# On What Really Matters

## *Arguing in Favor of Parfit's View on Personal Identity*

Sebastian Kristoffer Strøm-Helbekkmo

Supervisor: Mette Kristine Hansen

University of Bergen

Institutt for filosofi og førstesemesterstudier

Master thesis in FILO350

Autumn 2023

# Goals

The central aim of this thesis is to present a thorough overview of the field of personal identity and provide argument for why I believe Derek Parfit's position on personal identity over time, as written in his book Reasons and Persons (1984), gives the best answer to the persistence question. More specifically, I will argue in favor of what I believe to be the five main claims he makes:

1. Personal Identity (PI) can be plausibly explained only by Reductionist accounts.
2. The best account of PI is based on psychological Connectedness and Continuity.
3. You can fully describe a Person by using completely impersonal language – i.e., an objective account.
4. The nature of PI cannot be determinate in all cases.
5. PI is not in fact what matters in matters where we are concerned about survival or morality – instead it is Relation R that we should care about.

In order to make this case in a clear and convincing way, I will also provide a general overview of the field of Personal Identity, encompassing what I believe to be the most relevant works and ideas.

I believe that this work is not only important in a semantic way, but that it has wide-reaching implications for how we approach ethics and morality in our own lives, as well as how we apply these concepts in wider society. In short, I believe that accepting the main points of Parfit's theory will force us to accept these further conclusions:

1. The "special concern" we hold for ourselves cannot be as strongly justified as most people intuitively believe. As such, most of us will have less reason to be especially worried or interested in our own futures.
2. What divides one person from another is also based on far more trivial factors than most of us accept. As such, we will have more reason to care about others,

even complete strangers, than most of us currently believe.

These conclusions, if wholeheartedly believed, will almost certainly make the world a better place. This fact, that the effect of believing such a theory would be good, cannot be a rational reason for belief in Parfit's theory in of itself. However, for those that come to see the truth of the theory based on its arguments (such as myself), this fact provides a strong reason for further study and development, as well as arguing in favor of the theory in the ongoing debate.

# Thesis Structure

This thesis is divided into four parts. The first part will serve as a foundation for the discussion that take place in the succeeding chapters. The main goal of this part is to explain the most important and/or most easily misunderstood concepts related to this subject matter. This part includes some reflections that are relevant to my conclusions at the end of the thesis, but the main purpose of this chapter is to establish the general meaning of the term/concept being examined. Once this has been done, I will move on to Part two, where I will be providing a general overview of the philosophical field of Personal Identity over time. In this part, I will be examining and explaining the various types of theories that are most the most influential. I will also be explaining some of their strengths and weaknesses as a theory, but as there is far too many theories to discuss in depth, I will only discuss certain aspects that I think are most relevant. Some parts of these theories and their arguments will be examined more thoroughly in part three of the thesis, where they come up naturally as I discuss Parfit's work in detail. In Part Three my goal is to both present and argue in favor for Parfit's theory of Personal Identity. I will first approach this goal by looking at the key aspects of Parfit's theory as he presents it in part 3 of his book Reasons and Persons. This should serve as an overview of the theory, highlighting the most important arguments and thoughts expressed, as well as looking at the multitude of various thought experiments Parfit uses continuously throughout his book. At this point, I will be primarily focused on explaining Parfit's own view as he presents it, but as we encounter various objections to this view (both those that Parfit directly addresses in the book, but also objections that could have been raised, or objections which have been raised at a later date) I will be contributing to the discussion with some of my own arguments and reflections when I think they add to the topic being discussed. Once the relevant theory has been thoroughly discussed I will briefly summarize the main points of the discussion, and how I think this leads to a conclusion in favor of Parfit's theory. Finally, in Part Four I will be briefly looking at what the consequences of this metaphysical world view would be. In particular, I will be examining both the practical effects of believing in this theory (both on a personal and a societal level), as well as the implications

of this theory on our current understanding of ethics and morality. In this part I will make some reference to thoughts presented in part four of Reasons and Persons Parfit (1984), but this part will mostly consist of my own independent reflections. Due to the scope of this subject, I will only touch on some of what I believe to be the most important considerations, but this part will leave open the door to further discussion, as this area is in my opinion at the very least as large as the metaphysics of PI, but considerably less clear and more controversial. I will suggest that, while Parfit's theory has answered most of the metaphysical questions relevant to PI, there is still much work to be done regarding the implications of this theory.

# Contents

# Part I

# Key Terms and Concepts

As previously stated in the chapter Thesis Structure, the primary objective I will be workings towards in Part One of this thesis is to establish an understanding of the important concepts and theories within this field, and how they relate to one another as well as to Parfit's theory. The first chapters in this part will be dedicated to defining, explaining, and reflecting upon some of the more important concepts that are used with the metaphysical field of Personal Identity over time. This section will provide both the "standard" definitions of the concepts, but also examples of how these concepts can be controversial and easily misunderstood, and how we should approach discussions using these concepts later on in the text.

Once the groundwork has been laid, the second section of this part will focus on providing a general overview of the field. This means that we will look at each of the major types of theories that are held either by (1) a notable number of people working within the field or (2) a significant number of people with little or no connection to the field. The reason for including (1) should be apparent, as it is almost literally the definition of an influential theory, but I think theories that fall within (2) should be explored as well. Even if a type of theory is not held in high regard within the field, if it is believed by a significant number of people, then it still might have a great effect on people's lives and society as a whole. Of course, when it comes to theories in category (2) it is important to note that people adhering to said theory might not in fact be aware that their beliefs coincide with this type of theory. Most ordinary people might quite simply hold a world view that is based on experience and that they find intuitively plausible, without having reflected on it intentionally. With this in mind, most types of theories that we will examine fall quite reasonably within both categories, though there are some notable exceptions that belong in the one or the other camp.

# Chapter 1

# On the Nature of Personhood

The first issue we will explore is that which is commonly referred to in philosophy as the *Identification* question (or problem). What is a person? Or what is required for a being to attain personhood? The notion of a *person* will most of the time correlate one-to-one with the concept *human being* in everyday life. Here I think it is important to establish how we approach concepts.

**Approaching Concepts**

1. **I assume (as most people do) that there exists an external world.**
2. **I have various thoughts and ideas regarding elements of this world.**
3. **I organize my thoughts and ideas into a linguistic structure, in this case English, both for the purposes of my own analytical process, but also to communicate these ideas to others.**

When we perform this type of inquiry, we must be clear what we are in fact analyzing. The purpose of this thesis is to examine the metaphysical concept of personal identity, that is to say, how this concept relates to the real world. This would be the link between point (1) and (2) as written above. By "examining this link" I am attempting to highlight the fact that any thought, idea, concept, etc. we have is in fact just a mental construct, which may or may not correlate in any degree with a real world. I will not argue or discuss these points in detail, as it would require a thesis of its own, but I believe it is necessary to be aware of this before we proceed with relevant arguments in regard to any metaphysical topic. It is however entirely when exploring a topic such as this to end up confusing the link between point (1) and (2) with the link between (2) and (3). In doing so we would no longer be studying metaphysics, but rather areas such as linguistics. In the first part of this thesis, we will look at this

connection when attempting to ascertain what a normal person believes such terms to mean, that is to say, which concepts, thoughts, and ideas this person correlates with certain words such as "person". This, I believe, can provide us with useful insights into what basic intuitions we carry before exploring a metaphysical topic in depth, such that we may have a better chance at avoiding confusions due to choice of language, as well as knowing which concepts and beliefs that are important to address. Now, having said this, we will start by looking at language and specifically how we use the word "person".

All language is constructed; a word's meaning and how it correlates to the world is defined by the people who use it. As such, I could choose to define the term "person" and the term "human being" as being identical terms referring to the same phenomenon exactly. If I did this, and there would be nothing logically wrong in doing so, we would be left with two separate words for identifying the same. This would simply be an inefficient way of using language. The reason we do have these two separate words, and that they should not be regarded as mere synonyms – which will be shown clearly as we look at the various thought experiments to follow – is that these words correspond to two separate conceptual beliefs we have about the world.

