MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Is Mind-Uploading Possible?

A Thesis Presented in Partial Fulfilment of The Requirements
For the Degree Of

Master of Arts
in Philosophy

at

Massey University, Manawatu, New Zealand.

Nathaniel Honore Brown

2024

Contents

Abstract

This thesis is a philosophical evaluation of mind uploading. 'Mind-uploading' is a trope common in science fiction where we move our mind/identity to a computer. I evaluate three classes of mind-uploading: 'destructive', 'non-destructive', and 'gradual' uploading. I assess the different intuitions about whether each upload would enable us to continue as the same person. I consider the Hard Problem of consciousness - the conceptual gap between the description of a physical system and the description of a mind. I note that the apparent inaccessibility of consciousness from the third person creates a barrier to knowing whether we could survive a mind-upload. I briefly survey different philosophical accounts of the mind and note that dualism generally appears more compatible with mind-uploading than physicalism. I argue that our intuitions about mind-uploading tend to support a dualist view of the mind. However, substance dualism cannot be reconciled with a belief in mind-uploading without making implausible claims.

I then explain how a physicalist ontology can be rendered compatible with mind-uploading. To do this I rely on a robust presentation of computational functionalism, which argues that the mind is reducible to the right type of organisation and information processing. To support the compatibility of mind-uploading with physicalism, I offer the contours of how functionalism could refute the existence of the Hard Problem by appealing to the evolutionary origins of phenomenal consciousness and a Humean account of causation. I suggest this means we can have some confidence that a computer could have an identical phenomenal experience to our own, even though it seems impossible to know what conscious states someone is experiencing from the third person.

In the final chapter, I assess whether our personal identity could be transferred in a mind-upload. This is a crucial requirement for the success of mind-uploading, as even if a computer could replicate our conscious states, without the ability for our identity to transfer across physical systems, mind-uploading would not be feasible. I argue that our intuitions about survival (or otherwise) in mind-uploading cases would tend to undermine most traditional accounts of personal identity. I also discuss Derek Parfit's critique of personal identity and apply it to mind-uploading cases. I reconcile the different intuitions about mind-uploading by appealing to the closest continuer account which I argue provides a persuasive answer about whether we survive mind-uploading procedures. If the closest continuer account of identity is true, then it appears that we could upload our minds.

Acknowledgements

In preparing this work, I am indebted to two individuals in particular:

My supervisor, Dr. Gerald Harrison, has been generous with his time and provided a great deal of valuable feedback on both this thesis and all other work I produced in this Master's program. I enjoyed the time spent together at the pub debating issues within metaethics and the philosophy of mind. While we never agreed on much, I think my work and our conversations were more interesting for it.

Janelle Robinson has been my biggest supporter and best friend. She spent days at Massey University with me while I worked on this project and has listened to me drone on about philosophy for years. I am grateful for her love and support.

Overview

This thesis is about reconciling mind-uploading with a purely physicalist view of the mind. Mind-uploading is a common pop-cultural trope. It involves the creation of a computer-simulated copy of a brain that transfers consciousness and personal identity from the person to the computer. It has attracted attention in fiction and philosophy, and many scientists and philosophers believe that no fundamental barrier to mind-uploading exists. This thesis examines two questions - could a computer possess our type of consciousness? And could our personal identity be transferred to a computer? I tentatively submit that plausibly the answer to both questions is "yes". I advance an approach that suggests mind-uploading is possible without conceding ground to dualism. I also argue that both 'gradual uploading' and 'destructive uploading' constitute survival. While I believe that every argument I make in favour of mind-uploading is plausible, it's important to note that mind-uploading requires several conditions to be met. If any of these conditions are not satisfied, mind-uploading will not be possible.

This work is the result of extensive research, drawing particularly from the themes from the works of David Chalmers, Richard Swinburne, and the late Daniel Dennett and Derek Parfit. I argue that computational functionalism is a plausible account of the mind. If computational functionalism is true, then a computer could possess our consciousness. My argument is that the mind is plausibly a system of information processing, and that identical information can be expressed by different physical systems so long as those systems have the right form of 'functional organisation'. This suggests that non-biological systems could have our type of consciousness.

I acknowledge strong grounds for scepticism that our personal identity would be preserved if we attempted to upload our minds. Failure to preserve personal identity would make attempting certain types of mind-uploads suicide whether or not a computer could have our kind of conscious experience. I circumvent this risk by appealing to the closest continuer account of personal identity. The closest continuer theory posits that personal identity is non-branching, and the entity most closely and sufficiently linked to a previous entity is that entity. This view is compatible with successful mind-uploading. I endorse closest continuer accounts as a persuasive philosophical account of identity and the position best able to accommodate fringe cases such as 'partial brain transplants' which undermine other accounts of personal identity. So long as the mind-upload is our closest continuer we will persist as the mind-upload. Therefore,

we have reason to believe we would survive a mind-upload procedure if the technology became available.

The first half of my thesis explains why scepticism about mind-uploading persists and what the justifications for it are. I also outline destructive, non-destructive and gradual mind-uploading procedures. Each distinct method of mind-uploading raises different intuitions about whether we could persist as a mind-upload.

I make the point that for a mind-upload procedure to be successful, it would need to replicate our 'phenomenal consciousness'. This form of consciousness describes first-person subjective experience, such as the pain of pain. Chalmers (1997) argues that there is a 'Hard Problem' of consciousness. This is the alleged conceptual barrier between a full description of a physical system and a description of phenomenal consciousness which suggests the former cannot describe the latter. The Hard Problem suggests it is impossible to know (as an outsider) whether an attempted mind upload had been successful.

I then discuss dualist approaches to the mind. I argue that although most forms of dualism are compatible with mind-uploads and that mind-uploading itself seems dualistic, substance dualism is not compatible with mind-uploading despite appearances. I propose that reconciling substance dualism with mind-uploading requires the substance dualist to make implausible assumptions about how souls can be transferred from one objects to another. Since I am interested in supporting the possibility of destructive uploads while supporting physicalism, I offer a rebuttal to substance dualism relying on the phenomenon of split-brains. I offer split-brains as an interesting example of a real-world problem substance dualism struggles to explain. I use the example of the split-brain phenomenon specifically because it is thematically gels with this thesis topic.

I argue in favour of a physical explanation of the mind and provide a rough outline of computer functionalism. I note that functionalism explains how different kinds of physical organisation could give rise to our form of phenomenal consciousness. I offer a picture of the mind as the instantiation of a set of computations/formal systems. The mind, in this view, will be substrate-independent without being dualist (in a problematic sense). The same computations can be performed by different systems, which suggests that it may be possible to upload our minds. I

believe that computational functionalism provides a plausible explanation of the mind, but I recognize that the account I offer is basic and will require further development.

I then suggest the Hard Problem is answerable. I argue a computer could have all the properties we ascribe to minds, drawing on arguments made by Dennett (1991, 2017). I support his recent argument that phenomenal consciousness is linked closely to the necessity for organisms to communicate their internal states with others, which requires first person subjectivity. Dennett offers an evolutionary explanation of phenomenal consciousness. I also draw on themes from David Hume to explain why a full physical explanation of consciousness may not feel complete without this suggesting dualism. In aggregate, I suggest that these arguments are a start at answering the Hard Problem. If I am right, then we can be sure that our mind upload would have a mind identical to our own.

In the last part of my thesis, I explore the implications of attempting mind-uploading for personal identity. I consider different perspectives on personal identity and how they relate to various forms of mind-uploading. Additionally, I support an argument made by Derek Parfit (1984) in his book, *Reasons and Persons,* which also discusses brain transplants and split-brain scenarios. This argument challenges the traditional accounts of personal identity, ultimately suggesting that these views are not logically coherent. Parfit believed that we should disregard the significance usually attached to identity and instead care only about psychological continuity to us. Instead, I offer an alternative resolution to Parfit's challenge by endorsing the "closest continuer" account of personal identity. According to this approach, we could survive as a mind-upload if the mind-upload is our closest continuer.

1. What Needs to be True for Mind-Uploading to be Possible?

In this chapter, I discuss different methods of attempting a mind-upload. The relevance of this will not be obvious at first but the different classes of mind-upload will have implications further on. This chapter also discusses the necessary (and together sufficient) conditions for mind-uploading to be possible, and the relevance of the "Hard Problem" of consciousness to this assessment. This background colours the discussion further on within the thesis.

How To Upload Your Mind

David Chalmers (2014) provides a recent discussion of mind-uploading in *Mind-Uploading: A Philosophical Analysis.* Chalmers offers three possible methods of attempting a mind-upload which seem plausible in outline.

The first class of mind-uploading is **gradual uploading**. This is the method Chalmers thinks is most likely to preserve both consciousness and identity. As the name suggests, this process occurs gradually. Chalmers suggests we replace, individually, each neuron with a silicon chip which is designed as a "functional isomorph" (fulfils the same role) of a regular neuron. Chalmers suggests that nanotechnology robots be inserted into our brains to do this. The robots would then 'learn' to perfectly mimic the function of a neuron. As each neuron is replicated it is destroyed and replaced. Once the procedure is complete every neuron will have been replaced by its isomorph and all conscious activity is simulated by the nanorobots. The silicon robots could then transfer what they've 'learned' by radio to a computer to achieve full computation, or the robots could just replace the brain and be left in the otherwise organic body.[1] If a gradual upload is successful you are the same person throughout the uploading process.

The second class of attempting a mind-upload is **destructive uploading**. Destructive uploading refers to a method where your brain is quickly destroyed and then copied. Chalmers suggests this could be done through serial sectioning. This involves removing our brain from our body and then slicing the brain into extremely thin layers and analysing each layer of the brain. We would create a complete map of each neuron and its location within the brain. This map would

[1] This is my gloss, but perhaps, other organic body parts such as the heart, lungs and kidney could be replaced robotically over time as these pieces deteriorated as well.

then be uploaded to a computer which was built to model both neural behaviour and our brain structure. If successful, this computer would both perfectly imitate your brain and be conscious.

The third class of mind-uploading is **non-destructive uploading**. Unlike the other two classes, this method does not alter your organic brain. Chalmers suggests some form of brain imaging could be used to copy your brain's architecture without destroying the brain. This would then be simulated by a computer. Non-destructive uploading involves survival but is accompanied by a copy of our mind. Arguably this is "duplication" rather than "uploading".

The Conditions Required for Mind-Uploading To Be Possible.

We are not just interested in whether a computer can be conscious – we are interested in whether we can survive as a computer (Chalmers, 2014). We need to identify what facts about the world must be true for mind-uploads to be possible.

Daniel Dennett (1994) in his article *The Practical Requirements for Making a Conscious Robot* notes three possible objections to strong (i.e. conscious) artificial intelligence. While Dennett was not evaluating "mind-uploading", his objections remain relevant. I have amended these objections to be applicable to mind-uploading:

1. Computers are purely physical objects and consciousness requires immaterial "mind-stuff" which we would not possess after a mind-upload.

2. Only a biological system can sustain consciousness. Mind-uploading would transfer us to a non-biological system and therefore remove consciousness.

3. Computers will always be too simple to sustain a human mind.

The first is an objection that a substance dualist might raise, whereas the second is an objection that either a dualist or physicalist could offer. I do not engage with the third objection in this thesis. If it made as an observation about human limitations, then it is possible, but I would not know. But if the objection is that computers are inherently too simple to sustain minds then it does not seem credible. There are speculative computer systems which operate using the total power of a sun (a Matrioshka brain). Would such a system really be "too simple" to simulate a human mind? If the answer is "no" then computers can at least theoretically possess minds even if we have no idea how to build one that does. Simplicity is not a philosophical objection.

Considering Dennett's list of possible objections (although he endorses none of them) I submit the following two conditions must be satisfied for us to be confident that mind-uploading is possible:

1. The purported "mind-upload" must be capable of having our exact phenomenal experience if consciousness is to be preserved. This means that some form of functionalism about the mind must be true. Identical mental states, minimally, need to multiply realisable across different physical systems.

2. Our personal identity must be preserved. We must be the same person after the mind-upload as we were before. This requires that either a psychological or closest continuer account of personal identity is true. Our personal identity cannot be tied to our status as a biological organism.

A semantic challenge can be raised as to whether preservation of phenomenal experience is actually a necessary condition for survival in a mind-uploading case. The critique is that it is only identity, rather than consciousness, that matters since I seem to persist as the same person while in non-conscious states such as when asleep or even comatose. It follows from this that consciousness is not essential to continuing as the same person. By extension a mind-upload would not need to be conscious to be me. I consider this critique ineffective. Even if this constituted survival (which I doubt), it would be survival deprived the essential part of what makes survival worthwhile. Moreover, a "zombie" mind-upload would be importantly disanalogous to ourselves. When, for example, I make a verbal report I do so accompanied by conscious experience of that verbal report. If the purported mind-upload made a verbal report without conscious experience, then it might be so different from us that we could not seriously say that the mind-upload was a person at all – let alone the same person.

Out of a commitment to naturalism, when defending the essentiality of functionalism to the possibility of successful mind-uploading, I argue in favour of the ambitious thesis that mental states are reducible to different forms of functional organisation. I do this because I am interested in showing how a physicalist can defend mind-uploading. A commitment to reductive physicalism, however, is not essential to a defence of mind-uploading and a dualist can defend a less expansive form of functionalism. A dualist can be a functionalist so long as they do not claim functionalism is a full explanation of consciousness. This dualist would then

say certain forms of functional composition are always accompanied by mental states without mental states being exhaustively described by functional composition.

The advantage of the physicalist approach I offer, is that it provides absolute certainty that consciousness is preserved in mind-uploading cases. If consciousness is nothing more than the right type of functional composition, then as long as the mind-upload preserves this then by definition consciousness is also preserved. My approach also has the benefit of not conceding ground to dualism, if like myself, the reader finds dualism unappealing.

What Are We Trying to Upload? Defining Consciousness

It is a meta-cliché to observe that saying consciousness "is hard to define" is itself a cliché. The difficulty of describing consciousness is part of why some say consciousness cannot be explained purely through physical facts. A popular definition of consciousness is the one suggested by Thomas Nagel (1974) in his famous paper, *What Is It Like to Be a Bat?* Nagel states that something is conscious if there is something **"**it is like" to be that thing. It is generally agreed that there is something it "would be like" to be a monkey, a dolphin or, indeed, a bat. Intuitions are murkier about whether there is anything "it would be like" to be less a complex organism, especially an insect, or whether these organisms are in essence biological robots. On the "what it is like" account a mind-upload procedure succeeds at preserving phenomenal experience if being that mind-upload is exactly like being me.

Nagel's description of consciousness is attractive because it does justice to its seemingly subjective and first-person character. Few doubt that bats are conscious since they are intelligent and sociable mammals, however, Nagel argues that the contour of their experience is likely alien from our own. Bats are nocturnal and 'see' mainly through echolocation. They also fly, shriek, eat insects and hang upside down. They are conscious but that consciousness is different from ours. I cannot find out what it is like to be a bat through my imagination. My imagination only tells me what it would be like for me to behave like a bat (Nagel, 1974). Nagel argues that consciousness is inaccessible from the third-person perspective. While I could be completely informed about the behaviour of a bat, understand its neurophysiology, and monitor the firing of its neurons I would remain 'in the dark' about what being a bat is like. There is an answer to the question, "What is it like to be a bat?" but it is a-priori inaccessible to us. It is the "what it is like" description of consciousness I am referring to when I use the term "phenomenal consciousness" in this thesis.

The risk facing someone contemplating whether they should try and upload their mind is that it seems impossible to know in advance whether consciousness would be preserved. Just as I will never know what it is like to be a bat, an external observer will never know if there is anything it is like to be a mind-upload. Assume that I engaged in Chalmer's "destructive uploading" scenario. My brain is sliced apart, and a simulation of my brain is created. There is now a computer simulation that claims to have my mind. If the computer is conscious then it will claim that it is conscious, but if the computer is not conscious it will still claim to be conscious since a competently designed simulation of me would say that. An observer will have no way of knowing ahead of time whether their consciousness will be preserved or whether the purported mind-uploading procedure just kills them.

