universal fabric, this really could just be due to an infinite amount of space or time in the multiverse yielding countably infinite life. If this were the case, it would be very easy to imagine a separate set of infinity potential subjects that will never exist. The number of active subjects in this case should be countable infinity, because actual living organisms are discrete. However, the type of infinity describing the number of potential subjects would be less clear. One could make an argument that this is countable infinity, because even imaginary subjects are discrete and should be able to be lined up side by side like geometrical points. However, one could also make an argument that it would be uncountable infinity, because in principle one should be able to imagine a distinct potential entity corresponding to every possible real number on a number line[16].

Under some configurations, it can be shown that the chances are of any particular subject (such as *you* or *me*) ever existing are virtually zero. If there are an infinite amount of possible conscious entities for an organism to select besides mine, the chance of any particular developing conscious being becoming *me* is $1/\infty$. The probability of any particular conscious entity (me, you, etc) being drawn given n opportunities is thus[17]:

$$1-\left(1-\frac{1}{\infty}\right)^{n}$$

Let's take n to be the number of conscious systems that have existed, exist now, and ever will exist in the multiverse. *If n is a finite number, this probability is infinitesimally close to zero.* I do not necessarily take this as evidence against an only finite amount of conscious systems ever existing, because nature should not care about this tiny probability of me existing. However, this clearly demonstrates that our existence under these circumstances is nothing to take for granted, and many people probably overestimate the chances of existing naturally.

If we consider that perhaps an infinite amount of conscious systems will naturally exist over the course of multiverse space and time (though not the universal fabric view of consciousness) we can evaluate the following:

$$\lim_{n\to\infty} 1-\left(1-\frac{1}{n}\right)^{n}$$

This expression also yields zero. While the infinity represented in the denominator and the infinity represented in the exponent have different meanings, this does not affect the answer. It seems that even given a countably infinite amount of opportunities for your consciousness to be born into a conscious organism, you should have had virtually no chance of existing to read these words.

To eliminate this idea, the only feasible solution is to suppose that there are zero potential subjects. This would mean that *you cannot think of or imagine a possible conscious entity that isn't (at least periodically) conscious*. In other words, every entity in the *entire possible set* of conscious entities is active. This would be consistent with the hypothesis of consciousness being embedded in universal

fabric, because otherwise it would be possible to imagine a separate set of infinity potential subjects over and above the countably infinite conscious organisms. The type of infinity describing the number of active subjects in this case is open to debate though, as described above for counting the entire possible set of potential subjects.

7. What's Next?

In this section, I will make a basic outline for how we can proceed to utilize this model of the self scientifically. The basic goal is to determine what naturally occurs after death, and then figure out how we may be able to control the future of our conscious experience if we don't like the natural answer.

If naturally existing consciousness after death could be tested and proven, we need not bother with uploading or any efforts to create immortality. To start, some forms of qualitative research may be useful. For example, we can look at the phenomenology of transcendental experiences, which are considered by many to be indications of a world beyond. Transcendental experiences can include near-death experiences (NDEs), meditations, as well as effects from certain drugs. Namely, we should be looking for some experience of a self being connected to an infinite amount of other selves, as the model I proposed would suggest. While such reports have been described[18], it may be worth examining the phenomenology of their experience in greater depth to see how well it matches up to the ideas presented in this model.

But eventually we need actual proof of something. We need a complete theory of consciousness that can explain the mechanisms of the self and qualia the same way that the theory of evolution can explain the diversity of life. What are the options for solving the selection problem of how potential subjects become active subjects? We first may need to know more about how a subject (or type-2 self) forms. Research into altered states of consciousness, from both drugs and pathology, may still be invaluable. This is because we need to understand *all* the forms that consciousness can take, and what the brain is doing in each case. This will become easier with improving brain scanning technology.