## 1.1 Person vs. Human Being

So, in what ways does our concept of personhood differ from the concept of a human being? A fully functioning adult human being is clearly a person, whereas an amoeba is clearly not. There is a whole spectrum of beings in between these modes of existence, but most of us do not consider any of them to be persons in the general sense. Does this mean there is a species barrier for our notion of personhood? Well, let's imagine encountering some alien species. This alien looks, acts, talks, thinks, and is in any meaningful way indistinguishable from a normal adult human. The only way we do in fact know that it is not a human, is that when tested, its DNA reveals a sufficiently different genome from our own. Furthermore, this alien shows emotion (and let's imagine we have some instrument that makes us able to tell the difference between genuine emotion and mere acting), feels pain and pleasure, and is not only conscious, but also self-conscious Is this being not a person?

At this point I think a lot of people might agree that surely, this must be a person in any meaningful way. But from this simple representation I would also assume that there would be a significant minority of people who would disagree with this assessment. I think skepticism is due to the framing of the example – taking the third person perspective, we might keep an intuitive belief that this being is lacking *something* that makes us persons. So, let us instead imagine a different scenario.

You are sitting in a doctor's office, waiting for the results after a routine check-up. Finally, the doctor returns, with a serious frown on his face. "I'm sorry to tell you this, but while checking your blood sample, we became aware of certain anomalies. After extensive testing, we have concluded that you are, in fact, not human." After making sure that there are no hidden cameras, and having overcome the initial shock, you might ask yourself whether this really matters? After all, you've always regarded yourself as a person. Should this discovery change your assumptions about yourself?

If at this point one is still not convinced that they themselves would in fact be a person, I think the only explanation could be one of two. Either the person believes that only a human being *can* in fact have the capacity of being a person, that is to say that the thought experiment is mistaken, because a non-human being would not be capable of having this experience – or he has simply interpreted these words as being equivalent to the same concept in his mind. The first belief is an understandable position to hold based on experience, yet all the evidence points to the fact that there is nothing inherently unique with human DNA, and there is no reason to assume that a non-human would not be capable of having the same type of experiences. Even if this was wrong, however, and only human beings *could* be persons, it would only show that the quality required in order to attain personhood is only ever *found* in human beings. If we instead choose to reject the thought experiment based on the second belief, then we are simply saying that these two terms cover the same concept, and as such we have no use for the separation of these terms. Again, this is a defendable position, but I think that it goes against the intuitions of most people (best illustrated by the fact that we do actually have two separate terms, and we are in no hurry to get rid of either one), which we will see in more detail as we explore further thought experiments.

## 1.2 Persistence

The idea of the tele-transporter (or just Transporter for short) should be quite familiar to most science fiction fans, as well as a significant portion of movie-enjoyers in general. It is hard to say when the concept was first imagined, but it was first popularized to a large audience by its inclusion in the Star Trek series as it aired its pilot episode in 1966 (and continues, in various forms, to this day). For those not familiar with the show, the story surrounds the crew of the starship *Enterprise* as they go on various missions throughout (and sometimes; beyond) the galaxy. During many of these missions the crew would find themselves in need of leaving their starship to perform their duties in person on some planet, but rather than boarding some type of shuttle in order to safely reach the planet's surface, a character such as Kirk or Spock would simply step on to one of the ship's *transporter pads*, before

dematerializing and then momentarily (at least from the perspective of the audience) rematerialize at their target destination. This would, most of the time, work without issue. Any concern for using the transporter seems mostly to be correlated with the possibility that it might malfunction in some way, not with the concept itself. Is this a reasonable way of looking at it? What actually happens when a person steps onto the pad?

At this point, I think various people's intuitions might differ, and so we will look at each perspective in turn. First, some of us might say that surely this type of technology is not really an alternative for travel, rather it is a way of committing suicide in a discreet manner. To everyone else it appears you go on living, but as a matter of fact, you were killed by the transporter and then a convincing copy took your place (or rather, your *role*, as it would be placed at the target destination). This seems a quite reasonable perspective. After all, a person being dematerialized does surely seem to be equivalent to death (even if such a thing does not occur in quite this way in the world as we know it). Although no person has ever been actually dematerialized as described in this case, the closest would probably be a person unfortunate enough to be at the center of a sufficiently powerful explosion, such as the atomic blast of Nagasaki and Hiroshima, or the eruption of mount Krakatoa. In these cases, there were real people who came about as close to being dematerialized as might be possible. And in these cases, it seems intuitively plausible that these persons did in fact die, rather than take part in some alternative way of travel. We might say that having one's entire body obliterated is consistent with death, and this would arguably be the best description of what occurs in the case of the transporter.

However, as mentioned earlier we might also make the opposite conclusion in regard to the transporter. As the name suggests, we might think that this really is just a very fast and efficient way of travel. How would it be reasonable to assume this, when it also seemed quite clear that this must surely be a form of death on the view we just examined. Well, I suggest it all has to do with the perspective of choice. On the previous view we were "looking" at the event from the outside, our eyes fixed on the body being destroyed, and how a separate replica of that very body was produced at the same time. We could clearly see that Kirk (or whomever stepped onto the pad) was in fact disintegrated and copied, rather than his body moving to another location – which would be what we expect in the case of traveling. The other way of looking at this example would be from an *introspective* point of view. What would it be *like* to step onto the transporter pad?

## 1.3 Qualitative and Numerical Identity

At this point we might briefly observe a fact about our concepts of identity. On the one hand, Tom on Earth and Tom on Mars are clearly identical, in so far as they are at the moment of the cloning completely inseparable when looked at in a vacuum. This is commonly referred to as qualitative identity. However, they are also clearly not *one* person (in the ordinary sense), in virtue of their being quite literally two separate people on each their own planet. Despite them being qualitatively identical, we can tell that they are numerically different. This is what we refer to as (you guessed it) numerical identity.

These concepts tie in closely with our own intuitions regarding our personal identity. As we see in so far as we have explored the transporter case, even if two things are perfectly qualitatively identical, they cannot be numerically identical in virtue of their being two separate things. Numerical identity is that identity which holds, and only holds, between a thing and itself. (Noonan & Curtis, 2018) If we take a static picture of the universe, it would be a fairly simple matter of picking out that which is numerically identical to itself. In this case, we could look at Tom on Earth standing on the transporter pad and identify him as Tom on Earth, just as we could do the same with Tom on Mars. However, once we introduce the dimension of time things change. Let's imagine two tables, completely qualitatively identical to one another, yet numerically distinct. Let's call them table A and B. Then I might bring a bucket of paint and use it to color table B completely blue. Now, table A and B are no longer qualitatively identical, and furthermore, table B is no longer qualitatively identical to itself at the start of this example. Even so, we still refer to it as table B – we recognize that it is numerically the same table. On what basis do we make this assumption?

At this point we will temporarily leave our two Tom's at their transport pads in order to explore this idea further from a different perspective. This time we will look at a case far older and much less sci-fi than the one we have been preoccupied with so far.

### The Ship of Theseus

Our next journey takes us from the world of Star Trek in the far future, to Greece in the ancient past. The Ship of Theseus, as it is known today, played a major role in the founding mythos of Athens. According to the legends, after having slayed King Minos' minotaur, Theseus (who founded Athens in these myths) rescued the children of Athens and then escaped on a ship headed for the island of Delos. (Levin, 2019) According to the ancient Greek historian and philosopher (to name a few of the many roles he performed) Plutarch, this ship was maintained and preserved for

generations to come, the shipwrights replacing old planks and rotting timbers as necessary. The ship itself, or the story of it at least, became a symbol of change and growth for Plutarch and his contemporary philosophers. The questions asked are closely related to the questions we have been exploring so far. Generations after Theseus escaped to Delos, is it still the same ship? If at this point every single original plank and piece of timber had been removed so that there was none of the original matter present to date, how could we explain our intuition that it is indeed the very same ship that Theseus sailed all those years past?