I will eventually argue that consciousness is amenable to a reductive explanation, but I take Nagel's description of consciousness as accurately capturing the folk essence of what we want to preserve in a mind-upload and proceed with it for now.

The Conceptual Barrier to Knowing Whether A Physical System Is Conscious (The Hard Problem)

The Hard Problem is the problem of explaining how consciousness arises from non-conscious material. It alleges that there is a conceptual gap between a full description of any set of physical facts (neurophysiology, behaviour, oral report etc) and a description of phenomenal consciousness.[2] The Hard Problem is relevant to this thesis for two reasons:

a. The Hard Problem supports substance dualism, the truth of which I will argue cannot plausibly sit alongside mind-uploading.

b. The Hard Problem suggests that "zombies" could metaphysically exist. If zombies are possible the concern is that a mind-upload would not be conscious despite being relevantly similar to us. This concern is more acute when contemplating whether to

[2] There are, perhaps, other "Hard Problems" that make conscious states philosophically interesting which I do not have the space to address. For example, they seem to be non-spatial. Colin McGinn (1995) in *Consciousness and Space* argues that conscious states lack extension, mass and spatial properties. I cannot describe the width of a thought and cannot locate it in space except loosely as being "in my head". This separates consciousness from even unobservable things such as quarks and electrons which have spatial boundaries, albeit subject to quantum uncertainty. McGinn has argued that this highlights another aspect of the Hard Problem: how can the spatially bound describe the non-spatial?

undergo a mind-upload procedure because we have no precedent for conscious machines. Even a small risk of losing phenomenal consciousness may be unacceptably high.

The Hard Problem is contrasted with what Chalmers calls the "Easy Problems". The Easy Problems of consciousness are those likely to prove tractable to scientific explanation. Chalmers does not think solving these problems will be especially easy, but he believes that in principle they can be solved with scientific investigative methods. Examples of Easy Problems (Chalmers, 1995, p.5) include verbal report and the sleep/wakefulness distinction. To explain verbal report we need an account of the process by which information is retrieved and then made available for communication. The distinction between sleep and wakefulness can be explained by offering contrasting accounts of our neurophysiology between each state. While we do not have a full account of how either occurs, they seem like the sort of problems science can solve.

Chalmers alleges that once you explain how an Easy Problem is performed, a further Hard Problem is left unresolved: why is it accompanied by phenomenal consciousness? To take verbal report, an explanation of how we perform verbal report does not need to explain why we are conscious of giving that report. A "zombie" account is equally compatible with any scientific explanation. A reductive explanation of consciousness will therefore always fail. I could explain everything about what goes on inside someone's brain without answering the question of "what it is like" to be that person. Describing the firing of c-fibre strings that accompany pain does not actually describe the feeling of pain. This leaves an unanswered "further fact" suggesting phenomenal consciousness is unlike natural properties which are amenable to reductive explanation. Water, for example, is reducible to H_2O. While it is not immediately obvious that properties such as wetness or viscosity are captured by H_2O, once someone is informed of the relevant facts it is impossible to deny the identity without being conceptually confused. There is no "further water fact" left once you describe H_2O.

The concern is any physical explanation of phenomenal consciousness has a "zombie" counterpart. In the literature philosophical zombie versions are imagined as our exact duplicates both physically and functionally but lacking phenomenal experience. For example, we can imagine a zombie version of ourselves, who sleeps, procreates, writes books and collects Yu-Gi-Oh cards but who does this without a mind. The claim is not that zombies could exist. It is usually accepted that as a matter of nomological necessity zombies are prohibited

(i.e. the laws of nature don't permit it). What is disputed is that it is metaphysically necessary that a functional and physical duplicate must be phenomenally conscious (i.e we can conceive of a possible world where there are zombies). This is said to point to a conceptual gap between describing the physical and describing the mental.

Zombies are of interest because some real-world phenomena are sometimes taken to support their metaphysical possibility (Dennett, 1991). Blind-sighted people are cortically blind. Due to damage to their occipital lobe blind-sighted people are not able to see in part or all of their visual field. Nonetheless, they are able to respond to visual stimuli such as a flashing light or movement with much greater accuracy than chance alone would suggest. They do this while denying being conscious of either light or movement. This is sometimes taken to suggest that we can imagine people who respond as if they are conscious without being phenomenally aware.

A common physicalist answer to the problem of zombies "is the sense and reference" distinction. For example, the "Emperor of France in 1806" and the "King of Italy in 1806" have a different sense but the same reference (Napoleon). They pick out different attributes while referring to the same person. The physicalist could say that the phenomenal and the physical descriptions of the mind share the same reference but have a different sense. This is an identity claim that is true but is contingently so. It appears that it could have been the case that the Emperor of France in 1806 may not have been the King of Italy – perhaps Francis of Austria could have been instead. Likewise, it could be that "mind" and "(X) facts" (where "X" is your preferred form of physicalist reductionism) have a different sense but share the same reference. This could answer why we do not think we have fully described what is meant by "mind" when we describe the relevant facts.

Saul Kripke (1980) provided a logical argument against such replies (formal proof in the footnotes).[3] In some respects, the Hard Problem is just a compelling rhetorical expression of

[3] $(x)(y)\ [(x = y) \supset (Fx \supset Fy)$

$(x)\ \Box\ (x = x)$

$(x)(y)\ (x = y) \supset (\ [\Box\ (x = x) \supset \Box\ (x = y)$

$(x)(y)\ ((x = y) \supset \Box\ (x = y))$

Kripke's earlier argument against contingent identity from *Naming and Necessity*.[4] The argument is that all identity claims must be necessarily true and never contingently true because if "X" is identical to "Y" then if something is true of X then it must be true of Y. The argument (distilled) goes like this:

1. If X and Y are the same, then if X has a property Y will have that property also. If the mind and brain[5] are the same then if the mind has some property then the brain will have that property also.

2. Every object is necessarily identical to itself (Leibniz's law). The brain is identical to the brain and the mind is identical to the mind.

3. If the mind and brain are identical to themselves and the mind is the brain then the mind and the brain are necessarily identical.

This argument rests on premises that seem obviously true. The problem of course, is that the mind and brain do not seem identical and so if they are identical then it appears they are only contingently so (i.e in some possible world they may not have been the same). We think we can imagine minds without brains and vice versa, but if the mind and brain are identical then we shouldn't be able to do this. Kripke's argument strikes some as fatal for physicalism. It blocks the physicalist from asserting the identity between the mind and some set of facts is only contingently true, which has been a common strategy.

Kripke acknowledged that contingent identity appears common and he has an explanation for this. To establish identity between two objects which appear contingently identical you face the burden of explaining the contingency away. His justification for this is technical and rests on distinguishing **a-prioricity** from **necessity**, and **rigid** from **non-rigid designators**. If

[4] Which itself is just a very clever extension on the earlier mind-body problem.

[5] I use "brain" and "mind" here, but I do not endorse the mind-brain identity thesis. The problem is a generalised one and "brain" can be substituted in for reductive computational functionalism, although that is less pithy.

something is a-priori true, then it is something that we can find out by reason alone. If something is a-posteriori then we cannot ascertain its truth without sense experience.

Kripke explained through this distinction that the reason why identity claims are necessarily true but can seem contingent is that some identity is discovered a-posteriori. His example was that his lectern is made out of wood. To know what it is made out of we need to physically examine it which is a form of a-posteriori knowledge. Nonetheless, necessarily for the lectern to be itself it must be made out of wood otherwise it would be a different lectern. Kripke argued that necessity is often conflated with a-prioricity, but that something does not need to be a-priori true to be necessarily true.

Something rigidly designates an object if it could "not be other than". For example, the "square root of 25" rigidly designates the number "5". The square root of 25 could not designate anything else. The description "Interim President of the People's Republic of Burkina Faso" non-rigidly designates "Ibrahim Traoré" at the time of writing.[6] Proper names are rigid designators, so Ibrahim Traoere refers only to the person Ibrahim Traoere who is non-rigidly picked out by his position. We can make counter-factual statements about him such "as if Ibrahim Traoere had not couped Paul-Henri Sandaogo Damiba then he would not be interim president". Clearly, in these counter-factual statements, we are talking about the same person even though Ibrahim Traoere did something different. We therefore fix the reference of a name with some non-rigid description such as "interim President of Burkina Faso" but once the reference is fixed the name rigidly refers to that person. Even if the President of Burkina Faso was different, "Ibrahaim Traeore" would refer to the same person. In Kripke's terms, an expression is a rigid designator if it refers to the same item across all possible worlds. Non-rigid designators pick out contingent properties of an object but do not themselves establish identity.

Kripke explained that this is why the identity between "heat" and the "movement of molecules appears" to be contingent. We first knew of heat indirectly through our phenomenal experience of its sensation. We therefore fix the rigid designation of "heat" through the non-rigid designation of "the sensation of heat". Heat, however, is the kinetic motion of molecules. Even if we discovered aliens who were inert to the feeling of heat, "heat" and "movement of molecules" would still be the same thing. The "movement of molecules" is therefore a rigid designator for heat, and unlike the sensation of heat, is an essential property which heat must

[6] Given the frequency of West African coups this may change.

have in all possible worlds to be heat. Empirical investigation revealed this and so our knowledge that "heat" and "the movement of molecules" are necessarily identical but this discovery was a-posteriori. The contingency of identity here is an illusion arising from the conflation of a-posteriori knowledge and contingent identity. It appears contingently true because we think we can conceive of the movement of molecules without the contingent property of the sensation of heat, and the sensation of heat without the movement of molecules.

Kripke then extends this analysis to the mind. Pain is a phenomenal state and it has the essential property of being painful. It is not a contingent property of pain that it feels painful. Phenomenal pain is therefore a rigid designator. The necessity of identity requires that the physicalist assert that "pain" is necessarily the same as a physical description such as (for example) the "stimulation of c-fibre strings". The stimulation of c-fibre strings is also a rigid designator. The physicalist therefore needs to explain why we think we can imagine one without the other. Put another way (McLaughlin & Planer, 2014) we pick out heat contingently through the sensation of heat but we could also pick it out in other ways, but we cannot imagine an epistemic counterpart to pain that is not painful. Pain is essentially itself. The problem for the physicalist is that if you wish to assert that the mind is identical to some set of physical facts you then need to explain why this does not appear to be the case. Kripke's argument suggests that the physicalist cannot meet this burden.

2. What Do Our Intuitions About Mind-Uploading Suggest About the Mind?

In this section, I introduce dualism and physicalism. I weigh competing intuitions about whether mind-uploading appears more compatible with dualistic or physicalist approaches to the mind. I conclude that intuitions are mixed but at first glance mind-uploading is easier to reconcile with dualism. In Chapter 3 I will argue that soul theories of the mind (substance dualism) actually are incompatible with mind-uploading, but it is conceded that this is not obvious and requires argument. At the conclusion of this chapter, I explain why the only path from physicalism to mind-uploading is through a very strong form of functionalism.

Do Our Intuitions About Mind-Uploading Support or Undermine Physicalism?

Non-destructive Uploading

In the case of non-destructive uploading (or duplication if you prefer), there is a clear and strong intuition. The upload would not be 'you' since 'you' are left unaffected by the

uploading procedure. The mind would not relocate positions just because a duplicate of it had been assembled elsewhere whether or not the duplicate was conscious.

Destructive Uploading

Destructive uploading requires a period of biological death, and intuitions are more equivocal about whether upload would be you (assuming it is conscious) than they are in the case of gradual uploading. The intuition against this is that 'you' die in the procedure. If the upload is successful, then you have been resurrected inside a computer. The competing intuition in favour of the view that you do survive a destructive upload rest on the fact that the upload would be psychologically indistinguishable from you and also believe that it was you. On its own, this does not appear to be enough for destructive uploading to constitute survival because the upload also thinks that it is you in the case of a non-destructive upload.

An intuition that we survive a destructive upload seems inconsistent with the view that we are not the mind-upload in a non-destructive uploading scenario. In a non-destructive upload procedure, we have Person A and then create Person A* who is psychologically indistinguishable from Person A. In this scenario, we say that Person A ≠ Person A*. In a destructive upload procedure, we kill Person A and then create Person A* and say Person A = A*. This violates the 'necessity of identity'. It is not logically coherent to say that sometimes A = A* and sometimes A ≠ A* since personal identity seems to be a rigid designator. Without something further said, our clear intuition that the mind is not transferred in a non-destructive upload should prevail over an equivocal intuition about survival in a destructive upload.

A further challenge for the physicalist, is that on the face of it, the mind coming back from the grave and then being implanted in a new shell would appear highly dualistic. For physicalists who want to endorse destructive uploading as survival there is this further tension.

In Chapter 6 I will explain that this obstacle can be circumvented by appealing to a closest continuer account of personal identity. To foreshadow my argument, the difference between non-destructive uploading and destructive uploading is that in the latter case no one has a better claim to your identity. For the moment, this may strike the reader as unsatisfactory, and a lot more will be said further on. For now, I acknowledge that there is a burden that I am required to discharge if I am to reconcile physicalism with the claim you could survive a destructive upload.

Gradual Uploading

The appeal of using the gradual upload procedure is clear: if each neuron is replaced by a silicon isomorph then there is no moment where I appear to lose either my consciousness or my identity. This has obvious similarities with the 'Ship of Theseus' argument. If gradual uploading is possible (and our intuitions generally suggest it is) then this suggests one of two options:

1. Dualism; and/or
2. Functionalism

Gradual uploading supports dualism because the mind remains in place despite the total dislocation of the matter involved. This suggests that the mind is not matter which is precisely what dualism predicts.[7] Indeed, this would seem to hold if we did not destroy each neuron but instead moved them during the gradual upload procedure. As my neurons are slowly displaced my consciousness and mind would be preserved but ultimately none of its parts would be the same. If we reassembled my organic brain with the displaced organic neurons, we would ultimately have the same brain in a different location, but plausibly unaccompanied by the same mind. This brain could then even be placed into a new body. This seems highly dualistic.

To see how gradual uploading may support functionalism we can appeal to Chalmers' (1995) "Dancing Qualia" argument.[8] This argument is that, minimally, some form of non-reductive functionalism must be true because if we removed our neurons and replaced them slowly with functional isomorphs made of silicon (essentially a gradual uploading procedure) we would not lose our phenomenal consciousness. The reason we can be sure that we would retain our phenomenal consciousness throughout the gradual upload procedure is because the two other options are implausible. These are respectively:

a. Consciousness entirely dissipates at some arbitrary point.
b. Consciousness fades slowly.

[7] Dr Gerald Harrison drew my attention to this possibility of displacing the brain rather than destroying it during the gradual upload procedure.

[8] Qualia is a plural term for items of phenomenal consciousness. The singular term is "quale". I personally find it to be an ugly piece of jargon and try not to use it.

The first option (a) is said to be implausible because it would imply that one neuron represents the difference between consciousness and non-consciousness. This would mean when we go from 89,000,000,000 organic neurons to 88,999,999,999 organic neurons and 1 silicon neuron (or somewhere further on in the process) we suddenly become a philosophical zombie. The second option (b) is said to be implausible because it implies that although consciousness may fade this would be without a change in behaviour. Since the functional structure of the brain has been preserved, its behaviour would not be any different. The person would gradually lose consciousness but would not be able to say anything to indicate it. The logic is that phenomenal consciousness would dance as we both reverse and then reimplement the process. If we did this very quickly the metaphorical "lights" would switch on and off without any change in that person's behaviour.

The argument does not require that consciousness fades, it could be that consciousness is altered. Perhaps during the procedure the colours red and orange switch. Again, since there has been no change in functional organisation, the behaviour of the person should not change and they could not remark on the difference despite seeing opposite-ordered colours. Chalmers' argument is that because the other two options are implausible we should endorse the view that phenomenal consciousness is invariant across systems with identical functional organisation. This would support the view that a computer could share our phenomenal consciousness, although it would not show that consciousness is computation.

I have no issue with this argument when offered more modestly than it is sometimes presented. The reason why consciousness is philosophically interesting is that it does not seem to be something you should expect any physical system to have. If phenomenal consciousness is not reducible to some form of functional organisation (such as information process/computation) as Chalmers' argues, then we should be open to the possibility of something like dancing qualia, even if it would reduce consciousness to an epiphenomenon.