Ultimately, if we can figure out where and how the irreducible type-2 self breaks down, we may be able to explain the mechanism of the type-2 self. Fortunately, scientists are seemingly making headway with technology that can selectively turn off parts of the brain. For example, the Transcranial Magnetic Stimulation device can non-invasively affect parts of the brain by magnetic fields. Perhaps more sophisticated forms of this technology, when developed, could effectively simulate the disintegration process of the brain (without causing actual brain damage). A subject reporting their experience could give scientists useful insight on the scale of brain activity needed for coherent perception of an irreducible conscious entity. Further, if there are indeed better forms of consciousness than qualia attached to an irreducible self, we may be able to understand how those work as well, and consider those for uploading efforts in lieu of saving our differentiable selves.

Simply by knowing the mechanisms of the type-2 self, the answer to the selection problem might be right in front of us, or it may at least uncover some hints. Once we have the answers, we then may have complete control over turning potential subjects into active subjects, therefore giving us the ability to upload or resurrect *any* conscious entity. If this seems like too much of a slippery slope argument, one may also consider hope from the singularity, which is the idea that there will be an explosion of intelligence (and resulting technology) once a human invents a computer "smarter" than human itself. Such an explosion could help us out tremendously in not only developing uploading technology, but also determining how to upload in order to preserve identity.

8. Conclusion

While the underlying reality of the self may be vastly different than what our intuition tells us, the importance of the way we perceive it cannot be neglected. Thus, it makes practical sense to distinguish two types of self, with the irreducible entity type of self giving us objective answers to whether or not a conscious system survives under specified conditions. Given the further fact of differentiability, there is an explanatory gap between our knowledge of the existence of these entities and our knowledge that specific entities are mapped into specific conscious systems. The underlying scientific question to answer is how potential subjects become active subjects. When broken down, it can be seen that possible solutions are quite finite, and can likely be solved with the tools of science in the near future. When including the selection problem in the quest to explain why we are not zombies, perhaps the puzzle of consciousness will finally come together.

Footnotes

1. See Chalmers 1995.

2. By physical place, I am referring to how consciousness fits in with the physical laws of the universe.

3. Uploading can take many forms, but most types discussed in literature involve scanning the brain in its microscopic parts followed by recreating it in an alternate substrate. The alternate substrate is usually a nonbiological functional isomorph, such as a silicon-based computer system with individual circuits taking the place of individual neurons in their original locations with respect to the brain and spinal cord as a whole. See Chalmers 2010 for more background on uploading forms.

4. Once again, I would direct readers wishing to seek more background on this topic to begin with Chalmers 2010.

5. The term "further fact" is often used to portray the view that there are facts about a person's identity beyond the sum of their phenomenological content.

6. This is the essence of the hard problem. From a perspective outside of a living organism, it is difficult to imagine how looking at all of its microphysical components and interactions would lead the observer to believe that the system is conscious.

7. For the sake of simplifying the problem, we can assume that these situations will render a *complete* lack of consciousness, equivalent to the phenomenology of not existing.

8. By *in principle*, I do not mean that it is possible within the laws of physics. Rather, I mean it is theoretically imaginable.

9. "Numerical identity" and "qualitative identity" are two different ways of considering whether or not two things are "the same" in the famous Ship of Theseus thought experiment. For example, two different carbon atoms with the same properties would be qualitatively identical, but not numerically identical. A carbon atom is numerically identical to itself and only itself. Since a whole brain can only be numerically identical to itself, a view that the numerical identity of brain molecules holds ontology to the conscious entity selection is inherently pessimistic about uploading.

10. Brain cells continuously die and regenerate as our bodies grow. Numerical identity of all of our brain molecules is never preserved in everyday life. This dilemma is slightly less extreme for parameter B and perhaps even less for C, as it could take a bit longer for qualitative and functional changes (i.e. memory encoding) to occur, but these are still very short timescales and it is nonetheless an extreme view. However, this is not to say that an extreme view is necessarily wrong.

11. A conscious stream will be defined as a continuous and unbroken period of consciousness. While it is difficult to know with certainty whether some things (such as deep sleep, general anesthesia, etc) truly cause a complete lack of consciousness (and thus break a stream), I speculate that most people will experience multiple streams throughout their lifetime.