First, we might point out that it is nearly identical to the original. In so far as the work that had been done on it served only to replace the parts that were being worn out, and no new additions or renovations had been made. As we have seen, however, this is does not provide the right kind of identity relation. Our two tables, or our two Tom's, might be qualitatively identical, but not numerically identical, and as such this is not sufficient. Furthermore, even if the ship had received numerous upgrades, such as a different type of mast or a crane, we would still probably consider it to be the same ship. How could adding a mast change that fact? It does not appear that being qualitatively identical is a necessary condition for the right type of identity relation. Even though the butterfly and the larvae are vastly different in their qualitative identity, it seems apparent that the butterfly and the larvae are just various life-stages of a single organism.

I assume that at this point most of us still has a very strong intuitive belief that it is the same ship, despite the fact that the ship as it stands now contains none of the original matter, and we have concluded that its qualitative relationship is not sufficient in justifying this belief. What we are left with is what Parfit calls the "Standard view". He assumes, and I agree, that the reason we still believe this ship to be the very same that Theseus sailed on is due to the fact that the ship maintains the right kind of physical continuity. Now, as with many, if not most metaphysical beliefs and intuitions, most people will not consciously be aware of their belief in the manner that it will here be described, and yet it seems to be a nearly universal belief. The physical continuity consists of the ship tracing a continuous "spatio-temporal path" (Parfit, 1984, p.182) What this means, is that we can trace a figurative line from any point in space and time (or space-time, to be more specific) in which the object in question exists, to another point in space-time in which we believe the same object to still exist. If (1) at every point along this line there exists an object such as the one we observed at the starting point, and that (2) at every point this objects existence was immediately caused by the object at the preceding point, and that these two conditions are met along the entirety of the line all the way to the object in question, then the object in question is numerically identical with the object at the start of the line. This standard view describes the conditions of what we can call the *physical continuity* of an object – that which makes a physical object the

one and same over a period of time.

So, when we look at the ship laying at port in front of us, and we believe it to be the same ship as Theseus sailed on, it is for this reason. The ship traces a continuous spatio-temporal path (if we believe the myth) from the time of Theseus until the point in time in which Plutarch questions its identity, and along every point in this path a ship exists (in fact, a very similar ship), and the ship's existence was caused by the existence of a very similar ship at the preceding point on the path.

#### Billiard Balls

I reckon at this point that the standard view seems like both a plausible explanation for numerical identity as well as an accurate representation of how we in fact decide on the identity of a given object in our own lives, though I would assume most of us do not state in our internal voice that this is what we do, nor that most of us have even consciously reflected on the fact that this is what we are doing when identifying objects. Another observation we can make about the standard view as presented so far is that the object at the start of the spatio-temporal path is in this case qualitatively very similar to the object at the end of the path. Whereas we have so far been using the ship of Theseus for as our case study, Parfit uses a billiard ball for illustration. In order for the billiard ball in front of us to be the same billiard ball as the one we imagined (in his illustration of the standard view), there would have to be a spatio-temporal path between these two balls where at every point in between there exists a *billiard ball* which owes it existence to the billiard ball at the previous point. The emphasis here is on the fact that the in order for it to be the *same* billiard ball, then there must always have been a billiard ball in existence. This assumption leads to two conclusions.

First, we must revisit our assumptions regarding the relevance between qualitative identity and numerical identity. Whereas we previously concluded that qualitative identity had to be separated from the concept of numerical identity, it seems that they could not exist wholly independently. As we explored previously, the fact that a ship such as the ship of Theseus underwent qualitative changes (another coat of paint, different types of masts, more oars, etc.) would not break with our intuition that it was numerically the same ship, which is also consistent with the standard view. However, if we were to take the ship on land, turn it upside down, and secure it to the ground, we might not reach the same conclusion.

## 1.4 Necessary and Sufficient Conditions

Now, having provided arguments which I believe give us enough reason to dispute the claim that the term *person* is synonymous with human being, it seems we still

have a way to go in defining what the qualities that separate these terms are. In order to find these, we will attempt to find which conditions are *necessary* and/or *sufficient* for establishing a being's personhood. Necessary conditions are those that when absent from a being commits us to conclude that this being is not a person in the metaphysical sense[1]. A fairly non-controversial example of such a condition could be that of consciousness. If we built some sort of robot that could perfectly mimic any person, alive or dead, we might assume that it would still not be a person in a meaningful way if we also knew for certain that it was not conscious in any way. An important part of being a person seems to be tied to the fact that persons are *subjects of experience*. For something to be a person it must make sense for us to ask the question: "What is it like to be X?" If we ask this question of a rock, it seems immediately ridiculous, as we assume there is no such thing as having the experience of "being a rock" (not to be confused with the experience of being "The Rock", which I assume is pretty awesome). Any being that exists totally without consciousness seems to be excluded from personhood, and as such we can conclude that consciousness is a necessary condition. This conclusion does not mean that consciousness is a *sufficient* condition for personhood. If a being has the qualities that we deem to be sufficient for personhood, then we must claim that this being is a person. There are obvious reasons for why we might not wish to do so in this case.

In nature there are a number of species familiar to us which we do not presume to have the possibility for consciousness in any meaningful way, such as single-celled organisms, bacteria, jellyfish, etc. There are also species that we assume to have a conscious experience that is exceptionally close to our own in all the ways that matter, such as some of the great apes, dolphins, certain birds, etc. This latter category invites the possibility of extending our concept of personhood beyond the border of our own species in a very real sense, not merely as a hypothetical thought experiment. Between these two extremes on the scale of conscious capabilities there are a myriad of species which clearly seem to be conscious, and yet we would intuitively hesitate when asked to consider them as persons. Examples of this kind could be dogs, cats, deer, various types of fish, pigeons, etc. These animals show that, as far as our intuition is concerned, I think most of us would agree that consciousness is a necessary but not sufficient condition for personhood.

So, I think we could now return to our list of traits we generally identify as belonging to persons and attempt to adequately categorize them according to their sufficiency and/or necessity in establishing personal identity. In doing so, we will have to distinguish what traits are simply correlated with our notion of personhood with

[1]When I use the term person without further specification, it should be understood to refer to the metaphysical concept of the person.

those that do in fact create it. Here I would like to introduce an example made by Amy Kind in her book *Persons and Personal Identity* that I think perfectly makes the point:

"In Hollywood, in Madame Tussaud' s wax museum, there is a wax replica of the actor Johnny Depp. The replica looks remarkably like Johnny Depp himself, enough for you perhaps to be fooled for a moment or two, before the complete lack of motion made you suspicious. The wax replica of Johnny Depp is clearly not a person. But Johnny Depp clearly is." (Kind, 2015, p.19)

This example clearly shows that having a characteristic human appearance is not a sufficient condition for being a person, despite their being a strong correlation between this trait and who we normally regard as persons. Furthermore, we can also argue that a characteristic human appearance does not seem to be a necessary condition for personhood. Even if one is not convinced by the case against a human-centric conception of personhood, we can still see that not all humans have a "characteristic human appearance". If you ask a young child to draw a person, she will likely end up drawing some sort of stick figure, with two legs, two arms, a big round head with eyes and a wide smile. This, while being a somewhat crude representation, succeeds in capturing some of the defining characteristics of most human's appearance. There are however a serious minority of people whom, either at birth or due to some disease or accident, lack in multiple limbs or have seriously stunted growth. Some people, for similar reasons, lack eyes or lips or even major facial features at all (although this is significantly rarer). Despite not having a characteristic human appearance, if this is the only way they differ from the average person then their personhood is not in doubt. Even if we were just our heads being kept alive in a jar, like depicted in Matt Groening's Futurama, then we would likely consider ourselves to be persons, nonetheless.