Since this argument can support either dualist functionalism or physicalist functionalism, I will need to say something further to reach the further conclusion that reductive functionalist account is plausible. My approach will be to suggest that consciousness is computation. Computational accounts of consciousness are functionalist because different systems can implement identical computations..

Evaluating Gradual Uploading Against Destructive Uploading

As discussed, there is a clear intuition to prefer gradual uploading to destructive uploading. One option would be for this thesis to proceed by only defending the view that gradual mind-uploading constitutes survival and conceding that destructive uploading is just an interesting method of suicide. However, I wish to make the more ambitious argument that destructive uploading is just as good as gradual uploading in this thesis. While both are outside the realm of scientific possibility, I submit that it is easier to see how we could invent a method of destructive uploading.

Serial sectioning a brain and then simulating it is something we have already done. For example, we have serial sectioned parts of a mouse brain and mapped it. We also have computationally modelled the nervous system of the nematode C.elegans. These models have successfully showed how every neuron inside C.elegans "communicates". Computers emulating the nervous system of C.elegans are even capable of autonomous motion (Laakasuo et al., 2021). Now there is obviously a big leap between simulating the 302 neurons of a nemotode and simulating the human brain, but progress is fast. In 2023 researchers managed to map the 3016 neurons and 548,000 synapses within the brain of a fruit fly larvae which suggests we are inching in the right direction (Winding, et.al, 2023). The proposal that we could serial section and then computationally model a human brain in a destructive mind-upload is therefore not totally fanciful.

By contrast, we have no idea how we could perform a gradual upload. While Chalmers (2014) refers to creating silicon isomorphs of human neurons to replace our brain cells with, we have not the faintest idea how we could devise such technology in practice. Chalmers describes this as "nanotechnology" but this is very clearly within the distant realm of science fiction. It seems reasonable conjecture to say that the technology for destructive uploading is likely to become available sooner than the technology for gradual uploading.[9] I therefore wish to make the bolder case in favour of destructive uploading, because if we are ever confronted with the option of uploading our minds it is likely to be with this intuitively less attractive option.

The Classes of Dualism and Their Compatibility with Mind Uploads

Dualism is the position that the mind and the physical are in some way distinct. It is an unpopular view amongst both neuroscientists and philosophers. The few scholars who do endorse dualism generally resile from substance dualism in favour of more scientific presenting

[9] Assuming either of them becomes available, which is far from guaranteed.

varieties. An interesting article by David Chalmers and David Bourget (2014) surveyed the views of philosophers across a range of subjects. It suggests that 56.5% of philosophers are physicalists with only 27.1% identifying as non-physicalists. Of the non-physicalists, only some small number would be substance dualists. However, the unpopularity of dualism is no evidence of its falsit and if raw popularity is the metric then substance dualism surely wins. Since most people believe in an afterlife, it follows that they believe the mind can persist without any body which entails substance dualism.

Physicalists are committed to the view that a mind is nothing more than either a physical or functional system. The difference between physicalism and dualism is that the dualist thinks our ontology needs to be expanded to include consciousness, and the physicalist believes consciousness can be included without any revision. Physicalism has been in vogue since at least the early twentieth century although the type of physicalism that is popular has shifted over that period.

Some physicalists believe our type of conscious states are restricted to our exact biological isomorph (mind-brain identity theorists) and some believe it is how a system functions (functionalists) that matters. Functionalism is only a physicalist position when accompanied by the further claim that full description of a system's functional organisations suffices to explain consciousness and some functionalists are also dualists.

There are three types of dualism (with many subcategories in each) and I discuss them in increasing order of ontological commitment. Both a predicate dualist and a property dualist can be functionalists, and accordingly both views are compatible with the possibility of successful mind-uploading.

Predicate Dualism

Predicate dualism is the weakest form of dualism. Predicate dualism is the position that mental predicates are not reducible to physical predicates. This is because mental predicates play an indispensable role in our description of the world (Robinson, 2023). Robinson Howard illustrates the position by explaining how different items within our ontology can be more or less resistant to reduction. Robinson contrasts “water” with “hurricane”. It is plausible that H_2O represents a fully reductive explanation of water. If you knew every fact about H_2O you would know everything about water. By contrast, a “hurricane” might be “multiply realisable”. You can have many different types of hurricanes and a hurricane is defined by what it does rather than what constitutes it (Robinson, 2023).

A predicate dualist would say that you would therefore not be able to fully reduce "hurricanes" to their physical characteristics because their physical characteristics can differ while still being a hurricane. Psychological states, such as fear, lust and loathing are similarly irreducible in this sense. Note, that although the predicate dualist rejects that a hurricane is analytically equivalent to any term in physics, a hurricane remains just a collection of atoms. Predicate dualists hold that this is true of the mind. I am doubtful that predicate dualism represents an actual category of dualism and it strikes me as a formulation of reductive functionalism. In this sense, predicate dualism may be appealing to those who possess a naturalistic outlook while also being willing to admit the appeal of dualism. I will briefly refer back to predicate dualism again further on..

Property Dualism

The next category is property dualism. Property dualism says there are two types of properties (the mental and the physical) but only one substance. Mental properties are not physical but they arise from the same substance as physical properties. A working definition of property dualism is a little slippery. As I understand it, property dualism is the position that mental properties are an emergent property of the same substance as physical properties. If both mental and physical properties arise from the same "stuff" there should be no barrier to generating consciousness with a computer so long as a computer can have mental properties – which property dualism is fully agnostic on.

Both the property dualist and the predicate dualist may express the relationship between the physical and the mental in terms of "supervenience." This means you cannot have a change in the mental state of a brain without an accompanying change in the physical state of the brain. The property dualist goes further than the predicate dualist and says mental states contain some literal (not semantic) further property. Note that dualists who endorse some form of supervenience may make the caveat that although there could be no change in mental states without a change in physical states, there can be a change in physical states without a change in mental states – this would allow a different physical entity (a mind-upload) to share your mental states.

Pan-psychism

A further class of view associated with dualism is pan-psychism. It is not strictly dualist because it only posits one type of substance but it says that substance always contains consciousness (or quasi-consciousness). This view is dualist in <u>essence</u> because it wants to provide a unique role for consciousness in our ontology. David Chalmers has argued that the Hard Problem can be partially resolved by adopting a view of consciousness as a fundamental

property of the universe which accompanies information. On this view, even a thermostat would have some minimal amount of consciousness (Chalmers, 1997). If consciousness is fundamental then there is no problem explaining how physical processes give rise to consciousness since consciousness is embedded in everything. A pan-psychist would usually say higher states of consciousness are associated with greater levels of complexity and that an atom would have some inconceivably small amount of consciousness.

A pan-psychist would have no trouble explaining how a computer could be conscious. Since computers are built out of physical material, and all physical material is conscious, a computer simulation of your brain could also be conscious. Indeed, all computers are already conscious in some sense. Pan-psychism, however, faces a combination problem (Chalmers, 2017) in explaining how rudimentary consciousness gives rise to sophisticated forms of consciousness in different structures. It might be that a computer simulation of your brain would be conscious but have a different type of consciousness. For pan-psychists who identify information as the source of consciousness, this would be less of a problem. Provided the mind-upload contains/processes the same information as we do, it would have our kind of consciousness.

Substance Dualism

The type of dualism I will be most concerned with is substance dualism. This view, associated with Rene Descartes, is the position that the mental is ontologically distinct from the physical. Descartes believed that mental properties exist outside the body and that the body itself does not think. To those of a religious bent, substance dualism will be attractive since it is generally seen as compatible with the mind persisting after death in an afterlife. I discuss substance dualism in greater depth Chapter 3. Arguments for substance dualism often appeal to the Hard Problem. I take substance dualism to be a soul account of identity and the mind. Our immaterial mind (soul) is what bestows both our identity and phenomenal consciousness. My argument will be that this immaterial soul does not look like it will be amenable to being uploaded.

The Classes of Physicalism and Their Compatibility with Mind-Uploading.

In this section, I briefly discuss why reductive functionalism is the only plausible pathway from physicalism to mind-uploading. I also offer a few reasons why it might be more plausible likely than its physicalists rivals, in addition to the earlier "dancing qualia" argument.

Behaviourism

One formerly popular method of explaining the mind was to describe it behaviourally. Crudely, a behaviouralist explains pain by analysing the behaviour expressed. If my body is

damaged, and I react to that stimuli - perhaps accompanied by vocalisations such as “ow” - by trying to avoid that stimuli I am experiencing the conscious state of pain. Admittedly, behaviourism is compatible with mind uploading. All a mind-upload would need to do for it to have our conscious experiences would be for it to behave indistinguishably from us. Behaviourism, however, is simply implausible and has long been discarded. While the behaviourist can offer a much more sophisticated account of pain than the one I just provided, ubiquitously their explanations are considered to suffer from a failure to capture phenomenal experience. Behavioural explanations miss the obvious fact that pain hurts! Moreover, I can be in pain without exhibiting any behaviour at all, such as when I put on a ‘brave face’. The painfulness of pain is what needs described. Nothing about the behaviouralist explanation includes this and it is vulnerable to the “zombie” reply.

Mind-brain Identity Theory

The mind-brain identity theory is the view that conscious states are just neural events and so the mind and brain are identical. Every conscious state is entirely reducible to some set of complex neural events in the brain. Note that this view goes beyond making the obvious point that conscious states have neural correlates. It is not disputed that conscious states are accompanied by neural activity, what is contentious is the claim that consciousness is nothing more than neural activity. The mind-brain identity theory works by appealing to Ockham’s razor. Crudely, phenomenal pain has a neural correlate in c-fiber stimulation. Phenomenal pain is always accompanied by the firing of c-fibre strings. Since the two always accompany one another through an appeal to Ockham’s razor the mind-brain identity theorist says we can eliminate the view that “pain” and the “firing of c-fibre strings” possess a different identity.

The mind-brain identity theorist will have doubts that a functionally invariant but distinct system could be conscious. If conscious states are just neural states, then mind-uploading will not be possible due to the absence of neurons. The radical procedures of destructive and non-destructive uploading are therefore definitively precluded. The mind-brain identity theorist might not rule out gradual uploading, however. It is an empirical question whether silicon neurons (as offered by Chalmers’) could do the exact same thing as carbon neurons. I have no idea, but silicon is often offered as a periodic element that could form the basis of alien lifeforms since it does much of what carbon does. As Pigliucci (2014) notes, however, not just any chemical will suffice. Growth, synapse production and response to electrical and chemical discharge are vital parts of the role of neurons and obviously no one has yet designed an isomorph to a neuron.

Reductive Functionalism

As has already been alluded to, functionalism is the view that explaining mental states involves explaining their function (both internally and externally) rather than their constitution. A ladder is a functional system. You can compose a ladder out of wood, rubber, metal or plastic and still have a ladder. To say that consciousness is functional is to say that you do not need to be a biological system made out of carbon to be phenomenally conscious, so long as the components function, interact and are organised in the right way. To explain a mental state you therefore explain the role it has in the cognitive system it is situated within.

Non-reductive functionalism is endorsed by Chalmers and others and can sit alongside some other theories of the mind. It is the view that the correct kind of functional organisation is a sufficient condition for consciousness, but that functional organisation does <u>not explain</u> phenomenal consciousness (see my earlier chapter on the Hard Problem). Then there is <u>reductive</u> functionalism, which was endorsed by writers such as Daniel Dennett, which states that functional organisation <u>is</u> phenomenal consciousness. Both non-reductive and reductive functionalism guarantee our minds can be copied/simulated so long as the computer has the right form of functional organisation. Reductive functionalism allows for mind-uploading so long as functional organisation is invariant between us and the mind-upload. The fact it is proffered as a reductive explanation is what enables the physicalist to chart a course to mind-uploading.

This only matters if reductive functionalism is actually more plausible than its contenders. A reason to endorse functionalism against the mind-brain identity theory as a physicalist is Hilary Putnam's *multiple realisability* argument. Putnam (1960) argued that mental states cannot be reduced to physical states in the brain because mental states can be "realised" out of many different physical states. A bat and I have very different neural architecture but seemingly both myself and the bat are capable of experiencing pain. Plausibly all manner of animals, even those with very different kinds of brain, experience pain. Putnam's example is that of an octopus. Octopuses, and other cephalopods, are unquestionably intelligent but their evolutionary history departed from our own a long time ago. Our last common ancestor, the flatworm, lived 750 million years ago. Octopus brains are both large and well-studied and demonstrably different from our own. Some of these differences include larger neurons which lack myelination and more diffuse spatial distribution; two-thirds of octopus neurons are found in the tentacles. Plausibly, however, octopuses experience phenomenal pain not dissimilar to

our own. They may even experience complex emotional states such as fear and stress. If octopuses can have pain which is like our own, with very different brain composition, this supports the view that consciousness is functional.

I cannot demonstrate that the octopus feels my kind of pain if my interlocutor insists on a first-person point of view, but from a third-person perspective, octopus pain seems plausible. Octopuses are capable of avoidance learning and react to noxious stimuli. They can even make motivational trade-offs. Octopuses predate on hermit crabs which sometimes place stinging sea anemones on their shells for protection. In response, octopuses use less efficient hunting methods such as blowing water at the anemone and moving underneath it to try and get at the hermit crab. When they encounter a hermit crab without this protection the octopus is much more direct. All that can be said in reply to a denial of octopus pain is that no one can establish anyone else has phenomenal consciousness from the first-person point of view. It is taken for granted that others do have consciousness because a third-person perspective is usually satisfactory. I consider multiple realisability a plausible but not established claim. I will explain why I consider computational functionalism particularly promising further on.

Since an octopus is an animal and animals belong to the class of entities which are understood as having consciousness, we are less worried about the possibility of octopus zombies. Minds simulated through computers would be novel and computers are not typically conscious. However, if octopuses can have the same class of phenomenal experience as ourselves, that gives us reason to think functionalism is true. If functionalism is true, we are a step closer to showing that mind-uploads are possible.

3. Why a Computer (Probably) Could Not Have a Soul.

This section is an examination of whether we can upload the "soul" and not just simulate a brain with a computer. I take it that the view that machines are soulless (and humans are not) does a lot to animate objections to mind-uploading in the popular imagination.

The view that 'becoming a robot' would render us soulless is entrenched in pop culture. In the Paradox Game "Stellaris" your interstellar empire can choose to 'ascend' by engaging in a species-wide mind-upload. This results in your species bodies and brains being replaced by machines. Your species acts as it did before, with the added benefit of immortality, but the response from religious empires is to send you a message stating your species has committed

suicide. The idea is that your ascended robot species mimics your prior alien race without being your alien race.

In this section, I explain why substance dualism is only compatible with mind-uploading if the substance dualist is prepared to make some damning concessions. To achieve this, I appeal to an argument by Richard Swinburne who argues that partial brain transplants support substance dualism. I will argue that if Swinburne is correct that partial brain transplants support substance dualism then from his argument mind-uploading is false.

Why Richard Swinburne's "Partial Brain-Transplant" Argument for Substance Dualism Suggests Mind-Uploading Is Not Possible

Richard Swinburne (2018) in his article, *The Argument to the Soul from Partial Brain Transplants,* makes the following argument in favour of substance dualism. This argument is relevant because it also presents a challenge to the concept of mind-uploading. The purpose of this subchapter is not to undermine substance dualism nor is it to assert that a substance dualist is required to endorse mind-uploading. Instead, I am trying to explain why a substance dualist who wants to maintain their substance dualism will find it easier to deny that mind-uploading is possible than to endorse it.

The hemispheres of the brain are divided in the middle by the corpus callosum. Swinburne states we could sever the corpus callosum entirely and then transfer each hemisphere of the brain and place it into a different brainless body. Swinburne's argument is this: if I sever my brain in two and place the right hemisphere into one body and the left hemisphere in another body, then it is logically possible that "I" go with the left hemisphere, and logically possible that "I" go with the right hemisphere. Swinburne states both bodies would be psychologically identical and have equal claim to my personal identity. If they were placed in identical bodies they would be biologically identical. Swinburne concludes from this that there is no fact about bodily or psychological continuity sufficient to resolve the issue of what constitutes personal identity.