12. Parameters A-C by themselves involve only *implicit* mechanisms for the selection problem, because there is no objective reason to assume that any particular configuration of molecules would favor one particular entity over another. Saying that a person's identity is determined by their molecular arrangement does not eliminate the randomness discussed, and thus we are still left with an

explanatory gap. However, identifying a specific mechanism inside the brain that explicitly polarizes a specific entity would directly explain the selection problem.

13. “Universal fabric” is my umbrella term for features at the quantum-scale of the universe. Things like time and spatial dimensions can also be considered part of universal fabric.

14. I am referring to this version of consciousness as a “higher dimension” as an analogue to the nature of spatial dimensions. An infinite number of n dimensional objects stacked together create a shape in the n+1th dimension, so an infinite amount of irreducible conscious entities embedded in universal fabric, in a sense, constructs a higher dimension of consciousness.

15. See Cantor’s diagonal argument for an explanation as to why all numbers on a number line are unlistable. It is generally accepted that this is proof of uncountable infinity being larger than countable infinity. While the exact ontology of these types of infinity may be disputable, the concepts that follow on the entire possible set of conscious entities should hold.

16. Mapping of specific conscious entities to specific number values would be an entirely arbitrary process. The purpose of imagining this is to think about this problem from a mathematical perspective so that it can be grasped.

17. If the probability of any particular developing conscious organism becoming me is$\frac{1}{\infty}$, the probability of that organism NOT becoming me is$1-\frac{1}{\infty}$. In order for me to never exist or know of existence, this probability would need to be realized for every single conscious organism that ever has and ever will exist in the multiverse. We don’t know how many such organisms there are, so we assign a variable *n* to represent it. The expression $1-\frac{1}{\infty}$ can be raised to the nth power to determine the probability that I will never exist, because this probability will need to be realized n times. To change this to the probability that I *will* exist, this new term $(1-\frac{1}{\infty})^n$ is subtracted from 1 to get the overall expression: $1-\left(1-\frac{1}{\infty}\right)^n$

18. This is a universal description in many near death experiences and other transcendental states of consciousness achieved by experienced meditators. For an example that provides particular detail, see LaBerge 2012. See also Jourdan 2011 for a detailed study on the higher dimensional perceptions of the near death experience.

References

Blackmore, S. (2002). The grand illusion: Why consciousness exists only when you look for it. *New Scientist* 174 (2348):26-29.

Chalmers, D. "Facing up to the problem of consciousness." *Journal of consciousness studies* 2.3 (1995): 200-219.

Chalmers, D. "The Singularity: a Philosophical Analysis." *Journal of Consciousness Studies* 17.9-10 (2010): 9-10.

Clark, T. W. (1995). Death, nothingness, and subjectivity.

Dawkins, R. *Unweaving the rainbow: Science, delusion and the appetite for wonder*. Houghton Mifflin Harcourt, 2000.

Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR'theory. *Physics of life reviews*, *11*(1), 39-78.

Hayworth, K. (2010). Killed by bad philosophy: Why brain preservation followed by mind uploading is a cure for death,". *Essay published online at http://www.brainpreservation.org*.

Jourdan, J. "Near Death Experiences and the 5th Dimensional Spatio-Temporal Perspective." *Journal of Cosmology* 14 (2011): n. pag. *JournalofCosmology.com*. Web. 16 Jan. 2012.

LaBerge, S, & Brown, D. "Waking the Dreamer." *Mavericks of the Mind*. N.p., n.d. Web. 16 Jan. 2012.

Metzinger, T. (2009). The Ego Tunnel. The science of the soul and the myth of the self.

Radin, D., Michel, L., Galdamez, K., Wendland, P., Rickenbach, R., & Delorme, A. (2012). Consciousness and the double-slit interference pattern: Six experiments. *Physics Essays*, *25*(2), 157.

Ramachandran, V. S. (2004). A brief tour of human consciousness: from impostor poodles to purple numbers.