This could lead to further questions, such as whether a person needs to have a physical presence at all? For example, could an advanced operating system be a person? These kinds of questions are not so clear cut. Some people may have very strong intuitions in favor of either answer, whereas others may have no strong inclination either way. This, I reckon, has to do with this being such a foreign concept to the kinds of problems our brains have evolved to solve, that it may require some mental gymnastics to proceed clearly.

### The House of Theseus

We might call this new structure the "house of Theseus", as it would now serve this function. Would it still be the ship of Theseus? Here intuitions will differ, but now an affirmative answer to the question seems less plausible, after all, how could it be the same *ship* if it is not in fact a ship at all. This would break one of the

conditions (1) of the standard view, that there no longer exists an object along the spatio-temporal path such as the one we started with. This would imply that the changes that have occurred are not superficial, but substantial enough to say that the thing has not *changed*, rather it has *seized to be*, and is replaced by something else. This shows the relevance of the type of inquiry we made earlier in this thesis, regarding necessary and sufficient conditions. In that case we looked at the necessary and sufficient conditions for personhood in specific, but the same concept applies for all manner of objects which we choose to define. In this case, we would say that at some point along the spatio-temporal path we crossed a threshold where certain qualities associated with our definition of a ship seized to exist to such a degree that the necessary conditions for "ship-hood" were no longer fulfilled.

## 1.5 Temporary Loss of Necessary Conditions

Still, we might make a reasonable case could be made on the basis of our current understanding of numerical identity, that it is in fact still the same ship. After all, at this point we are simply discussing what the necessary and sufficient conditions of ship-hood are. A ship's purpose is generally to traverse seas and waterways, and while the House of Theseus is currently incapable of doing so, it would not take significant changes (removing that which secures her to the ground and turning her over) to make her seaworthy. In fact, even though a ship's main purpose is to traverse water, the ability to do so might not be regarded as a necessary condition *at all times*. If a ship sails through a severe storm, for example, it may become sufficiently damaged so that it could not safely perform its sailing duties and would have to spend time at a drydock receiving the necessary repairs. Still, we would clearly not think that this is no longer a ship on account of this, simply a damaged ship that is in need of repairs before being returned to service. This seems to imply that a quality which we regard as being a necessary condition for objective identity, such as the quality that a ship needs to be seaworthy, may not be a necessary condition at every point in an object's spatio-temporal path. So, to what degree does the necessary condition apply?

If a ship is never, at any point in time, able to traverse bodies of water then it surely cannot be a ship in the normal use of the word. If the object has not yet been able to achieve seaworthiness, then it seems that it cannot yet be regarded as a ship, but rather a "hull that will soon become a ship" of some sort. So, we can state that the object has to become seaworthy for it to be regarded as a ship. What happens once this has been achieved? Will it retain its ship-hood indefinitely? As we have already seen, having temporarily lost this quality does not mean it immediately ceases to be a ship, but what if it were to be lost completely?

It might be hard to define a clear boundary between a temporary loss and a permanent one. In the case of a ship, at what point does a damaged vessel cease to have the opportunity to be repaired and is instead permanently disabled? The first example that comes to mind will probably be a sunken ship. At this point, we usually stop referring to object as a ship and instead use the term shipwreck. A shipwreck is the remains of the parts that constituted a ship at a previous point in time but is no longer capable of meeting the necessary conditions that made it a ship prior to its sinking. Another example might be a ship that has served for a long time and has become obsolete. At this point the ship might be brought to the shipyard in order to be scrapped for its useful parts and to break down its valuable metals, whereupon we would say that there is no longer any ship, but rather just the various parts that once constituted a ship. It seems clear that whether a quality is permanently or temporarily suspended makes a difference in our identification of objects, but this still leaves open further questions regarding the boundaries of these concepts.

**Shipwreck Case**

How is it that the shipwreck is permanently disabled whereas the damaged ship is not? In practical terms, it would simply be so expensive and challenging to raise and repair such an object, that it simply would not be done, and so for our normal use of language it is a fairly clear difference. But say we were to raise the wreck, and proceed to repair the damage sustained and replace the parts that were beyond simple repairs. This might be a larger repair than the one performed to the ship that was simply damaged, but it seems to be essentially the same process just with at a larger scale. Once the repairs had finished, we would be left with a fully functional ship. What are we to claim about the identity of this ship?

If we look at it from the standpoint of the standard view, its spatio-temporal path would look nearly identical to that of the damaged ship, the only significant difference being that at one point we claim that is ceased to be a ship. If this is true, then it could not possibly be the same ship on the standard view, but as we have seen so far, this interpretation relies upon the distinction we made between the damaged ship and the shipwreck, which again relied on our notion of *temporary* vs. *permanent* disability. If our idea that the object ceased to be a ship due to it being permanently disabled was correct, then we must also claim that what we previously described as a series of repairs were in fact substantially different from the repairs we described on the damaged ship, that instead of repairs it was a process of building a completely new ship simply based on the skeleton of the shipwreck we raised from the ocean floor. Is this a reasonable conclusion?

**The Scrapped-Ship Case**

What if we were to attempt to rebuild the ship that was scrapped for parts? Here it seems that the task is more significant. With the shipwreck we had at least a clear structure to restore, whereas in this case all we have is a series of various components that once constituted a ship.

Here we might quickly observe that the standard view on physical continuity can also apply in scenarios where we are not just trying to establish whether a present object is numerically identical with an object in the past but can also be used to explain identity relations between various objects. The claim that a bunch of "various components" at some time "constituted a ship" can be verified using the standard view. First, the various components we see today must hold the right kind of relation (relations 1 to 3 on the standard view) to the relevant component we believe constituted a ship in the past, as we have explained previously. If this is so, then we could say our claim is correct if this object did form part of an object that fulfills the necessary and sufficient conditions for qualifying as a ship.

So, returning to the task of restoring the ship that has been scrapped for parts. Here it will be more intuitive that what we are doing is not repairing the old ship (as it appears to not exist at all anymore), but rather building a new ship by recycling the parts that once constituted the old one. Are we able to make a reasonable defense of this intuition? Again, we seem forced to rely upon the same argument. In order to make a claim about the identity relation of the ship we just assembled to the ship that was previously scrapped, we must either be able to trace the right kind of spatio-temporal path between the two ships. For this to be the "right" kind of path, the path must be in accordance with the requirements stated previously. As such there must be a ship at each point along the path that was caused by the existence of a ship at the previous point on the path. Here, if we follow our first intuitions, we would likely conclude that this chain of relations is broken due to the ship ceasing to exist upon being scrapped. As such, if we believe the standard view to be the correct way of approaching identity relations such as this, then we will have good cause to believe that the best way of describing this case is the one where we simply build a new ship out of the parts of the old, and that the identity relation between the two ships is only in regards to the parts they shared.

However, we might question whether all the elements of this argument have been correct. First, we might question whether or not the chain of identity relation is in fact broken in the way we assumed. Our reason for thinking that it was is based on our claim that the original ship ceased to be once it was scrapped. However, as we explored previously, our argument for the ship ceasing to exist is based on the fact that it suffered a permanent loss of its ability to fulfill the necessary conditions of ship-hood. This, in turn, was based on our intuitive notion that largely seems

connected to the way things practically work out in the world as we know it. A ship that is scrapped or sinks to the bottom is usually not repair or restored, and so we are usually not challenged on this intuition regarding the permanent destruction of the object. However, when we do perform such repairs and rebuilds, the ground for our intuition becomes less clear. If the reason that the ship is not the same after its parts is reassembled is the fact that the ship ceased to be once it was scrapped; and the reason, we say it "ceased to be" rather than simply being temporarily disabled is based on our assumption that its disability was of a permanent nature; then we seem to be involved in a circular argument.