Swinburne's argument is that there must therefore be a "further fact" that establishes which hemisphere (if any) is "me" and that the answer is that possession of an immortal soul (mind) is what establishes personal identity. I am me over time by virtue of possessing the same soul. In Swinburne's words (2018, p.17):

> since they have the same body and so brain and exactly the same physical and mental properties… made of the same component parts, in the same way. So it could only be

> that one of two different persons could result from the operation if it had a part which the other lacks… they must different in the respect that one of them has a non-physical part which the other lacks, which I will call a "soul"… Only someone who has my soul will be me.

The reason why this is relevant to the prospect of mind-uploading is that the same intuition that underlines the view that there is no sufficient biological or psychological fact to establish continuing personal identity will also show that there is no set of biological or psychological facts sufficient for us to survive a mind-upload. Substance dualism is therefore incompatible with mind-uploading unless our soul can be transferred.

A substance dualist who wanted to reassure a person contemplating mind-uploading that their mind/soul would be preserved could naively suggest that we ask the mind-upload of someone else who had undergone the procedure whether they were conscious and still the same person. This is, by common consensus, the only way we know other people are conscious in day-to-day life. This strikes me as unsatisfying. If a purported simulation of my mind was created then of course it would state that it is me. This would not clarify whether or not the mind-upload was a zombie.

Moreover, on substance dualism, even if the mind-upload did have a soul it may not be my soul. In Swinburne's thought experiment he answers the question of 'which, if any, transplant would be me?' by replying 'whichever has your soul'. Perhaps the other transplant (in Swinburne's worldview) now has a newly created soul. An external observer would not be able to tell which soul had gone where and I do not even see a way the transplant patients would know themselves.

I will now explain why Swinburne's argument for substance dualism undermines the possibility of mind-uploading. Consider **non-destructive mind-uploading**. In this instance, my brain is preserved but an uploaded duplicate is made. If the duplicate is conscious then according to substance dualism the duplicate must have an "immaterial" soul. Since "my" body is unaffected presumably "I" keep the same soul. The alternative view is that my soul has left my body and entered into the computer. This would mean my former body is either a zombie or has a new soul. I presume it is agreed that this is not plausible and there is no reason to believe this would happen. Most people's intuition would be that the soul remains with the body in a non-destructive upload. I assume Swinburne would see no problem with this.

Problems arise when we consider **destructive uploading**. To reiterate this process would involve my brain being sliced into very thin layers with each layer being scanned and then uploaded. In this case, a period of biological death precedes the mind upload.[10] Like before, the upload is either conscious and has an "immaterial part" or the upload does not have an immaterial part and is a zombie. This immaterial part can either be my soul or a new soul. For mind-uploading to succeed this soul must be my soul.

To explain why the upload is unlikely to contain my soul - if substance dualism is true - assume that after the mind-upload my body is very carefully reassembled using futuristic technology. All the same physical parts that my brain was constituted of before the procedure are carefully put back into the same place down to the atom. My brain is then carefully placed back into its body which had been preserved during the procedure. "I" then wake up with no memory of what has happened since. Is this me? If mind-uploading is possible and substance dualism is true, then there are two options. Either my soul previously moved into the upload and then returned to my body or my soul had moved into the computer and remains in the computer. I submit that neither view is attractive to the substance dualist.

The view that my mind was uploaded but then suddenly returned once my body was reassembled is absurd. Since the mind and not the brain does the thinking then once my soul returns I should expect to have all the memories of my time on the computer. This cannot be true though. Since my brain is reassembled as it was before the upload, my memories from my time as a computer will not be reflected in the neural structure of my brain. This causes an obvious interaction problem. How could I have these memories without these memories being physically stored as information in my brain? It is also inconsistent with our former intuitions about **non-destructive uploading**. In the case of non-destructive uploading, we presume that the soul would not shift locations just because a duplicate of the brain is built. Assuming the soul shifts locations now would therefore be inconsistent and arbitrary.

The second option is that the soul moves during the destructive upload procedure but does not shift locations when my body is restored. This is also implausible. From the perspective of my revived body, he has just "woken up" and must have a new soul. My revived body would be

[10] If dead, you may also go to the afterlife if you also accept Swinburne's theism. This would imply you are clawed back from the Pearly Gates if uploaded. A surprising result. I don't adopt this line of analysis because I wish to avoid critiquing substance dualism through the backdoor of theism and one belief does not actually require you to hold the other – though they do seem to be a common pairing.

left entirely mistaken about his personal identity. Yet he would adamantly claim he was me, and there would be no good basis to refute him since everything appears the same as it was before the upload procedure. Moreover, we would intuitively be very inclined to prefer the "waking up account" to the "new soul" account if we undid the destruction of the body sufficiently quickly. It therefore seems the cleanest position is for the substance dualist to deny that destructive uploading would succeed in transferring the soul.

A substance dualist might agree with this, but then argue that **gradual uploading** does not involve a change or loss of soul, and this is a bit more reasonable. Since gradual uploading does not involve a clear moment of death, or the creation of a copy of our mind, it is easier to assert that the same soul persists throughout the entire process. Nonetheless, the only distinction between gradual uploading and destructive uploading is speed and method. This does not seem like a comfortable basis to assert that the soul would be preserved and then transferred to a digital medium. If after the gradual upload is complete the result is "radioed" to a computer as Chalmers suggests, the end state is the same as with a destructive upload.

Moreover, substance dualists are typically interactionists and believe the mind is a soul sustained or interacts with the brain. Once the brain is gone the mind may follow. Possibly the substance dualist could say some other non-biological material could sustain a mind, but the substance dualist says "mind stuff" is a special substance currently only found in organic systems. Say we go through a gradual mind-uploading process but then a third party switches the silicon neurons running our new brain off (perhaps by interfering with their power source somehow) consciousness would presumably cease. If this person walks away and never comes back the mind-uploaded person is dead, but if the third party came back and then turned the silicon neurons back on, the mind-upload is not dead. This seems to trap the mind-upload's soul in an eternal state of un-death. A third party could even continuously flick the silicon neurons and off, or even change (and then reverse) the mind being simulated by replacing each silicon neuron with a copy of someone else's neurons. The result is that not dissimilar to Chalmers' "dancing qualia" we get "dancing souls" with different souls vanishing and then reattaching to the same object. This is a very strange conclusion and I take it is as a set of claims the substance dualist is unlikely to want to endorse.

The only reply, on behalf of the substance dualist, I can offer is to analogise flicking the switch to reviving a person who is biologically dead. The substance dualist could reasonably say that reviving a person in those circumstance reattaches the soul to that body (or never left). The

distinction is that the same body can only be revived from so much currently. A mind-upload could be switched off" for centuries. It seems an enormous stretch to say that the same immaterial mind waits in limbo for centuries for the on switch to be pushed. The physicalism I am defending says that all the mind happens to be is the right form of organisation, so it is able to explain what has happened more neatly by simply saying functional organisation has been preserved. The substance dualist needs to offer a believable explanation of what is happening to the soul throughout this time.

To be clear, none of this rebuts Swinburne's argument nor is it intended to. I leave a rebuttal of substance dualism for the next chapter, and further critical analysis for the section on personal identity.

4. Substance Dualism Is Incompatible with The Phenomenon of Split-Brains.

My preferred critique of substance dualism appeals to the phenomenon of split-brains. This is not the only way of critiquing substance dualism, but I select it because it thematically coheres with the rest of the thesis. Split-brains are often taken as a "Hard Case" for immaterialists (Schechter, 2018, p.15). They are particularly interesting in this context since Swinburne relies on partial brain transplants to justify substance dualism. It is therefore dialectically interesting to reply by showing how a partially severed brain, inside a single body, undermines it.

The argument I will present in this chapter is that substance dualism is not well-equipped to explain the phenomenon of split-brains which has been extensively corroborated and studied. This argument will be evidential rather than logical. It is not that it is impossible to reconcile split-brains with substance dualism but rather through inference to the best explanation the split-brain phenomenon suggests we should prefer physicalism to substance dualism. I am not convinced substance dualism can be disproved in any stronger sense than by bringing out its implausibility and then appealing to parsimony. It posits something immaterial, invisible and indetectable with scientific investigative methods. This seems quite difficult to falsify.

A Brief Explanation of The Split-Brain Phenomenon

The term "split-brain" is a colloquialism for a form of surgical treatment where the corpus callosum is partially cut in two. This was historically a treatment for severe epilepsy. The corpus callosum connects the hemispheres of the brain and relays information between each of the two hemispheres which themselves perform different tasks. The hemispheres

themselves have contralateral executive control over the body. So the right hemisphere typically controls the left side of the body and the left hemisphere the right side.

Despite the radical-sounding nature of the procedure, split-brain patients experience only a few problems day-to-day. Most continue to live regular lives and hold down jobs although in experimental contexts a wide array of deficits become apparent. What is striking is that some of these experiments give the strong impression that within a split-brain patient, there are two subjects of experience. The impression conveyed is strong but difficult to reconcile with the fact that otherwise split-brain subjects present as a unified person (Schechter, 2018, p.72). Exactly what these experiments prove is the subject of debate and the procedure is now very uncommon.

To explain how the intuition of parallel streams of consciousness arises examples need to be given. Experiments that bring about this intuition include the cross-comparison test. Schechter (2018) provides the following idealised example. In the cross-comparison test, an object is placed in each hand of the split-brain patient. For example, a pipe in the left hand and a pen in the right hand. You then test whether each object has been recognised. You then take away these objects and have them select the object that compliments the one they were holding from a series of objects (pencil for the pen, a lighter for a pipe etc) while remaining blindfolded. The right hand selects the pencil and then the left hand selects lighter. It would appear the subject has recalled and identified each object. Say you then take the lighter and pencil away and then present the original two objects in the opposite hands and ask if they are of the same or different kinds. A split-brain subject cannot do this. What they can do, however, is a comparison where they hold both the pen and pipe in a single hand then they can then easily indicate that both are different objects (Schechter, 2018, p.30). A split-brain subject is able to give these different answers simultaneously. They are also capable of drawing different objects at the same time in a way most find extremely challenging.

Another well-reported example is where a compound word whose components are not semantically related to the compound word is shown to each half of the visual field. A patient is shown the word “keyring” in such a that only the word “key” is received by the left visual field, and the word “ring” is received by the right visual field (Bayne, 2008, p. 279) (Schechter, 2018, p. 12). Visual inputs received by the right hemisphere of the brain are not usually able to be verbally reported and visual inputs received by the left hemisphere are not able to be expressed using the right hand. If you ask the patient to describe what they see they may say

"key" but point to a picture of a "ring". The split-brain patient is not aware of the object the "keyring" but instead each hemisphere of the brain seems to have interpreted the instruction differently.

Interestingly, in one experiment a split-brain patient was asked to draw what he had seen using his left hand. When the split-brain patient was allowed to watch themselves draw so that both the left hemisphere and the right hemisphere could provide direction they sometimes drew both components of the compound word but never drew the full compound word (Schechter, 2018, p.30). This was possible because the left hand is mostly but not exclusively controlled by the right hemisphere of the brain and so both were able to give instruction.

The suspicion therefore arises that there is a distinct subject of experience within each respective hemisphere. Supporting this, split-brain patients even report inconsistent preferences. A split-brain subject can be asked what their favourite colour is and say and write two different answers.[11] If you are getting two different subjective preferences, from each of the two hemispheres which are acting independently, then it seems that there are two minds present. If this is so, then the surgery has either unshackled a second subjugated person who was always present, or alternatively, the surgery created the second perspective because the two hemispheres can no longer communicate and bridge their differences. If each hemisphere has its own viewpoint, and those viewpoints do not form a common whole, then we might say that there are two minds rather than one. The claim that there are two minds, however, is belied by the fact the split-brain patient otherwise presents like one person. While much hay is made of the interesting differences between the hemispheres in day-to-day life they act with common purpose. Different philosophers assign different weights to each intuition.

There are further instances which give rise to the intuition of dual consciousness too. One example is **alien hand syndrome** where a person's hand operates independently of the subject's reported conscious control. For example, you may attempt to put on your pants and find that the alien hand is trying to pull them off. This is a symptom that some people who have had split-brain operations experience and also sometimes occurs in people whose corpus callosum has been interfered with in other ways such as through a tumour or stroke (Hassan &

[11] If the reader is interested, you can find snippets of some of these experiments contained within the very brief YouTube video You Are Two (youtube.com)

Josephs, 2016). It is easy to attribute the "alien hand" to an "alien mind" who shares the body with the person experiencing the syndrome.

Why The Split-Brain Phenomenon Is Difficult to Reconcile with Soul Theories of The Mind

I take the following to be an apt description of what it means for the mind to exist distinct from the body (McMahan, 2002, p.21):

> If the soul is understood as the subject of consciousness, its boundaries are determined by what it is conscious of. All conscious events occurring simultaneously in a single soul must be co-conscious. If, for example, my soul is the substance coextensive with this field of consciousness, then any conscious events that are occurring now that are not within this field—any conscious events of which I am not now conscious—must be events within a different field of consciousness, a different soul. …a single soul cannot have a divided consciousness.

Split-brains seem to result in a bifurcation of consciousness with two "streams of thought" coexisting inside a single body. Subjective consciousness requires a soul. This would suggest each hemisphere has a soul. One answer is to deny that both hemispheres are the subject of consciousness even though both appears to report experience. Traditionally this is the "right" hemisphere because the left hemisphere has greater say over verbal report. Therefore, although the "right hemisphere" seems conscious it is actually a zombie. This is hard to disprove, but it sits uncomfortably alongside the fact that the hemisphere exhibits all the normal indicia of phenomenal consciousness. All the things that cause us to attribute consciousness to each other are possessed by both hemispheres. Each hemisphere can make observations, offer preferences and even act in conflict with the other.

The substance dualist can say that the procedure has created a second soul inside a split-brain patient but this seems absurd. There seems to be no good reason why this procedure should have created a second immaterial mind. It would also make reunifying the two hemispheres murder which seems to be an odd conclusion.[12] Likely, some split-brain patients would prefer to have a "normal" brain if it were not for the risk of other medical complications. The split-brain subject also appears to have unified phenomenal consciousness in non-experimental contexts and so on balance, seem like one person and so attribution of one soul is more

[12] Well, perhaps that is because "lefty" does not care about "righty", and since "lefty" controls verbal report "righty" cannot plead for his/her life, but I digress.

plausible. The “one soul” answer seems to be the most common approach for the substance dualist (Hershenov and Taylor, 2014).

However, saying there is only one soul in the split-brain patient makes no sense either. If the mind is the source of consciousness, then consciousness ought to be unified. You cannot have a conscious experience that you are not conscious of. Bifurcated conscious experiences are happening simultaneously so the problem is not that the split-brain patient is forgetting its experiences, or behaving inconsistently but rather that it is having a conscious experience it is not conscious of.

Each answer the substance dualist offers has problem. If there is only one soul within the split-brain subject, then why does consciousness seem bifurcated experimentally? Since the soul's boundaries are determined by what it is conscious of, then it should be conscious of all its conscious experiences. Moreover, one mind doing two simultaneously contradictory things or offering two incompatible preferences would be unlikely. But again, if the substance dualist instead says there are two souls, then why do both souls otherwise seem to form an integrated person outside of experimental contexts? The substance dualist therefore is caught in dillema, attributing either one soul or two souls to the split-brain subject is implausible. By extension, substance dualism is implausible.

One resolution to the problem of split-brains is to concede bifurcated consciousness while defending the unity of the split-brain patient as a single person (Shechter, 2018). It might be that bifurcated consciousness is too ephemeral for us to say there are two persons inside a split-brain subject. But this option is only open to the physicalist. The physicalist does not believe in souls and so does not believe souls bestow identity. As Swinburne noted, souls for the substance dualist are the source of identity as well as and consciousness. You do not have consciousness without a soul, and without a soul you do not have an identity. If there are two subjects of experience, there must be two minds according to the substance dualist. The physicalist has the option of appealing to an alternative account of identity which allows them to defend the unity of the split-brain subject as a person.