## 1.6 Circular Reasoning

There are many ways in which a statement, argument, or definition, can be circular. The smaller the circle the more obvious it becomes. If you ask me what a ship is, and I explain that it is a ship, I will have said something truthful, but also completely irrelevant as the statement does not bring any new information. Normally, in order to actually explain a term, we will have to use other notions and terms that are familiar to the person(s) we are talking to. The fact that the first statement is useless is in this case due to the fact that we are simply placing the same term on either side of our definition. This is the smallest kind of circle, in fact it is hard to even call it circular, as it is simply self-referring; there are no other points of reference other than the point we started at.

A larger kind of circle may be more difficult to spot and may require a step-by-step breakdown. A good example of this type of circular reasoning is that of the dictionary definition of language, which Amy Kind presents as follows: First she presents a definition of the term, which states that *language* is a "body of words and the systems for their use common to a people who are of the same community or nation, the same geographical area, or the same cultural tradition." (Kind, 2015, p.39) She then asks us to examine this definition. Is it circular? It does not initially appear to be so. However, once we then examine what the definition of "word" is, we find this definition: "a unit of language, consisting of one or more spoken sounds or their written representation, that functions as a principal carrier of meaning." Here it becomes obvious that these definitions are in fact circular, as the definition of one term is explained by the other, and vice versa. This would be the same as if we were to explain what a ship is by saying "it's a type of boat", and when asked what a boat is we would say "it's a type of ship". This type of circularity does add *some* degree of useful information to the statement, namely in confirming that these two terms stand in relation to each other, though lacking in any meaningful degree of detail the way it is currently stated. In the case Kind presents we learn that words are smaller pieces that together make up language, so although we have not necessarily

been given an adequate definition for the *substance* of the terms themselves, we can gather some important facts about their relations to one another. If we were to say that a ship is akin to a "large boat", or that a boat is likewise similar to a "small ship", we would add more useful relational information, despite the core definition still being circular in nature.

The argument we made in order to defend our intuition regarding the identity of the scrapped ship seems to be relying on this secondary type of circle at its core. While it does not appear that our premises are circular when looked at alone; once we lay them out together it becomes apparent. If we are to structure the argument it would look something like this:

**Premise 1:** A ship that permanently loses a necessary condition for its ship-hood ceases to be a ship.

**Premise 2:** An object or a grouping of objects that used to be a ship, but is now permanently lacking a necessary condition for ship-hood cannot be repaired, as only something that exists can be repaired.

**Conclusion:** Therefore, we have to accept that anything that assembles a repair-process will in fact be the construction of a completely new ship, as their comes into existence a ship where previously there was none.

We see that premise (1) is built on the assumption that a core trait which forms a necessary condition for identity is permanently lost – i.e. it *cannot* be repaired. Likewise we see that premise (2) claims that it cannot be repaired based on our claim (1) that the identity of the object is permanently lost. Does this mean that our conclusion is wrong? Not necessarily. The conclusion might technically still be accurate, but by showing that it relies on the circularity of the argument which we have laid out so far, we have no reason to believe it as it stands, other than mere intuition. Proving that an argument is circular does not necessarily *disprove* the claim it supports, it simply removes the support. As such we could attempt to find new, reasonable arguments that can support the claim, if we find it intuitively plausible, but for now we will set it aside.

## 1.7 Reductionism vs. Non-Reductionism

Any theory that proposes a definition of PI needs to base their necessary and sufficient conditions on facts about what constitutes a person. The various main theoretical positions in the philosophy of personal identity can be allocated to one of two main camps of thought: they are either *reductionist* or *non-reductionist*.

Reductionism is a position in the philosophy of personal identity that asserts that

personal identity can be reduced to a collection of more basic, non-personal phenomena and facts. A reductionist theory of personal identity suggests that personal identity is not something above the physical and psychological components and events of a person. It states that there is no need to assume or claim the existence of a metaphysical *self* or *soul* that remains constant over time to explain personal identity. Instead, personal identity can be fully explained in terms of more basic entities and facts, such as the brain, body, and interrelated physical and mental events. Whether or not "Person A at time T1 is the same person as a Person B at a later time T2" is simply a matter of these simple facts and how they correlate. It is not necessary to use the language of personal identity in describing a person' s existence; a simple summary of all the relevant facts regarding this person's life would be sufficient in telling us all there is to know. Psychological continuity theories, physical continuity theories, animalism, and bundle theory are all examples of reductionist positions on personal identity. (Kind, 2015, pp.26-33)

Non-reductionism is a position in the philosophy of personal identity that asserts that personal identity is tied to a fundamental feature of reality that cannot be reduced to anything else. According to this view, personal identity is not merely a product of facts about psychological or physical states of our being, but rather is a fundamental aspect of our existence as persons. This may be something like a *soul*, as is commonly believed by billions of people, or it may be something like a *cartesian ego*, which has been a highly influential view since the days of Descartes, although it is no longer seen as favorably in the field today. In the case of a non-reductionist theory, what makes it true that "Person A at time A is the same person as Person B at time T2" is due to some irreducible property that both of these persons share. If Person A and Person B both have the same soul (which is something unchanging) then we can tell that they are in fact the same person; and we can apply the same logic to a theory relying on a cartesian ego or any other non-reducible fact. A non-reductionist theory may include some or all of the same facts that a reductionist theory relies on, but it must also include some fact or notion that are not reducible to physical or psychological properties. The fact that we remain the same persons over time despite of various changes in our physical and mental experiences is that there remains something that is unchanging, and this element is what makes us who we are. (Olson, 2002)

# Part II

# An Overview of the Field

In this chapter we will be examining the various categories of theories that are commonly held within the field of personal identity. As I explained earlier, the main element we are interested in when looking at these theories is to provide an answer to the question "What makes a person A at time T1 the same person as a person B at time T2?" In other words, we are looking at the *persistence conditions* provided by each of these theories.

As we discussed in the chapter on Reductionism and Non-Reductionism, all theories regarding personal identity over time fall either within a reductionist or a non-reductionist framework. I have chosen to start this overview by looking at the various types of non-reductionist theories before we finish by examining the categories of reductionist theories. I think it is natural to start with the non-reductionist category for two reasons: First, this makes sense in a chronological way, as most of the early theories of persistence were based on further facts and non-reducible elements, before gradually falling out of favor in place of reductionist theories. Secondly, although these theories are not as influential today as they were previously, they are also (I think) the more intuitive view, both for people outside but also within the field.

So far, we have been examining various cases in order to get a better grasp at the concept of identity. These cases have confronted various of our intuitions regarding both the identification question (what makes something a thing of category X) as well as the reidentification question (what makes something the same particular thing over time). We started by looking specifically at our intuitions regarding personal identity, dealing with various versions of the transporter case in order to figure out our core beliefs (whether conscious or unconscious) about the nature of persons and personal identity. Then, once we had arrived at various observations regarding these views, including multiple inconsistencies and challenges to the way many of us view these concepts, we proceeded by closely examining various cases regarding the identity of objects rather than persons. This was done in order to provide another perspective to identity, one that might be clearer as it is more removed and objective (pun not intended) in contrast to our primary discussion regarding persons, where our own subjective point of view will likely impact our intuitions in ways we might not be actively aware of. Using a varied set of hypothetical and real cases we examined what Derek Parfit calls the "standard view" of object identity, being essentially a theory that he assumes most people intuitively rely on when dealing with object identity. Having seen various issues, the standard view seems to cause, we will now transition from looking at object identity back to the primary topic of personal identity, applying the observations we made during our analysis of these object-oriented cases to cases regarding actual persons. Here we will see whether our same intuitions hold, and whether the views we have regarding objects can be reconciled with our views on persons.

# Chapter 2

# Theories of Personal Identity Over Time

## 2.1 Physical Continuity

The standard view, which we have been exploring in the previous chapters, is what we can call a *physicalist* view on identity. This should come as no surprise seeing as we have been dealing with our view on *objects*, which most people assume are purely physical in their nature, and thus there is no real alternative view. There are exceptions to this way of looking at the world, some people believe that spirits or other types of non-physical entities inhabit parts of the world, such as rivers and forests. This, however, is no longer the case for most people, and therefore we can assume that this type of physicalist theory of identity will be familiar in dealing with objects.

When it comes to personal identity, our intuitions will differ more strongly. It will not seem obvious for many people that identity is a concept based purely on facts about our physical nature. Basing personal identity on a physicalist theory does come with certain advantages, however, such as the fact that we can more easily reconcile our intuitions about how an object retains its identity over time with that of a person.

We will now examine some of the more prevalent types of physicalist theories of personal identity and look at how they can be used to answer the questions we have been asking so far.

First, we might consider what a theory of personal identity based on physical continuity might look like. As we explored with the ship of Theseus, according to the

standard view, an object is the same as an object at another time if the two objects are connected in the right way through a spatio-temporal path. What happens if we apply this same assumption to persons?

If we are to take this physical approach to identity, we must be aware of the fact that this excludes various of the common ways in which we might define a person. Concepts such as consciousness, psychology, personality, etc. will not be included in our definition, unless we claim they are purely physical phenomena. This does not mean, however, that we must claim that these concepts do not *exist* whatsoever (if it did, I believe we would have such a strong case against the physical approach that we might brush it off in a paragraph or two), as these concepts can related to various physical entities in a way such that they coincide with our conceptions of a person. At least, this can be done in theory. Whether or not it succeeds in this manner is part of what we will be examining in this chapter.

Physical continuity in the context of theories of personal identity refers to the idea that personal identity is grounded in the continuity of one' s physical body over time. There are some different variations of the physical continuity theory that we will look at one by one, but what they all have in common is that some physical element(s) that make you a person has to trace a continuous spatio-temporal path, similar to the standard view on object identity that we have been exploring. (Parfit, 1984, p.182)

The first and simplest version of this theory suggest that what makes you the same person over time is the persistence of your physical body. According to this view, as long as the same body persists through time, the person remains the same, regardless of changes in memory, personality, or other psychological states. This version of the theory would answer the persistence question by stating that "person A at time T1 is the same as person B at time T2 if and only if person B has a body that is physically continuous with person A's body". This means that personal identity persists as long as there is a continuous physical connection between a person's past and present self, usually constituted by the physical continuity of the body and/or the brain. Olson (2002) Physical continuity theories provide a clear and straightforward account of personal identity, which can be quite intuitively appealing.

The simplest way in which we can make a physical approach to personal identity is to view the identity of the person as being equal to the body. This we might call the *bodily* theory of personal identity. On the bodily theory of personal identity, a person remains the same person over time if and *only if* they have the same body over this period of time. (Olson, 2002)

What are the conditions for the persistence of the body? First, we might specify what we mean by the body in the case of identity. In common speech we might

signify a difference between the body and the head as being distinct (just as a text contains both a "body" and a "header"), separated conceptually somewhere around the neck despite being functionally inseparable. Even though we could probably think of some sci-fi cases which challenge our normal experiences, for the moment we will rely on our common conception of a person of having (or "being") both a body and a head. Another distinction that is sometimes made is that between the body and the brain, as in the body being all the matter that constitutes a human being apart from the brain. In this case, when we talk about the persistence condition of the body and how it relates to personal identity, all of the physical being is of relevance. As we explore various cases and challenges to the theory, we will specify if we are making any relevant distinction between certain parts of the body (such as "the brain"), but for now we will begin examining this idea with the assumption that the body as a whole is what matters.

As we begin this query, we must first conclude what the necessary conditions for being a body are. Most of us will probably have a fairly similar conception of what a human body looks like, but does this image correctly represent the essential elements? Let us imagine the life of an average person living today. This will provide a model by which we judge the intuitive appeal of our theory, as this average or "normal" life will be, almost by definition, what the existence of a person entails.

As we describe this life, we will have to make various relevant metaphysical assumptions. I will attempt to describe the life as closely to the way I think a normal person would describe it. Once this has been done, we will then begin to analyze the relevant parts of the life and see how our theory and our assumptions hold up.

The life of our person begins at the hospital, as a newly born infant held in its mother's arms. For the sake of this example, let's call this person Adam. At the moment of birth, Adam weighs around 3kg and is incapable of doing much other than screaming and crying. As Adam grows up, he eventually takes his first steps, and speaks his first words. Soon he starts going to kindergarten, where he learns to play and socialize with other kids his age, and he makes his first friends. More time passes, and Adam starts his journey through the school system, all the way from the first grade to the point he graduates from college, he continues to learn new skills; both as through his education as well as outside. He loses touch with old friends and gains new ones, he has his first romantic relationship and his first breakup, he plays sports, and he suffers multiple minor injuries and broken bones. Upon completion of his education, he heads out into the world and starts at his first full-time job. After a while he meets a girl called Eve, they fall in love, and together they start a family. Adam has a stable career, and together with his wife they raise their children together as they grow older. Eventually Adam retires and becomes a grandparent. Having lived a long life of some 80 years, Adam takes his final breath

in the same hospital where he took his first.

This admittedly very generic description of a life will nonetheless fit fairly well with a significant number of lives' having been lived in the last century or so, as well as millions of lives that are being lived today. As we will use this life as a baseline, the physical theory we will now test must be able to do one of two things; it must either be able to satisfyingly provide a coherent explanation for why Adam remains the same person across this span of eighty years, or it must claim that this is in fact not what the case describes, at which point the theory will have to provide an even more convincing account of events seeing as this would break with our common intuitions.

As we have seen, the same-body theory claims that Adam is the same person over this period of time only if Adam has the same body over this time period. This means that the body of the infant and the body of the 80-year-old has to be one and the same body. Of course, as we have explored previously, this does not mean that they have to be qualitatively identical, but rather that the existence of the 80-year-old body must trace the right kind of spatio-temporal path all the way back to the existence of the infant's body. In the case of Adam, as is the case for most people, we would have no trouble determining that our bodies do indeed trace the right kind of path. Our bodies change over time, but they do so slowly, and we have no trouble recognizing the body as it changes. If we were just to look at a picture of the infant and the 80-year-old, we would probably not be able to tell if they hold this connection, but were we then to be shown a continuous video following the infant as he grew up until the point he reaches the age of 80, we would have no problem ascertaining the relation.

Not every life lived happens to follow the exact same path as that of Adam, however, leading to more challenging examples. Every year millions of people are involved in car accidents. Some of these result in severe injuries for one or more of the people involved, including various cases where limbs have to be amputated, as well as cases where the victims suffer significant bodily disfigurement. This leads to a scenario where in a matter of seconds, we would not necessarily be able to recognize the person as having the same body they had previously, and yet we still clearly think of them as the same person. Is this a problem for the same-body theory? Not necessarily. Whether or not we are able to (in the ordinary sense of the word) recognize the continuity of the body is not what matter – it is the actual continuity that counts. Imagine having not seen your childhood friend for a decade. You might no longer be able to recognize them, but this is no reason to think they are no longer the same person. What matters in this case is that it is still the body of a living person, and that this body has traced the right kind of continuous spatio-temporal path. We might take this example further. The people who had to have limbs amputated

might choose to replace them with artificial prothesis. These can often serve as a close match functionally to the original limb (depending on the sophistication of the replacement), but will of course not be made of biological matter, but rather a composite of materials such as aluminum, various metals and plastics.

Does this make a relevant difference? Well, first we might obviously observe that we do not think these people to be any less of a person for having a prothesis. In fact, it seems that it would be absurd to think that a person with such an artificial limb would be any less of a person, considering we have already made the assertion that a person who is lacking one such limb is still wholly a person in the ways that matter – how could adding something in lead to a reduction in this case? It seem implausible that, unless the addition of such an element impeded some other function, this could be the case. We could extrapolate this type of case further. It is not only limbs that have to be replaced, sometimes a person suffers some type of organ failure and is need of a replacement. This, as of the time this thesis is being written, may include all the major organs: The heart, lungs, kidneys, liver, and skin. Replacements of this kind come in various forms; the ideal circumstance is usually to receive a donor organ from another human, preferably from a relative or someone closely matching oneself so that the organ is not rejected by the immune system, but when this is not possible it may be viable to use animal organs (such as the heart of a pig) or synthetic variants, though this field is still fairly new and the current capabilities of such replacements are limited.