Hershenov and Taylor (2014), on behalf of the soul theory endeavour to reply to the threat posed by split-brains. They start by reminding us that the substance dualist says the mind and brain interact, and so the mind being affected by a brain operation is to be expected (2014, p.6):

> Let us first point out that emergent dualists such as Hasker (1999) and Zimmerman (2010), as well as those who believe souls are divinely paired up with bodies like

> Plantinga (2007), or those who accept Unger's (2006) dispositionally paired soul, all posit a dependence of thought on the brain. Zimmerman goes so far as to say that "All contemporary dualists (among philosophers, at least) admit that the ability to think depends on a properly functioning brain" (Zimmerman (2010), 135). The emergentists argue that consciousness arises whenever the brain reaches a certain threshold level of organizational complexity. In less complex states, matter exhibits no consciousness. But when properly organized in the brain, it gives rise to the conscious soul.

The way Hershenov and Taylor try and circumvent the challenge is by arguing that although thought may be bifurcated, the same soul can be aware of both streams of thought. Hershenov and Taylor's argument is that the soul can have access to mental states "A" and mental state "B" but that mental states "A" and "B" may not have access to another ("C".) The soul can only reflect on "A" and "B" or "C" when there are physical connections in the brain which enable it to do so. Hershenov and Taylor argue that the soul can only think the thoughts produced in the brain so it cannot produce "C" by combining "A" and "B". Since it sits above each mental state we do not need to posit two different minds/souls to explain split-brains (2014, p.12):

> If such causal ties between A, B and C are not in the brain, then there won't be causal relations between mental states A*, B* and C* in the soul either. A will produce A* in the soul, and B will produce B* in the same soul, but mental states B* and A* will not interact if physical states B and A don't. B* will not influence or be about A*, or vice versa, nor will thought C* be about A* and B*. Rather A* and B* will be as isolated from each other in the soul and A and B are in the brains.

I agree that this can show that split-brains are not *logically* incompatible with substance dualism but that does not make their explanation very convincing either. On their account, the immaterial soul now sits as an explanatorily redundant add-on to the brain thinking thoughts that seem to have already occurred physically in the brain. Hershenov and Taylor by acknowledging the soul functions through the brain accommodate very obvious scientific realities such as brain damage, but the problem is not that the split-brain patient exhibits impaired behaviour, but rather that it has conscious experiences it is not conscious of. If the soul is the source of consciousness, then the soul should be unified even if executive function is significantly impaired.

Moreover, interactionist accounts such as those offered by Hershenov and Taylor seem to make the soul otiose. If the soul needs brain tissue to work properly then why suggest the soul is needed at all? Consider my contrary account that omits reference to souls:

> The split-brain patient is one person who experiences mental states "A" and "B" but mental states "A" and "B" cannot be combined unless they are physically connected in some way inside the brain.

Alternatively:

> There are two subjects of experience inside one organism who each respectively experience mental states "A" and "B". Since mental states "A" and "B" are experienced by different persons a combined mental state "C" is inaccessible to the organism.

My first redescription says that there is one person inside the split-brain patient and my second redescription says there are two. Which interpretation is preferred by the reader will just depend on your intuition about whether personhood as well as consciousness is bifurcated. Regardless, my explanation entirely removes reference to a soul without loss of explanation.

Hershenov and Taylor's account also seems contrived. If the soul is aware of both mental states "A" and "B" then why can't the soul put them together – an immaterial mind should not require physical processes to reflect on its immaterial thoughts. The soul is doing no explanatory work and on Hershenov and Taylor's account is in fact something in need of explaining. The substance dualist makes a very strong claim about the types of entities that we ought to include in our ontology. In this regard, substance dualism is more radical than its competitors such as property dualism and pan-psychism which say there is just one "substance". Ordinary parsimony considerations in theory selection suggest that we should prefer more modest revisions to our ontology than extravagant ones, and fewer entities to more.

Undoubtedly the physicalist also needs to explain the split-brain phenomenon, but the burden the physicalist faces in doing so is much smaller. If the mind is in some sense just the workings of the brain, then significant disruption to the brain is an unsurprising result of severing the corpus callosum. The physicalist can simply say that sometimes the hemispheres communicate and sometimes they do not and this is why consciousness bifurcates. It seems a good starting position that every physical effect has a sufficient physical cause, and to my knowledge, no one has ever proved otherwise. If this is the case – there is no need to posit an immaterial mind as a further cause of the split-brain phenomena because in doing so it seems that the relationship

between the mental and the physical becomes overdetermined (Robb, Heil & Gibb, 2023). If we have a sufficient physical cause to explain the split-brain phenomenon then there is no need to offer a nugatory mental cause.

5. Phenomenal Consciousness Is Substrate Independent and This Supports The Possibility of Successful Mind-Uploading

In this section I tentatively defend computational functionalism against its rivals; a view that naturally lends itself to the possibility of mind-uploading. I offer computational functionalism as a provisional explanation of what sort of functional system might be required to explain consciousness. Computational functionalism is advanced as a solution, partly because it thematically coheres with the topic of mind-uploading, and partly because there does not appear to be a better candidate the physicalist can point to. If some form of computational functionalism is true, then we can have a high degree of confidence that our phenomenal consciousness could be generated by a computer.

Defending Computational Functionalism

In the following section I lay out some plausible reasons why the mind might be software on the brain's "wetware". These are intended as general remarks designed to enhance the overall plausibility of the mind being functionally realisable over a variety of physical states. They are not intended to establish this conclusively. This is a topic requiring some expertise and I approach it from the vantage of an interested outsider.

In my discussion of how the mind might be computation, I am going to make some simplifying assumptions in the interests of brevity and clarity.[13] When I write of formal systems this is a rudimentary model. The mind (if it is a formal system) will be comprised of numerous highly complex formal systems, not just one. A formal system is a functional system. Chessboards, for example, can be realised with many different materials and using many different states (Carter, 2007). If the mind was a formal system, it would be realisable across a variety of physical states.

A formal system needs the following things (Carter, 2007).

1. Rules.

[13] See *Minds and Computers* (Carter, 2007) for an account of how basic human activities could be performed by computational.

2. An initial state.
3. States.
4. Entities

The states govern the entities in question, and the rules specify how we move from one state to another. If a task can be achieved by formally specified instructions in finite time, then it is **effective**. We need rules to generate different states. A formal system can be devised using only symbols, but symbols are not necessary. A chessboard has an obvious initial state. There is a set of rules that apply to it over a finite number of entities. While a Chessboard has a finite set of states – infinite states are possible for other formal systems. However, not all states can be realized according to some rules. In a formal system, a state becomes terminal if there are no further rules which apply. Formality is generally considered a constraint on computation (Chalmers, 2000).

The quintessential formal system is the Turing machine. Turing machines are composed of an infinite string of tape which is divided into segments. The segments may have symbols. The write-read head moves from segment to segment following a set of instructions telling it when to write a symbol in, erase it or leave it alone. Turing showed that such a machine can solve calculations and logic problems. The idea for a Turing machine is a precursor to the contemporary digital computer where the tape is "memory", the symbols are the "data" and the write-head is the "central processing unit" (Katz, 2008).

The advantage of such a model is it can explain what Fodor (1975) (Rescorla, 2020) calls "productivity". Loosely, mental processes preserve the truth. If you start out with a true premise and reflect on it you will generally come to a true conclusion. While undoubtedly humans make errors, our true beliefs about the world dwarf our false beliefs. Any person who had mostly false beliefs would not last long. The human mind then, like a Turing machine can reliably transform true input symbols into true output symbols. This led Fodor to conclude there is a mental system of representation rather similar to language performing the role of the symbols in the Turing machine. There are atomic components (rather like words) which can be arranged to form complicated modes of representations (like sentences). This language-like structure, implemented in a Turing-like system, helps explain the rationality of thought. We follow some set of rules in moving from one thought to another which explains why we generally hold true beliefs about the world.

Computation can also explain the systemic performance of the mind. For example, while we have only a finite number of thoughts across a lifetime, we could theoretically think of an infinite number of thoughts (Rescorla, 2020). Consider:

> Napoleon had a son who became Napoleon II
>
> Napoleon II had a son who became Napoleon III
>
> Napoleon had a son who became Napoleon IV…

And so on until the universe's heat death. We have the ability to entertain infinite thoughts but our capacity to do so is bounded by time, attention and memory. The reason why this is taken to support the computational theory of the mind is this. By implementing a set of rules (perhaps recursively) we explain how we can have infinite numbers of potential thoughts from a finite base. For example, a computer could simply perform a simple "add one" operation and perform a theoretically infinite number of calculations although it could never finish it (Rescorla, 2023).

The way the idea is normally expressed is through the ability to reconstruct sentences in different ways without any diminishment in understanding. If I can understand the sentence "Napoleon defeated General Wellington at the Battle of Waterloo" I can understand, "General Wellington defeated Napoleon at Waterloo". General Wellington and Napoleon make the same contribution to the sentence although they pick out different people. Your internal system of mental representation picks out each of the component parts and so re-ordering them will not diminish your understanding, although it gives you the precisely opposite meaning.

From this, we can roughly understand how a formal system might instantiate a mind. All computers have if-then primitives (Dennett, 1991). Obviously, a slightly contrived explanation can be crafted which explains what humans do in terms of IF-THEN functions: "if you are hungry, then seek food" and this can occur at a variety of different levels. There would therefore be meta-computations that break ties and institute weighting for a variety of commands and actions for example [[[If you do not have an opportunity to reproduce then [If there is no danger then [If you are hungry then seek food]]]

The brain is different from a computer following IF-THEN commands in a number of important ways. For example, the brain clearly implements many commands parallel to one another (Dennett, 1991) (Dennett, 2017). There would not just be a single computation but many higher-order and lower-order computations. These occur inside the brain at a much slower rate than a desktop computer performs its computations but speed is dwarfed by volume.

The result of a computational account is that consciousness is diffuse. One way of expressing what this looks like is the "Pandemonium architecture" metaphor. The Pandemonium model is slightly dated but still elucidates how consciousness could be decentralised across the brain while being performed by something resembling a formal system. Disagreements are resolved by principles of conflict resolution which include weighting, specificity, and degree of match. Each layer of the model is constituted by a discrete cluster of neurons called "demons". Here is how it would work in the case of visual identification (Dennett, 1997).

1. Image demon – A single demon records the image received into the retina.

2. Feature demons – Unlike the single image demon there are many feature demons. Each of these responds to particular features (legs, arms, a UFO) and yells more loudly the more their feature is detected.

3. Cognitive demons – They hear the yelling from the feature demons. The cognitive demons get more or less excited depending on how many features they hear that respond to their pattern. If it "hears" enough legs, arms, eyes, and bipedal feature demons screaming then the "human" cognitive demon might get really excited. The "tree" cognitive demon is likely to remain nearly silent.

4. Decision demon – Hears the yelling of the cognitive demons and then selects the loudest.

This metaphor for the mind can be supplemented by Global Workspace Theory ("GWT"). GWT suggests that the brain is divided into different specialized modules for different functions (VanRullen & Kanai, 2021). Each specialized area has long-distance connections to each other. When a process requires attention this information is shared across the distinct modules. The most important information conveyed by modules is what gains control over the global workspace. This is distributed to the other modules. The shared information in the "global workspace" is what constitutes conscious awareness. Information can be accessed if another system inside the workspace can use that information in its computations/processing. Not all neural states are accessible to the workspace, and these are the ones that fall outside of consciousness.

I should caveat my analysis by noting that there has been a move away from formal systems towards connectionist models. Connectionism models the brain using units linked together with a measurement of the strength of connections between units and a system of weighting and omits reference to symbols. A very basic model contains input units, hidden units and output units. Cameron and Garson (2019) provide an example of the human nervous system. The inputs would be sensory neurons, the output units would be the motor neurons and the hidden units would be the remaining neurons. The hidden units serve to mediate between the inputs and outputs (Carter, 2007).

Artificial neural networks can resemble neural maps of the brain and neural networks do not necessarily fall vulnerable to critiques of traditional computational models of the mind. It is worth underscoring that connectionist models are sometimes just construed as another implementation of a formal system (Rescorla, 2023). Whether this is so, is not something I well placed but to answer. An answer to whether connectionism or formal systems better model the mind is not required for my argument to proceed. Both suggest the mind is an information-processing system and information appears to be "substate independent".

The exact method of computation the brain uses is not relevant, and my prior remarks are mostly to give the argument for more reductive functionalism some relevant colour. The brain clearly processes information and information is substrate-independent. The exact process the brain uses to do this is less significant than the fact it does. It does not matter whether the information conveyed by "cat" is written, spoken, thought or contained as an adorable mammal. The same information can be expressed albeit through the different physical mediums (Dennett, 2017) of neurons, symbols or vibrations within the air. Information is discrete from matter although it is contained within it. To steal an invention I don't need to take anything physical at all. A corporate spy can commit the sought-after design to memory and then take the design with them in their mind. The mode of representing the spinning top can vary but the information conveyed in either case is identical.

Information therefore appears substrate-independent. It is occasionally implied that using this as a basis to endorse mind-uploading leads to some form of unacceptable dualism. This is because it suggests the mind can exist without the body. In his critique of mind-uploading Massimo Pigliuccci makes the following argument (2014, p.125):

> [This] brings us right back to a curious form of dualism, since it essentially assumes consciousness is substrate-independent. I find this position downright bizarre and not

at all disanalogous to claiming that photosynthesis, or life itself, is likely to be substrate-independent. Here I follow Searle's (2008) biological naturalism position, since after all consciousness – so far as we can tell – is a biological phenomenon.

But if this is dualism it is not an objectionable kind of dualism. It would be dualism stripped of the unscientific commitments. Dennett memorably said it is dualism in the sense that "the software/hardware distinction implies dualism" (Dennett, 2017). No one thinks the fact that different systems can implement the same software requires us to radically alter our worldview. Information can be nested within a purely physical ontology,

Responding To the Chinese Room Reply to Computational Functionalism

I wish to briefly reply to the most common critique of computational accounts of consciousness. Variants of John Searle's (1980) "Chinese Room" argument are often cited in reply to the idea that something implementing the right kind of formal system results in a system being consciousness.

The idea is that inside the room there is an English speaker who does not know Mandarin. He is locked inside a room with a large batch of Mandarin writing. The English speaker is then given a second batch of Mandarin writing together with rules (in English) for correlating the first batch with the second. The English speaker is then given a third batch of Mandarin writing together with some further rules in English which allows the English speaker to correlate the third batch with the first two.

The English speaker is sometimes given written questions in English which he, of course, understands and answers by writing in English. On other occasions he is asked questions in Mandarin which he answers by following the instructions. After a time, he gets so efficient at following the instructions that he returns answers indistinguishable from a person who speaks fluent Mandarin. However, unlike the answers returned in English there is no comprehension. All that happens is that the English speaker matches sets of uninterpreted Mandarin symbols together. However, from the view of the person outside the room the answers in English and Mandarin are as good as each other.

Searle argues that an artificial intelligence would appear to be in the same position as the man inside the Chinese room. Although an artificial intelligence might be able to convincingly appear to understand English but it never could. It would simply be matching symbols with other symbols according to a set of rules (Searle, 1980).

I endorse the 'system consciousness' reply to Chinese room variants. While the English speaker does not understand Mandarin, it could be reasonably said that the "Chinese room" does. Similarly, no individual neuron is conscious, but the whole of the brain is (Dennett, 2014). Dennett (2014) points out that this thought experiment and others similar to it, trade off of a failure of imagination and hidden premises. To imagine the English speaker in the Chinese room you need to image someone unimaginably fast and efficient. The occupant of the room would have to travel and comprehend at enormous speeds, selecting each book perfectly, with a reaction time indistinguishable from a normal person's capacity for written report. Unless you imagine these (omitted) details you have not correctly conceived of the conditions required for the thought experiment to proceed. And these conditions are really difficult to imagine and unlike anything we typically experience. Our intuition is unlikely to assist once the experiment is appropriately revised because this is not a situation like anything we can experience. It is the failure to imagine these details which leads to the intuition that the Chinese room must not be conscious.