It would not be hard to imagine that in the not-too-distant future the technology has advanced to such a point that all of these organs could be replaced by artificial replacements capable of functioning just as well or even better than the original organ in its role. As we have seen, replacing one such organ seems like it clearly doesn't change the facts about personal identity. So far, our general intuitions about the case and the same body-theory seem to correlate closely. But what if we were to take this type of case even further? What if we changed two organs? Or three? Or half the organs? Or all of them? Is it still the same body? We can go even further, and imagine that ALL of the body, apart from the brain, is replaced by artificial components. All of the organs, bone and muscles, completely replaced by artificial counterparts, so that all that remains is a brain encased in a composite body, like something akin to a T-1000 in the Terminator movies. Now it seems less intuitive to claim that it is the same body, and as such the same person that we started with. This example, as we can see, closely mirrors that of the ship of Theseus by gradual replacement of parts. What we experience in a normal life, such as that of Adam, is akin to the standard maintenance process we looked at with the ship. The cells in our bodies (with the exception of certain brain cells) are continuously dying and being replaced by new ones that fill the same role, just like the rotting planks of the ship is replaced. In the case of the terminator, we see the exact same process, except

the parts are replaced with non-organic matter. Should this make a difference? If we replaced the rotting planks on the ship of Theseus with a type of plank made from aluminum, but which served the exact same purpose, would this matter?

In this case, we might think that it does. For many people, this type of change would seem significant. In the case of the person, it might be hard to reconcile one's conception of a person as a human being with an organically human body with that of the artificial replacement. Likewise, our idea of the ship might (though we have not consciously stated it) subtly include ideas such as "wooden". If this is the case, then the replacements would constitute a break in the physical continuity of the objects in question.

In summary, physical continuity theories claim that the persistence of the same physical body (or brain) over time is the key factor in maintaining personal identity. These theories offer a somewhat intuitively plausible approach to understanding personal identity, but face challenges in addressing psychological changes, as well as both some types of radical and even gradual changes of the matter constituting the body.

## 2.2 The Soul Theory

The "Soul Theory" is perhaps the oldest type of personal identity theory that we will be discussing throughout this thesis. The concept of a soul should be familiar to everyone, and the posited theory is likewise quite simple and straightforward. Soul theories assume that what is essential to a person is that he or she has a *soul*, and that this soul is immortal and unchanging. (Kind, 2015, pp. 96-97) This view predates even the earliest written records and has developed independently in various different cultures across the globe. The view played an important part during the development of early European philosophy during the time of the ancient Greeks, being the primary view posited by philosophers of the time such as Plato and Socrates. Today it is no longer a popularly held view among philosophers but remain at the core of religious thought and doctrine in most of the major religions. This places this theory in an interesting spot, as it has relatively little influence within the field itself, but outside it holds massive sway, as literally *billions* of people adhere to religions which include this concept. This makes it probably the most polarized of the theories we will discuss in terms of influence within versus outside the field.

The soul theory, quite obviously, provides an answer the question "What makes a person A at time T1 the same person as a person B at time T2?" by claiming that this is so if and only if A and B both share the same soul. In plain language, adherents of the soul theory would say that you are the same person as your former

self because your former version of yourself and your current version of yourself are both carriers of the same soul. In other words, even though your body and psychology may have changed to varying degrees of similarity, your unchanging soul continued to exist in this body up until this day. This underpins the common way of speaking about ourselves, that "my body" and/or "my personality" may have changed, but "I" have remained – we talk about "my" body as something we may own and control, but that is separate from "me". This, according to the soul theorist, is what explains the difference between qualitative and numerical identity, as our numerical identity is determined by the unchanging soul (that is what we truly are), while our "possessions" such as the body may undergo qualitative changes.

This theory has the clear advantage of explaining personal identity by relying on what seems to be a quite intuitive way of thinking about ourselves. There are however a number of drawbacks, which have led to the theory losing its dominant position in the field. Kind raises concerns about the coherence of non-physical souls, questioning how non-physical entities could be distinguished from one another without matter to differentiate them. She points out that if souls are unchanging, they cannot rely on psychological properties for distinction, as these properties change over time (Kind, 2015, pp.95-97) To illustrate this point, we can imagine that two different people, Jon and Tom, perform a "soul swap". That is to say, the soul of Jon enters the body of Tom, and vice versa. How should we describe this case? Well, according to the soul theory we should conclude that, since Tom's soul is now in Jon's body, the *human being that* we have previously been calling Jon is now in fact Tom, and the human being that we have called Tom is now in fact Jon. Is this view plausible?

Well, if we imagine the soul as sort of a ghost or ball of light that literally flies into the body, our picture of this case seems to make sense. And I think that this type of imagery is in fact what we are naturally inclined to imagine when we think about this scenario without further reflection. However, what we imagine in this case is not representative of reality as described – the point being that the *soul* has to be something immaterial and unchanging. This means that there would be no actual way to distinguish between the human we call Tom being inhabited by "his own soul" or that of Jon's.

For all we know, "my" soul could be leap-frogging between a thousand people each second, and there would be no way for anyone to tell that this was happening. This, among many other challenges, have led to this view falling out of favor within the field.

### Cartesian Ego Theory

A similar, but not entirely overlapping view, is that which claims that a person is really a *cartesian ego*. This concept gets its name from Descartes, who through

his famous meditations claims that the only thing one can be certain about is the existence of one's *self*. A cartesian ego may be a soul, as Descartes himself thought to be the case, but it can also be a non-theological entity. (Olson, 2002)

The way we often think of ourselves correspond to this conception. We might imagine our self being "some entity" separate form the brain and body, but which controls everything. For the same reasons, the concept of a Cartesian ego matches with our strong intuitions regarding ourselves, but seems highly implausible when looked at more carefully. We will look at further arguments against these views in Part Three of this thesis, as we examine how Parfit argues in favor of a reductionist approach.

## 2.3 Memory Theory

Next, we will briefly examine what is often referred to as the "Memory Theory", not because it is highly influential in contemporary philosophy, but because it was an important stepping-stone in developing the psychological continuity theory, which we will look at next.

The memory theory, while probably having multiple origins, is largely credited to John Locke. The theory, as the name suggests, is based on continuity of *memory*.

If we are not convinced by the various versions of physical theories of personal identity, as I suspect we might be worried about the multitude of problems it seems to cause. Having explored the alter natives of personal identity based on facts about the purely physical nature of the body (brain included), we might next consider those other elements that we generally think are important to a person, such as personality, consciousness, abilities, memories, etc. These traits have in common that they are not physical objects, but rather play a part in what we might refer to as the mental aspect of our existence. In the context of personal identity, identity theories based on these types of elements are usually referred to as "psychological theories".

The first one we might look at is one I think is the easiest to understand, as it makes the smallest leap from the physical theories we have been exploring. This is the theory of personal identity based on memories, famously proposed by John Locke in his 1690 essay "*An Essay Concerning Human Understanding*". In this essay, Locke claims that experience-memories (memories of things you personally experienced) are sufficient for providing a criterion of personal identity. (Piccirillo, 2010) This is because, according to Locke, the nature of one's identity extends only as far as one's consciousness. He argues that this is so on the basis that a person, or a "self", is really a "thinking thing" with abilities such as reason and reflection, and these

qualities are always accompanied by consciousness.

Regarding this claim, Locke states that "as far as [a] consciousness can be extended backwards to any past action or thought, so far reaches the identity of that person; it is the same self now as it was then; and it is by the same self with this present one that now reflects on it, that that action was done" (Piccirillo, 2010).