Why Currently Only Biological Systems Are Conscious.

Douglas Bilodeau (1996) offered sceptical remarks about the prospect of machine consciousness in *Physics, Machines and the Hard Problem*. Bilodeau argues that computers and minds are importantly asymmetric with the latter being only found in organic systems. His examples are unlike a machine, organic systems arise naturally and self-maintain. My brain and body have a capacity for self-repair that a computer does not have. He also submits that the function of a machine is one of design whereas the function of an organ is the result of "biological imperatives attributed to the organism itself". The mind often disregards things that are in the interest of either the body or reproduction. Bilodeau argues that even if we created a machine that replicated a neuron or a cell – it took in energy and emitted a model of waste we would not have created a cell. Any model we made of a neuron would be simplified and only replicate the parts of the neuron that are obvious to us. Pigliocci (2014) makes the familiar observation that consciousness seems to be restricted to organic systems and says this supports the view consciousness is substrate dependent.

Insofar as Bilodeau's critiques carry any weight I think they rest on a misapprehension. Organic systems are the product of a (type of) design – natural selection which non-organic systems have not been. While organic systems are not the product of an intelligent designer, they are the product of a process which has given them a unique degree of complexity. My answer to why consciousness only (seems) to appear in organic systems without being restricted to that

class of objects is complexity. Natural selection is the unintelligent process through which highly complicated systems came into being without being designed. With the advent of intelligent designers (humans), we can begin to see how non-organic systems of similar complexity might come into being by being intentionally designed.

There is competence with comprehension and competence without comprehension (Dennett, 2017). A termite colony building a mound is extremely competent but has no comprehension. In Nagel's terms, there is "nothing it is like" (probably) to be a termite or a termite colony. If a termite is conscious its consciousness is faint. A beaver, building a dam, is competent and probably also has a little bit of comprehension. Most of the beaver's knowledge is instinctive, but there is also some imitation and learning that goes on as well since Beavers get more skilled at their tasks over their life. At the far end of the "competence with comprehension" scale, we have Napoleon, Beethoven or Antoni Gaudi. Dennett (2017) describes some interesting visual parallels between the termite mound and Gaudi's magnum opus the Basilica Sagrada Família. Nonetheless, the Sagrada Familia (which is still an ongoing project) is the product of an intelligent designer (top-down design) rather than the bottom-up process like the termite mounds.

The process of evolution is a process that gave us reasons. Dennett calls these reasons "free-floating rationales". The infamous cuckoo bird places its eggs in the nests of other birds who then raise the cuckoo's children. Cuckoo eggs hatch earlier than the other eggs and when the cuckoo hatches it pushes out the other eggs from the nest. It has a "reason" for committing fratricide, but the reason is not truly the cuckoo's reason. The reason is that by doing this the cuckoo can monopolise the attention of its foster parent and increase its odds of survival, and so this trait was selected for. Evolution brought reasons into existence although those reasons organise around the survival and reproduction of the organism (2017, p. 343). A cuckoo does not need to know why it kills its adopted siblings for this to be adaptively useful behaviour so it does not. When Antoni Gaudi designed for Sagrada Familia his reasons for doing so were his reasons; whether they be vanity, passion or piety. This was extreme competence alongside extreme comprehension.

The reason for the distinction between the cuckoo's reasons and Mr Gaudi's reasons is that Mr Gaudi is a person. Language and communication are a defining feature of human activity. Communication requires internal introspection. To communicate our thoughts and actions we need to be aware of them. Knowledge of our reasons permits us to reflect on whether we should

share them or demur. We do not need to be aware of the causes of our consciousness, however, which is why the origins of our thoughts can be opaque. Evolution was previously the only method that could provide reasons (without reasoners) in need of communication, hence why only biological organisms exhibit consciousness. However, the complexity that unintelligent design endowed biological organisms with can be replicated by human intelligent design. Natural selection never created intelligent machines for the simple reason that machines do not appear in nature or reproduce and so are not subject to selection pressures. This is not a bar to human made conscious machines.

Consciousness can be described as a "user illusion" (Dennett, 2017). A user illusion is a process where the workings of a system are presented in a metaphorical way that makes it accessible to the user. My desktop is currently portraying the current iteration of word processor by Microsoft. The vast number of computations inside my computer are distilled in a way that makes the system accessible for my use. But the pieces of paper on my screen do not have a literal representation in the computations giving rise to them. The trash icon on my computer gives me a visual representation of deleting a file. I can therefore use the system without understanding it. On this view, consciousness is an edited digest of our activities made available to ourselves for, at least in part, external communication.

Consciousness is therefore edited. Large parts of what my body and brain do fall outside of my consciousness (or global workspace) simply because I do not need to understand how they work. I do not need to be conscious of my digestion for my digestive system to be useful. If breath control needed to be consciously controlled all the time that could be fatal. This tracks nicely with global workspace theory and its analysis of items that fall within and outside of the global workspace. This begins to make the ineffability of phenomenal consciousness less surprising. We see a TV and the pandemonium architecture of our brain gets set off triggering a reaction of "Oh, it's a TV!". Our ability to describe an item is what makes us conscious of it, and nothing further.

Chalmers and others would add that this does not answer the Hard Problem. Why is this edited digest/user illusion personally represented? A reply to this critique needs to avoid positing what Dennett (1991) called the "Cartesian theatre". This is where conscious experience is 'presented' to an audience, with objects outside of the theatre also being outside of consciousness. The problem with creating a Cartesian theatre is that the subject watching the theatre of consciousness is itself conscious and so its consciousness must then be explained. This results

in an infinite regress. This is exactly what the user illusion metaphor seems to posit at first glance.

One way to avoid an infinite regress is to suggest that consciousness is spread across space and time. There is not one single narrative of consciousness but rather "multiple drafts" (as Dennett sometimes put it). This is unintuitive since we think we experience consciousness as an uninterrupted narrative. But that is not the case. One way of illustrating this point is the **phi phenomenon**. The phi phenomenon involves two balls being presented to a viewer. These balls then appear to be moving and changing colour in the process. However, the appearance of movement is false. The balls are perfectly stationary. What occurs is that a ball flashes in front of the viewer, followed by a pause, with the other coloured ball then appearing. We see uninterrupted movement between the first flash and the subsequent flash. The balls are stationary, but we see movement even before we see a subsequent flash! Absent attributing foresight to ourselves this is very hard to make sense of.

Dennett offers two explanations as to what might be happening here. He calls one explanation "Orwellian" and the other "Stalinesque". We are either misremembering what happened to the ball (Orwellian) or the movement is being inserted into our experience as we go along (Stalinesque). Dennett argues there is no way we could distinguish between which of these two options is the truth and that because of this there is no fact of the matter about when exactly this event falls into consciousness.

It seems to me plausible that we could have a change in phenomenal consciousness without being aware of it. People are wrong about their subjective experiences frequently. Famously, Daniel Kahneman's "prospect theory" observed that people evaluate definitively net more painful experiences as less painful provided the pain gradually decreases at the end. If I am tortured for 1 hour I will rate that as less painful than an identical 1 hour torture plus another 40 minutes of gradually diminishing torture. Likewise, dysphasia can involve people speaking gibberish without being aware that there is an issue. Split-brain subjects must also be mistaken about the contents of their experience. What this illustrates is that reports of our phenomenal experience are not authoritative on what our phenomenal experience actually is. This may explain why (in Kripke's terms) the relationship between consciousness and computation appears contingent.

How A Reductive Account of Consciousness Might Proceed

Computation coupled with the assertion that we are mistaken about our experience may seem unsatisfactory as a description of phenomenal consciousness. To explain why this explanation seems inadequate and how consciousness could be caused by computation we can appeal to Hume's scepticism about causation. Human beings are primed to see causation even when there is none. Seeing causes and patterns was essential to the survival of our ancestors. Unless we drew a link between eating the mushrooms and feeling ill, we might do so again. Nonetheless, the feeling we understand causation is deceptive.

Mark Price (1996) in his article, *Should We Expect to Feel as if We Understand Consciousness?* argues that causal gaps in our ontology are ubiquitous. This, he says, undercuts the Hard Problem. Although we do not understand how consciousness is caused by physical processes, we equally do not understand how other physical processes cause other physical processes. It is simply this gap is more obvious with consciousness.

What we are seeing when we perceive causation is "A followed by B" and not "A caused by B" (Dennett, 2017, p.354). We perceive "A followed by B" as causation through a mixture of habit and genetics. Price (1996) notes that we will usually need a complex range of "regularities" before we can speak of causes, and not all regularities will be causes. For example, a rooster crowing as the sun rises does not cause the sun to rise (Price, 1996). Instead, we often need to speak of what are referred to as "INUS" conditions – "insufficient but necessary parts of an unnecessary but sufficient condition". Multiple things, in aggregate, may be one of multiple different things which together we say "causes" something. Obviously, due to the complexity of the human brain, the regularities in question which precede consciousness will be especially complicated. It will therefore be easy to see a gap between the causes and the effect, but this does not render consciousness ontologically unique.

This tendency to perceive causation naturally leads to frequent misattribution of causation. We can see this clearly with religion. People of differing and incompatible religious affiliations all seem to find that their prayers are answered. Of course, they often find that God answers their requests with a "no", yet the pious still attribute the outcome to their prayer. If two rival groups each pray for one of two mutually exclusive outcomes then of course one will be right! The efficacy of prayer can never be falsified because whatever happens will be seen as the result of the prayer. More prosaically, TV requires us to misattribute causation otherwise we would not see a series of flashing still images as movement.

With phenomenal consciousness we can see similar misattribution. The mind projects itself onto other objects and so we see sweetness as a property of sugar, redness as a property of a tomato and funniness as an object of the joke (Dennett, 2017). The true origins lie inside neural events occurring within the brain. The misattribution in causation which leads to view there is a "Hard Problem" is that when we make subjective judgments we attribute our subjective experiences as arising from a subjective property that is the source of our judgments. Whereas in reality is our ability to describe our judgments that are the source of our convictions about them. Both Chalmers and Dennett would agree reductive explanation ends somewhere, and Dennett asserts that this is where this is.

Price makes a further observation that even if we had a fully scientific explanation of consciousness, we may nonetheless still feel like we have not understood phenomenal consciousness. Psychological causation is concerned with normal conceptions of causation. Philosophical explanations of causation try and explain what causation truly is. The folk conception of causation is that one thing makes another thing happen. We believe we have identified a cause when we identify a 'causal nexus' that leads to a necessary connection between the cause and effect. We feel we have understood something when we have identified the causal nexus (Price, 1996). The absence of this feeling of a causal nexus is both what leads to the Hard Problem and by extension the feeling that any identification of consciousness with a set of physical properties would be contingent.

Chalmers (1997) observed in reply to such arguments that if we acknowledge causal gaps are common in our ontology and concede this is an example of one of them, then consciousness is clearly not amenable to reductive explanation. The essence of the problem is that although there are blunt causal gaps within our ontology they are considered exceptional. Reductive analysis usually suffices. For example, biology can be reduced to genes. Genes can then be reduced to DNA, and DNA can then be reduced to chemistry. Chalmers' says the reason things such as gravity cannot be adequately explained by science is that they are "fundamental". They are fundamental because they are explanatorily indispensable but cannot themselves be further explained. They appear to be brute facts about nature. If we are taking consciousness as an instance of a casual gap Chalmers' argues that consciousness must be fundamental and that this leads to either dualism or pan-psychism.

I think Chalmers' reply misses the mark. Consciousness does not have to be fundamental for us to not feel like we understand the causation involved. The absence of this feeling of

understanding causation does not imply that there is no actual understanding. When we narrow in on complex issues in quantum mechanics any feeling someone has of understanding is surely chimerical. We have maths which undoubtedly has amazing predictive power and we take quantum theory as a causal explanation because it is predictive. But for (at least some of us) reifying maths, into a feeling of understanding of what the maths explains, is impossible. Some mathematical models of how the universe ends suggest that stars will eventually turn into large lumps of iron floating in an endless void. The feeling of causation here will be elusive for possibly everyone. Science searches for a coherent articulation of universal laws – those universal laws need not be causes we find satisfactory psychologically. Human minds evolved to answer the problems of Palaeolithic Earth and may not be well placed to understand consciousness.

In an earlier section, I noted that predicate dualism suggests mental predicates might not be reducible to physical predicates. I expressed doubt that this was truly dualistic in its ontology. If this makes me a dualist it is only a trivial semantic sense. A soft form of predicate dualism might also help illustrate the appeal of the "Hard Problem". Our physical descriptions of a system might poorly latch onto mental descriptions of a system because we experience mental systems from the "inside". Before it is objected that this opens the door to more significant forms of dualism, it can be observed that forks are not reducible to any single class of physical items. You can have a fork made of wood, glass, ice, plastic or metal. No one would assert we ought to be (seriously) dualistic about forks though. Minds may just be an exceedingly complicated example of the same sort of principle. Referring back to computational functionalism, many current computer programs and algorithms have acquired significant complexity through machine learning. This renders their workings impenetrable to their makers. Once the algorithm starts acquiring new information and correcting and adjusting its own behaviour, the system becomes impenetrable to outsiders. Nonetheless, no one would suggest that the computer is not operating according to predictable physical laws.

Moreover, it might be that consciousness is amenable to human understanding but that this understanding will be diffuse owing to the complexity of the human brain. Different specialist groups will understand different components, but no one has the aggregate expertise to put the picture together. It might be that distributed competence allows for group understanding. Science is replete with works completed by multiple authors (Dennett, 2017). For completing complicated tasks collaboration is vital between theoreticians and experimentalists – each of whom may not have a full understanding of the skillset and details the other party works on.

Given the complexity of the brain, it may be that a full understanding of consciousness escapes any individual's comprehension but as a whole, it is eventually explained by science (Dennett, 2017, p. 354).

6. The Implications of Different Philosophical Accounts of Personal Identity For Whether We Survive a Mind-Upload Procedure

I have argued that, plausibly, a computer could share an identical conscious state with ourselves. I have tentatively endorsed reductive computational functionalism. However, mind-uploading does not just require that our consciousness can be simulated. The mind-upload must also be me. It would be no good if we could copy out consciousness perfectly but die in the process. Simulating our mind must permit us to continue as the same person.

The question of whether we would survive digitizing our mind is a broader question about what constitutes our personal identity. Replication of phenomenal consciousness is a necessary but not sufficient condition for surviving a mind-upload. I have some doubt that there is any combination of facts sufficient to describe our personal identity. Each philosophical account of personal identity appeals to some intuitions but disregards others. I will argue in favour of a "closest continuer" view of personal identity, but I have my doubts. It is not that the mind-upload would be a "zombie", it just may not be you!

Different Criterion of Personal Identity

Chalmers (2010) states that there are three different accounts of personal identity:

1. Facts about psychological continuity; and

2. Facts about biological continuity; and

3. Closest continuer theories.

To this list I think we can add two more:

4. Further fact accounts (for example, soul theories and Parfit's "R" account of identity); and

5. Error theories: There is no such thing as personal identity.

It should be noted that hybrid views are possible. One could argue that both psychological and biological continuity are both necessary conditions. Pigliocci (2014) endorses such a position. Such a view is incompatible with mind-uploads for the same reason the biological account is incompatible with them.

Most accounts of identity succeed equally in explaining day-to-day personal identity. It is only in fringe cases they break down. Most people would say that Joe Biden the President is the same man as Joe Biden the Senator and essentially all theories will agree on that answer. He is the same organism he previously was, and he shares memories/psychological connectedness between his time as a Senator and his time and as President. He is also his own "closest continuer" and if souls existed, we would attribute the same soul to Joe Biden's body across its existence.

Neither The Psychological Criterion nor The Biological Criterion Offers a Plausible Account of Identity

It is plausible that creating a replica of your consciousness would not be sufficient for the preservation of your personal identity. This is the area where I submit the risk of a mind-upload being death is most pronounced.

The biological criterion is the view that you are the same person as you were in the past if you are the same biological organism. This view is straightforwardly incompatible with the view that you continue as a person during a mind-upload. A computer simulation of you would not be you since a computer simulation is not a biological organism. A computer simulation might be a different class of person, however, since the biological criterion is only committed to the view that we are biological organisms.

A contrasting view, the psychological criterion, holds that personhood is mental. I am the same person as I was yesterday by virtue of some psychological connection to that prior version. A common proposal is *memory*, since I remember being the same person yesterday and in ordinary conditions have memories from our whole life. John Locke (1694) offered the famous example of the "Prince and the Cobbler". Locke states that if the "soul of a prince" which contained the knowledge of the prince's past life was placed inside the body of a cobbler then the person of the prince would now inhabit the body of the cobbler, and accordingly be responsible for everything the prince had done. This is clearly similar to a mind-upload. The psychological account of identity provides a basis for endorsing mind-uploading.

One argument in favour of the biological criterion is that it captures the view that we would persist as the same person over time even if we lost all our memories. If you asked me whether 'I' would like to undergo a painful procedure which results in me having an entirely different personality and memory set, or whether I prefer to painlessly die, I would choose the former. If a psychological criterion was the true source of personhood, then I should prefer the second since both are death. It is not obvious that this is an error. The biological criterion naturally makes sense of the view that we continue as the same person even with total amnesia which is also a common intuition.

Nonetheless, the biological criterion is clearly open to criticism. For example, if I have a full brain transplant (unlike Swinburne's partial brain transplant) we have the obvious question – do I go with my brain or do I stay with my body? (Olson, 2023). Imagine that my friend Seth and I engage in a philosophically motivated brain transplant operation. If Seth and I stay with our body, and not our brain, then we are left with the view that Seth (who thinks he's Nathaniel) and Nathaniel (who thinks he's Seth) are both wrong about who we are. Because of this, the biological criterion is sometimes narrowed to the view that we are just brains. I suggest this is arbitrary since a brain without a body is just a decomposing organ. In this thought experiment we seem to favour a psychological criterion because we seem to think we go with our brain.

The psychological account of identity does not seem to be much better at avoiding unintuitive answers, however, than the biological criterion. There are clear problems with memory accounts of identity. Thomas Reid provided the example of a brave officer who was flogged when he was a boy for robbing an orchard. The brave officer took a standard from the enemy in his first campaign, and then he was made a general late in life. When he took the standard he remembered robbing the orchard, and when he was a general in his dotage he remembered taking the standard but had forgotten being flogged. The memory criterion suggests that the brave officer was the same person who robbed the orchard, and that the general was the same person as the brave officer, but the general was not the boy who robbed the orchard. This creates a logical flaw which is fatal for the memory criterion.

The flaw in the memory theory cannot be corrected by specifying that the memory theory of personhood simply requires the less onerous condition that we "could" remember what we did under non-tautologically specified conditions. This is because the general, for example, may be senile and so would not remember the flogging no matter how much prompting he is given. Further, no help can be gained by modifying the memory condition to include the ability to

remember a time where you remembered it ("quasi-memories") because there are many things you cannot quasi-remember (Olson, 2023).

The psychological criterion is sometimes weakened to just a requirement that there be the right kind of "psychological continuity" across a life. This continuity could be habit, preferences or dispositions. For example, if I had some form of dementia there should be some form of psychological continuity between myself over time even if my memory is significantly impaired. For example, perhaps some of my habits or language skills continue. This would be sufficient. Alternatively, we can be temporally indifferent about the ordering of our memories. So, on this view, so long as there is some rearrangement of our memory or psychological states that would ensure direct continuity then we persist as the same person. This meets the brave officer objection but at the risk of becoming too amorphous.

I submit in the context of **non-destructive** uploading all psychological criteria get the wrong answer. In destructive and gradual uploading there is only one conscious being with connected psychological states. In non-destructive uploading, there are two. The biological criterion of identity seems to be more in harmony with our intuitions in this instance. Most people would say that if the biological organism and the computer simulation existed concurrently then the biological organism has the greater claim to being my person. Identity is a straightforward 1 – 1 relation. It is not likely that both the computer simulation and the biological organism are the same person, and in any instance, their psychological states will branch in different directions after the upload. Nothing has changed with us – so there should not be any break in personal identity. Nonetheless, if psychological continuity is the only criterion then there is no sensible way of distinguishing between ourselves and the mind-upload when evaluating who has a claim to the pre-upload identity. I take this as a good reason to assume that the psychological criterion, on its own, seems implausible.

By contrast, with **gradual uploading** the biological criterion gets the wrong answer. If I replaced all the neurons in my head with silicon ones, and all the "flesh" parts of my body with computer parts, and did this over a very long period of time I would seem to be the same person throughout (for reasons already given). Nonetheless, the biological criterion says that I am not the same person and will have to offer some arbitrary point throughout the transition and say that is when I ceased to be the same person. This is not very compelling.

Evaluating Further Fact Accounts of Personal Identity

I disagree with Swinburne (2018) when he says we can take it as a given that some fact constitutes personal identity. His basis for this assertion appears to be that we must have it because we think do. Swinburne writes "Each of us normally hopes to survive an operation, and we know what we are hoping for – that is "I" will continue to have a conscious life" (p.15). He says because of this, there must be a clear answer to his partial brain-transplant experiment about what happens to identity. Swinburne states that partial brain transplants justify believing in a soul account of personal identity, since if two separate people each had half of a whole brain (post-brain transplant) there is no basis for ascribing continuing identity to one of these people over the other. Swinburne (2018) also offers a second argument in favour of his further fact account. He suggests that if small parts of our brain were replaced very gradually, in a way not dissimilar to Chalmers' method of gradual uploading we would never be able to say when that brain had ceased to be us (or whether it would be us) therefore facts about our mind cannot be facts about the brain.

The view that we could not be mistaken about identity seems to me like it belongs as a conclusion to an argument, rather than offered as an assertion in support of one, but I grant Swinburne is right that souls would provide an elegant account of personal identity. Although we may not know who has what soul in the case of a partial brain transplant, there would be a fact about the matter, and that fact would suffice to establish personal identity. Souls, however, are themselves a fact in need of explanation. For reasons outlined earlier, there is no reason to believe they exist. I would go further than Swinburne and lightly suggest that our notion of continuing personal identity may derive from Abrahamic conceptions of a soul. But if the soul does not exist (as I have argued) we may need to accept there is no concrete set of facts that describe our identity and make do with more malleable criteria. Our description of personal identity should try to do justice to our intuitions about what our identity is, but since all anyone is, is a gradually changing collection of atoms, our identity may not be clearly fixed.

Derek Parfit (1971, 1984) offered a similar thought experiment to Swinburne's which I think highlights the flaw in identifying concrete criteria for identity. Parfit imagined that a brain is transported into another (brainless) body. The person arising from the brain/body transplant has your personality and memory. The ordinary intuition people have is that this person would be you. The next element of his thought experiment is that half our brain could be destroyed. Parfit notes that people have survived with half their brains destroyed. While this is not an ideal situation, these people usually behave similarly to how they did before the destruction of half

their brain's hemisphere. A transplant of only half a brain should be sufficient to preserve identity.

The challenge arises if we divide our brain into two parts similar to how Swinburne suggests. Each half of my brain is then placed into a different brainless body. The two individuals arising also have your personality and memory.

Derek Parfit observes that there are three possibilities.

(a) You do not survive at all; or
(b) You survive as one of the two people; or
(c) You survive as both people.

Parfit says that none of these three options is satisfactory. The claim you do not survive is strange – if we admit that I could survive if my entire brain was transported into a body or if only half of my brain is transported, then why would I not survive if my brain is split and transported into two bodies? Parfit asks how a "double success" could be a failure. The second option is equally untenable. It seems deeply implausible to argue that you survive as only one of the transplants (unless there is a soul to break the tie as Swinburne suggests). Each person would be psychologically identical to you – it is therefore impossible to distinguish which of the two of these persons would be you. The third option is also not tenable. Personal identity is a-priori a 1 to 1 relation. Two organisms with distinct consciousness cannot both be me by definition. I note that the biological criterion does not adequately answer this problem. If we retreat to the view that we are a brain (rather than a whole animal) to make sense of our intuitions that "we would go with the brain" in a brain transplant then which half of the brain are we if our brain is split in two?

I do not see why the question "Which hemisphere is me?" in a partial brain transplant must have a clear answer. I tentatively submit that there is no hard set of facts that constitute our personal identity. Both the biological criterion and the psychological appeal to some of our intuitions and disregard others. There might be an evolutionary explanation for this and we can craft an "evolutionary debunking argument" similar to those used in metaethics to undercut the claim that our identity is rigidly demarcated by some concrete fact.[14] The debunking argument

[14] These are used to rebut moral realism. The idea being that since moral reasoning is an evolved trait we would be disposed to make moral judgments whether or not there are any moral facts to know in

appeals to the fact that any intelligent organism would evolve a strong notion of personal identity since personal identity propels self-preservation. By being aware of a 'me' to care about I am more likely to take steps to keep myself alive. We therefore will have evolved a strong feeling that we have a clear identity whether or not there is a coherent set of facts that specify that identity. It does not follow from this rough debunking argument that we need to abandon the notion of personal identity altogether. Ordinary colour judgments are often taken to be false, but some philosophers believe our use of colour language can be saved.

Parfit's solution to the problem is to suggest what we care about is the degree of psychological relatedness "R" rather than identity. Both hemispheres would be equally psychologically related to me and so that is why we would prefer both exist rather than neither. This explains why we would want both hemispheres to survive in a double brain-transplant. It is not the same as identity which is 1 to 1 because conceivably multiple beings could have an "R" relation to you. It has been noted that Parfits "R" relation runs into the same Brave Soldier challenge as Locke's memory account, but the same answers used to circumvent those challenges can be used to circumvent challenges to the "R" relation. We can amend Parfit's relation to clarify it does not matter what order the psychological connectedness appears in so long as there is psychological connectedness.

I agree that Parfit has raised a significant challenge to both the psychological and biological account of personal identity, but I am unpersuaded by his "R" solution. In Parfit's analysis "R" is the degree of psychological connectedness between yourself and some other entity. Parfit acknowledges the consequence of this view is that we gradually become less related to earlier iterations of ourselves over time. I have a lower "R" relatedness to 'myself' 8 years ago than I do to the version of 'myself' from one day ago. The problem is – of course –that we think we are the same person over time. If Parfit is substituting one unintuitive conclusion for another then it is not obvious that he has resolved the problem.

If Parfit was right that "R" is what matters then this would have interesting implications when applied to **non-destructive** mind-uploading. When applied to mind-uploading it suggests we should maximise the number of non-destructive uploads to maximise the number of entities which persist with regards to relation "R" to us. Taken to its logical conclusion this becomes absurd. No one would care about creating 1 million non-destructive mind uploads versus 1

---

the world. Similarly, we would make very different moral judgments if our evolutionary history had been different.

million and 1. While diminishing marginal returns would creep in very quickly, it is unobvious that having even a single non-destructive upload in existence is a benefit. I will die regardless, and my mind-upload will have its own set of experiences which will not be my experiences.

If Parfit is right that "R" is what matters, then what we would care most about is that one entity persists in a chain of psychological connectedness to us. This would still explain why we would prefer both hemispheres of our brain survive albeit in different bodies (what reason would we have to want both destroyed?) while also more strongly preferring that at least one hemisphere continues. A non-destructive upload would therefore be a good thing. This brings me to the closest continuer accounts of identity, which I suggest are "the least bad".

Endorsing The Closest Continuer Account of Identity: We Might Survive A Mind Upload

In *Uploading and Branching Identity*, Michael Cerullo (2015) argues that arguments similar to Parfit's when applied to mind-uploading should push us towards accepting a branching view of personal identity. Cerullo (2015) describes that an entity could be qualitatively identical to me without being numerically identical and it is numeric identity that is embedded in our notion of identity. Cerullo suggests we abandon this and accept branching identity. His basic argument is that traditional views of personal identity cannot accommodate mind-uploading and so we should be willing to countenance branching personal identity

I do not endorse branching identity. While our notion of identity may be flawed, if anything is fundamental to it, it is the notion that there is only "one" me. This is the intuition that underpins Swinburne's and others' works. Instead, I support a view of personal identity where it is composed of a basket of things that are not on their own necessary or sufficient conditions. These would include psychological connectedness, biological connectedness, and temporal-spatial connectedness. I am sure there are other properties I am missing. Whichever entity has the most items from this basket is "you".

My approach is therefore to endorse Robert Nozick's (1981) closest continuer (and therefore *non-branching*) view of personal identity. This view is the one that most neatly accommodates all three forms of mind-uploading. The closest continuer view states that whichever version (in a loose sense) of myself is closest to the previous one has the claim to my personal identity. In the case of destructive and gradual uploading this view would state that we survive the upload process. In the case of non-destructive uploading, we do not become the mind-upload since the mind-upload would not share all the attributes of myself I possessed before the upload process - it would not be a biological organism.

The problem with closest continuer accounts of identity is that the insertion of a mandatory non-branching condition is arbitrary. Sometimes "X" is me and sometimes "Y" is me – depending on the circumstances. I concede that such a condition is arbitrary, but that has to be weighed against the fact that the closest continuer view consistently gets the right answer in challenging fringe cases. I persist as myself if I have amnesia, and I persist in a brain transplant.

I also persist as myself in the interesting case of tele-transportation. In Derek Parfit's (1984) famous 'tele-transporter' thought experiment you have a tele-transporter which puts you to sleep, breaks you down, records your molecular composition, and then relays the information to Mars. 'You' are then reassembled from deposits of carbon, hydrogen, and nitrogen etc on Mars. Each atom in the reassembled 'you' occupies the same relative position as it did previously, but no atom is the same as those that constituted 'you' previously. The tele-transporter experiment asks whether tele-transportation is death, or merely travel.

Intuitions differ although a slight plurality of philosophers say that the tele-transporter is survival rather than death (36.2% vs 31.1%) and a slight plurality favours the psychological criterion over the biological and further fact criterion (33.6%, 16.9% and 12.2% respectively) (Bourget & Chalmers, 2013). One intuition in tele-transportation is that although a new person has been created and that person is indistinguishable from you, they are not you. You, the 'organism' have been killed and a copy has come into existence. Although there would be unbroken psychological unity between you and the copy, continuing identity would be an illusion. There is just a clone of you that happens to think it is you. After all, if the tele-transporter had not disintegrated you but still assembled a 'copy' of 'you' on Mars then most people would assert you remain on Earth and that there is only one 'you'.

The closest continuer account can defuse the challenge of Tele-Transportation. Tele-transportation is not death if I am disintegrated and then assembled on Mars because there is a closest continuer on Mars. However, in the case where the tele-transporter makes two copies of me, I am the copy on earth because that is the version of me that is the closest continuer temporally and spatially. Although the two copies are externally indistinguishable, only one occupies the correct position in space-time.

When applied to mind-uploading closest continuer accounts offer the right answer. In the case of non-destructive uploading our personal identity does not go with the upload. My closest continuer is the entity that is biologically and psychologically continuous with myself before the upload. This would not be the mind-upload. However, in the case of the destructive upload

there is no longer a unit that is biologically continuous with me and so the closest continuer is the mind-upload. However, the destructive upload is less temporally and biologically connected to me, so it has fewer items from "the basket". This explains why some people are more equivocal about destructive mind-uploading. Naturally, in the case of gradual mind-uploading there is only one closest continuer throughout the entire procedure so I continue as myself throughout the upload process.

Closest continuer views, perhaps, can meet the problem of partial brain-transplants. The answer to which of the two severed hemispheres of the brain would be me would be "whichever hemisphere is most similar to who I was. The natural objection is both hemispheres when ensconced in a new body would be identically similar to me (as Parfit suggests). Cerullo describes each hemisphere post-transplant as "equally psychologically continuous" with the pre-transplant brain/person (2015, p.21) and so submits that closest continuer views cannot give a compelling answer to Parfit's thought experiment. In one sense that is true since both hemispheres trace were previously part of an integrated person, but in an important sense it may be false.

I think the claim that both hemispheres would be equally psychologically continuous to us misleading. While obviously, each hemisphere plays an important role in our cognitive functioning it is not obvious that they have an equal role in all functions. Whichever hemisphere contributed most to our behaviour pre-transplant and was most similar to us post partial-brain transplant would be an empirical question about the degree of psychological connectedness. I see no good reason to believe this does not have an answer. We might adjudicate the dispute by stating that whichever version of us post partial brain-transplant has the attributes that we would have valued most before the transplant is the version we continue as. The left hemisphere offers the greatest contribution to language and logical reasoning and so we might sensibly say we go with the left hemisphere.

It therefore appears that the closest continuer account is plausible and consequently that there is a good chance we would survive a mind-upload.

Implications Of the Closest Continuer Account: Mind-Uploading Would Be Desirable.

Approval of mind-uploading is linked to dark triad characteristics, Machiavellianism and utilitarian outlook (Laakasuo et al. 2021). Since some recoil when contemplating mind-

uploading it is worth sketching an affirmative case for it. My argument is that the closest continuer account of identity suggests that mind-uploading is desirable.[15]

Mind-uploading has obvious advantages to biological immortality. Even if anti-ageing technology is invented our bodies would still be susceptible to death from accident or homicide and so on a long enough timeframe death remains inevitable. Mind-uploading, is possible, makes us immune from physical death. If our hardware was terminated, a new computer could be built and continue to generate our mind. The software behind our mind could presumably be stored in the "cloud" for safekeeping. Mind-uploading offers a form of immortality that is more secure than biological immortality.

That does not matter if immortality is bad. Objections to the desirability of immortality often cite intolerable boredom. Bernard Williams (1973) in *The Makropulos Case: Reflections on the Tedium of Immortality* argued that eventual death provides life with meaning even though dying is still always bad. Bernard Williams suggests that a world full of immortals would be (paraphrasing the inimitable French diplomat Talleyrand), a world of Bourbons who "learnt nothing and forgot nothing". Williams argued that after a sufficient period we would run out of categorical desires. New experiences would be rare and eventually non-existent. Painful boredom would then set in. Infinite time may also deprive us of creative energy and drive to achieve that mortality instils in (some) of us. A society of immortals could lead to people so entrenched in their habits that neither art, politics or science would progress very much.

Other arguments against immortality focus on the death of friends and family, although these do not matter if your friends and family are immortal too. They also ignore that obviously, you can just make new friends. Replies to Williams often focus on whether you would be able to amend your desires over time. Some argue that an immortal life would come with the prospect of having new desires related to goals that would take longer than a human lifetime to complete. These goals could be endless such as maximizing human well-being or reflecting on more

[15] There are other reasons to endorse mind-uploading. If you are a utilitarian, you might see mind-uploading as a way of maximising aggregate well-being (Laakasuo et al. 2021) (Bostrom, 2003). The idea is that death is an "astronomical waste". Death causes enormous unhappiness to those who watch their loved ones die, and it also reduces the number of people, lowering overall well-being. Having more beings alive will raise overall well-being. Construed this way, mind-uploading might be a positive moral obligation.

perfect mathematical theorems (Burley, 2009). Of course, having the option to live forever is not the same as having to live forever. Anti-ageing technology or mind-uploads would not force the person to live longer than they want to. As long as very long lives are something we might want, then we have reason to endorse mind-uploading. Moreover, even unhappy people usually do try and avoid death, or at least do not court it. It might be that death is a very bad thing regardless of your state of well-being.

The advantage is that mind-uploading, in a highly speculative way, circumvents these objections. Loneliness would scarcely be a problem if other people uploaded their minds, and perhaps we might eventually program away out of loneliness if needed. We are unable to remould our organic brains in the same way we could remould computer programming. A mind-upload would also be able to make incremental improvements to itself to achieve new goals which offers endless opportunity for achievement. Much discussion of artificial intelligence focuses on the risk of an AI making gradual changes to itself and surpassing general human intelligence entirely. Nothing precludes a mind-upload from doing something similar. For example, an uploaded mind might gradually increase its computational power by adding simulated neurons. Assuming computational power scaled accordingly, if our uploaded mind found itself at risk of unendurable discomfort it could anticipate and make incremental adjustments to itself.

The concern in doing this would be that we would eventually be so different to who we were once that it no longer makes sense to describe us as the same person. But I argued earlier with gradual uploading cases that so long as changes to ourselves are made slowly there is no loss of identity. This is not that radical. There is a wide gap between the infant and the adult in both personality and intelligence, but the changes happen so slowly that we ascribe shared personal identity to both. In any case, improving ourselves would not require us to abandon things we think are core to our identity. Someone who is passionate about mathematics could improve their mathematics ability to better achieve their current goals. Like in the case of gradual uploading, there would no moment in time where we could sensibly point to a change so abrupt it ruptures identity.[16]

Conclusion

[16] The psychological criterion for identity can also get this answer.

My argument has been that mind-uploading using some plausible assumptions about both the mind and identity. Despite its dualistic appearances mind-uploading can be harmonised with a purely physicalist ontology. A belief in its possibility does not require any concessions to unattractive dualism.

To alleviate the concern that phenomenal consciousness would disappear in a mind-upload I suggested that a reductionist account of consciousness be adopted. I provided a detailed discussion of the objection to reductionist accounts of consciousness which is what David Chalmers has called "the Hard Problem". I explained why the Hard Problem is so difficult to solve and explained how this buttressed substance dualism.

I then crafted a (to my knowledge) novel argument advocating for the incompatibility of substance dualism with mind-uploading which drew on recent work from the philosophy of mind. This gave expression to the view that if souls existed then they could not be uploaded. I then provided a critique of substance dualism. For thematic reasons I used the example of the split-brains to explain why substance dualism appears implausible. The soul can be reasonably expected to be aware of all its conscious experiences yet split-brain patients seem to have bifurcated consciousness. Unless, the substance dualist wants to imply that after the experiment the patient has two souls then substance dualism does not seem able to neatly accommodate this. Positing two souls is implausible because for the most part split-brains present as a unified person and it is only in experimental contexts the disunity becomes pronounced.

To explain how mind-uploading can be accommodated with a physicalist ontology I argued that computational functionalism was a plausible explanation of the mind. The brain is a system which processes information and information is substrate-independent. It follows that it is plausible that the some set of computations or information processing would be sufficient to instantiate a human mind. I do not offer computational functionalism as the definitive answer to the Hard Problem of consciousness, but I pointed to some arguments that suggest that the Hard Problem might be built on a misconception of phenomenal consciousness. In particular, I offered the contour of an evolutionary account of phenomenal consciousness from the works of Daniel Dennett. I noted that introspection and an awareness of your conscious states is a necessary part of communication which explains how phenomenal consciousness might have evolved. I also discussed a Humean argument which suggests that a lack of understanding about how physical states fully explain consciousness does not constitute a full explanation. This is a human limitation that does not necessarily imply dualism.

In the final chapter, I supported the closest continuer theory of personal identity as the least bad explanation of personal identity. I noted that it can meet common objections that other accounts of personal identity struggle with. I particularly respondent to Derek Parfit's (1984) argument that personal identity is not what matters because if both your hemispheres were removed there would be no fact about which is you. I submitted that there are good odds that one hemisphere would exhibit more behaviour that we link our personal identity than the other hemisphere. We can therefore prefer that hemisphere as our closest continuer.

I also argued that the closest continuer accounts of personal identity give the intuitively right answer to whether we survive all three possible mind-upload procedures. It is obvious that if we non-destructively uploaded our mind we would continue as our biological self and not the computer upload. In the case of a gradual upload, there is no significant break in continuity at any point throughout the upload process and so we remain the same person throughout. Finally, although the case of destructive uploading creates more equivocal intuitions as to whether we survive it, it is certainly plausible that we do and that is the answer the closest continuer view offers.

Bibliography

Bayne, T. (2008). The unity of consciousness and the split-brain syndrome. *Journal of Philosophy* 105(6), pp. 277-300,

Bilodea, D. (1996). Physics, machines, and the hard problem. *Journal of Consciousness Studies. 3*(5-6), pp. 386-401, https://philpapers.org/rec/BILPMA

Burley. M. (2009). Immortality and boredom: a response to Wisnewski *International Journal for Philosophy of Religion* 65(2), pp.77-85. https://philpapers.org/rec/BURIAB-3

Bourget, D & Chalmers, D. (2014). What do philosophers believe? *Philosophical Studies* 170(3), pp. 465-550. https://philarchive.org/rec/BOUWDP

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas* 15(3), pp. 308-314. https://philpapers.org/rec/BOSAWT-3

Buckner, Cameron and James Garson, "Connectionism", The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>.

Carter, M. (2007) *Minds and computers: an introduction to the philosophy of artificial intelligence*, Edinburgh University Press

Cerelullo, M. (2015). Uploading and branching identity, *Minds and Machines 25*, pp. 17-36. DOI 10.1007/s11023-014-9352-8

Chalmers, D. J. (1995). Absent qualia, fading qualia, dancing qualia. In T. Metzinger (Ed.), *Conscious experience*, pp. 309-328. Ferdinand Schoningh

Chalmers, D. (1997) Moving forward on the problem of consciousness, *Journal of Consciousness Studies, 4*(1), pp. 3-46. https://philpapers.org/rec/CHAMFO

Chalmers, D. (2003). "Consciousness and its place in nature" in (eds. Stich S & Warfield T (eds.) *Blackwell Guide to the Philosophy of Mind*, https://philpapers.org/rec/CHACAI

Chalmers, D. (2010). The singularity: A philosophical analysis. Journal of Consciousness Studies, 17, 7–65.

Chalmers, D. (2014). Mind-uploading: a philosophical analysis. In R. Blackford & D. Broderik (Eds.) *Intelligence Unbound: The Future of Uploaded and Machine Minds*. https://doi.org/10.1002/9781118736302.ch6

Chalmers, D. (2014). A computational foundation for the study of cognition. *Journal of Cognitive Science*, 12(4), 323-357.

Chalmers, D. (2016). The combination problem for panpsychism. In Ludwig Jaskolla and Godehard Bruntrup, eds, *Panpsychism*, Oxford University Press

Clarke, Randolph, Justin Capes, and Philip Swenson, "Incompatibilist (Nondeterministic) Theories of Free Will", The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2021/entries/incompatibilism-theories/

Crick, F & Koch, C. (1995). Why neuroscience may be able to explain consciousness. *Scientific American* 273(6), pp. 84-85. https://philpapers.org/rec/CRIWNM

Dennett, D. (1991). *Consciousness explained.* Penguin Books.

Dennett, D. (1994). The practical requirements for making a conscious robot. *Philosophical Transactions: Physical Sciences and Engineering* 349(1669), pp. 133-146

Dennett, D. (1994). The practical requirements for making a conscious robot. Philosophical Transactions: Physical Sciences and Engineering 349(1669), pp. 133-146

Dennett, D. (2013). *Intuition pumps and other tools for thinking*, New York: W. W. Norton & Company

Dennett, D. (2017). *From Bacteria to Bach and Back: The evolution of minds*. Penguin Books.

Fodor, J. (1975). The *Language of Thought*. New York: Thomas Y. Crowell

Gordon-Roth, Jessica, "Locke on Personal Identity", The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/locke-personal-identity/> .

Hershnov, T., Taylor, (2014). A. Split-brains: No Headache for the Soul Theorist. Religious Studies 50(4), 487-504. doi: 10.1017/s0034412514000109

Hassan, A. & Keither, J, (2016). Alien Hand Syndrome. *Current Neurology and Neuroscience Reports; Philadelphia* 16(8). pp. 1-10. DOI:10.1007/s11910-016-0676-z

Jackson, F. (1986). What Mary didn't know. *Journal of Philosophy* 83(5), pp. 291-295 https://philpapers.org/rec/JACWMD

Katz, M. (2008). Analog and Digital Representation. *Minds and Machines* 18(3), pp. 403-408. https://philpapers.org/rec/KATAAD

Kirk, Robert, "Zombies", The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/zombies/>.

Kripke, S. (1980). Naming and necessity. *Philosophy* 6(217), pp. 431-433

Laakasuo. M. Repo, M.Drosinou, M. Berg, A. Kunnari, A. Koverola, M. Saikkonen, T. Hannikainen, I.Visala, A. Sundvall, J. (2021). The dark path to eternal life: Machiavellianism predicts approval of mind-upload technology, Personality and Individual Differences 177, https://doi.org/10.1016/j.paid.2021.110731

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies* 6(8-9), pp. 47-57. https://philpapers.org/rec/LIBDWH

Locke, J (1689). *An essay concerning human understanding*. Oxford University Press.

McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. Oxford University Press. https://doi.org/10.1093/0195079981.001.0001

McGinn, C. (1995). Consciousness and space. *Journal of Consciousness Studies*, 2(3), pp. 220-230. https://philpapers.org/rec/MCGCAS-5

McLaughlin, B. P., & Planer, R. J. (2014). The contributions of U.T. Place, H. Feigl, and J.J.C. Smart to the identity theory of consciousness. In A. Bailey (Ed.), Philosophy of mind: The key thinkers (pp. 103-128). Bloomsbury Academic.

Miguens, S. (2021). Animal brains and the work of words: Daniel Dennett, *Topoi* 41(3), pp. 599-607, https://philpapers.org/rec/MIGABA

Nagel, T. (1974). What is it like to be a bat? Philosophical Review 83(October), pp. 435-50. Doi: 10.2307/2183914

Nozick, Robert (1981). *Philosophical explanations*. Cambridge: Harvard University Press.

Olson, Eric T., "Personal Identity", The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/identity-personal/>.

Parfit, D. (1971). Personal Identity. *Philosophical Review* 80, pp.3-27, https://philpapers.org/rec/PARPI

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Pigliocci (2014), “Mind uploading: A philosophical counter-analysis”. In Russell Blackford & Damien Broderick (eds.), *Intelligence Unbound*. https://philpapers.org/rec/PIGMUA-2

Price, M. (1996). Should we expect to feel as if we understand consciousness? *Journal of Consciousness Studies* 3(4), pp. 303-12. https://philpapers.org/rec/PRISWE

Putnam, H. (1960) Minds and Machines. In Sidney Hook (ed.), Dimensions Of Mind: A Symposium. NY: NEW YORK University Press. pp. 138-164

Rescorla, Michael, "The Language of Thought Hypothesis", The Stanford Encyclopedia of Philosophy (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2023/entries/language-thought/>.

Rescorla, Michael, "The Computational Theory of Mind", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/computational-mind/>.

Robb, David, John Heil, and Sophie Gibb, "Mental Causation", The Stanford Encyclopedia of Philosophy (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/spr2023/entries/mental-causation/>.

Robinson, Howard, "Dualism", The Stanford Encyclopedia of Philosophy (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/spr2023/entries/dualism/>.

Searle, J. (1980). Minds, brains, and programs. *Behavioural and Brain Sciences* 3(3), pp. 417-57 https://philpapers.org/rec/SEAMBA

Schechter, E. (2018). *Self-consciousness and split-brains: The minds “I”*. Oxford University Press

Swinburne, R. (2018). The argument to the soul from partial brain transplants. *Philosophia Christi* 20(1), pp. 13 – 19

Swinburne, R. (2019). *Are we bodies or souls?* Oxford University Press.

VanRullen, R & Kanai, R. (2021). Deep learning and the global workspace theory. *Trends in Neurosciences* 44(9) https://doi.org/10.1016/j.tins.2021.04.005

Weatherson, Brian, "The Problem of the Many", The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/fall2023/entries/problem-of-many/

Winding M, Pedigo BD, Barnes CL, Patsolic HG, Park Y, Kazimiers T, Fushiki A, Andrade IV, Khandelwal A, Valdes-Aleman J, Li F, Randel N, Barsotti E, Correia A, Fetter RD, Hartenstein V, Priebe CE, Vogelstein JT, Cardona A, Zlatic M (2023. The connectome of an insect brain. Science. 379(6636) doi: 10.1126/science.add9330.

Williams, B. (1973). *The Makropulos case: Reflections on the tedium of immortality*. In John Martin Fischer (ed.) (1993) Stanford University Press, pp. 71-92