This memory theory would propose that "A at time T1 is identical to B at some later time T2 if and only if there is continuity of experience memory between B and A." (Kind, 2015, p.36) On this view, being able to remember doing something means that you are the person that did it. This seems a quite reasonable claim to make, and intuitively plausible. There are some challenges to this claim, which we will explore in more detail, but first we will look at the claim that *is* immediately controversial. In order to maintain consistency with this first claim, Locke's memory theory also makes the claim that if you *cannot* remember doing something, then it could not be you who did it. (Parfit, 1984, p.184) This is a much less intuitively plausible claim to make for obvious reasons. After all, there are plenty of instances where we cannot remember having done something, and yet when we look at a recording of us doing it, or having a person who was there recount the events for us, we conclude that we must have done it. In these types of instances, it is important to examine how we react to this "new" information. If I cannot remember leaving dirty dishes in the sink, but my girlfriend tells me "It had to be you, I haven't been home before now" then my reaction will be a realization that my memory was flawed, and I had in fact left the dirty dishes in the sink, as opposed to claiming that "somebody else did it". So, seeing as Locke's memory theory has to fight this intuition from the start, does it provide a useful perspective on identity?

There are some elements which might attract us to a memory-based theory of personal identity. Having the ability for consciousness is clearly a necessary condition for personhood, and as such it seems reasonable that some sort of continuity of consciousness is required for maintaining personal identity over time. Here there can be two kinds of continuities in mind. First, we might have what I will refer to as "objective" continuity of consciousness. This is closest to how we normally use the term continuity; there being a series of conscious experiences following each other, one moment after the next. There is, on this definition, continuity of consciousness over a period of time, if at no point during this period consciousness ceases to be.

The other type of continuity is what I will refer to as "subjective" continuity of consciousness. This is the type of continuity the person *actually* experiences. Whether or not consciousness ceases for a period of time before returning will not (by definition) be experienced by the person losing consciousness. If a person in the middle of a thought is "frozen" in such a manner that all elements of his body

and mind are kept perfectly as they are in that moment with no change, we would assume that there is no conscious experience to be had. However, if we then were to "unfreeze" the person after some period of time (and let's imagine that there this is an absolutely perfect process with no side-effects), consciousness should resume from the point it left off. For the subject of this experiment, the experience would be no different to that one experiences in normal day-to-day life. The only way it would be different was if, while being frozen, external circumstances had changed, such that there would be a noticeable difference in the surrounding environment. If you were, for example, having a conversation with someone when frozen, and then while you remained frozen the other person moved to the other side of the room, it would appear to you as if the person suddenly and instantaneously teleported from one spot to the other.

This second kind of continuity, what I have referred to as *subjective* continuity of consciousness, is completely reliant upon individual memory. If one does not have the ability to remember or recall prior conscious experiences, then one does not have this type of continuity. As such, it is this kind of continuity that is central to Locke's theory of personal identity.

We can highlight the importance of this by imagining a case where this kind of continuity is broken. Imagine a normal day in your life. Seeing as I don't know what your life is like, I will just present it as a fairly average adult persons daily activity. It may start by getting out of bed and following your morning routine, such as showering and having breakfast. Then, you head to work or school, where you spend around 8 hours engaged in various activities, having social interactions, etc. After work you probably return home, eat dinner, maybe watch some TV and spend some time with the family.

If you are like most people, you will have a decent recollection of your day. Although your memory will by no means provide a perfect recollection of every detail in your conscious experience, for the most part you will remember the key experiences leading from one moment to the next, so that if you were asked you could provide an outline of the day from your perspective. But what if there was a major blind spot in your day? What if, from the moment you stepped into your place of work, to the moment you stepped out again, there was no memories at all? Instead, for the roughly 8 hours spent at the office, there would simply be nothing; your last memory being stepping into work, before then seemingly instantaneously stepping out again.

This is the premise of the critically acclaimed TV Series *Severance* (2022-), where you follow a group of people working at the mysterious company called "Lumon". The effect on the people working at Lumon is that effectively, they are divided into two persons, or at least this what the experience is like for them. When speaking about themselves, the employees refer to themselves as "innies" or "outies" based on

whether they are currently inside of Lumon (and thus only have work-memories) or outside of Lumon (and therefore have no memories of work). These different versions of the same people maintain for the most part their personalities but due to the difference in environmental factors and their perception of the world around them, these personalities evolve and are expressed at times somewhat differently.

If we base our view of personal identity on memory theories such as the one Locke argues in favor of, then we might conclude that these two versions, the innie and outie, are in fact separate persons. We can see that the innie has a continuous chain of memories which only cover the experiences the human being had while working inside the company. For the innie, this set of memories correspond to their entire life. They know nothing about the outside world, or what kind of person their outie is, apart from clues on their bodies and assumptions they can make based on their own personality. The same goes for the outies, living what we might consider a relatively normal life, apart from a roughly eight-hour gap in their day, from the time they enter work and until they leave. What they do, and who they are at work is completely foreign to them, apart from what they might be told by those in charge of Lumon. For the innie and the outie, their experience might best be described as two separate persons sharing a body.

Whereas memory theory does provide a quite intuitive way of addressing personal identity, as it concerns itself with the *actual* conscious experience, we have of living our lives, it suffers from various inconsistencies that weaken it significantly. Today in contemporary philosophy, it has largely been replaced by various psychological theories (Kind, 2015, p.43) and we will examine these weaknesses in more detail as we look at these theories in general, as well as Parfit's theory in specific.

## 2.4 Psychological Continuity

The psychological continuity theory attempts to describe personal identity in terms of psychological connections. In contemporary discussions, this theory of personal identity has largely replaced earlier theories such as one's based solely on continuity of consciousness and/or memory, which we discussed in the previous chapter. The psychological continuity theory asserts that personal identity is based on the continuity of psychological states or experiences over time. When answering our question "What makes a person A at time T1 the same person as a person B at time T2?", a psychological continuity theory would claim that person A and person B are the same if there is *sufficient* psychological continuity between these people. It posits that an individual at one time is identical to an individual at a later time if there is continuity of psychology between them (Kind, 2015, p.43).

In other words, if someone's current psychological state can be traced back to earlier

states, then they are the same person. The theory emphasizes that the connections that run through an individual's life may encompass elements such as beliefs, desires, character traits, and habits – and that these connections are the essence that makes up the person's identity. Psychological continuity consists of overlapping chains of such connections over time, extending beyond mere memory to include a broader range of psychological features, such as, but not necessarily limited to, the features mentioned above.

The terms connections and continuity are essential to this theory, so I think we should look at them in a bit more detail. We might first look at the term psychological *connectedness*. When talking about connectedness, we mean the actual direct psychological connections that apply between a person at two different times. These direct connections can be such things as memories, desires, intentions, etc. In the case of an ordinary person, there will be a large number of direct connections between that person and themselves on the previous day. Of course, this number will be considerably smaller than the number of direct connections between themselves only a minute ago, and much larger than the connections they have with themselves a year prior. Parfit says that if the "number of connections, over any day, is at least half the number of direct connections that hold, over every day, in the lives of nearly every actual person... there is what I call strong connectedness." (Parfit, 1984, p.185) Being "strongly connected" is an important definition. If we simply talk about a singular connection with some previous self, then this degree of connectedness might be so weak that there really is no way in which it affects the actual life of the person, so that it might as well be seen as two strangers. Sharing no or very few psychological connections with a person is about the same as not being that person.

Using the term strong connectedness, we can then make a more precise claim regarding the criterion of psychological continuity. Parfit defines: "There is psychological continuity if and only if there are overlapping chains of strong connectedness." (Parfit, 1984, p.185) For a person A today to be the same as some person B at a previous time, then A has to be psychologically continuous B. This explains why, even though we have very few direct psychological connections with ourselves as children, we can still claim to be the same persons as long as there has been this overlapping of strong connectedness throughout our lives.

There are however two primary versions of the psychological continuity theory. The first is the standard version, which we just described. The second versions states that this psychological continuity must only hold between in a one-one manner. When answering our question "What makes a person A at time T1 the same person as a person B at time T2?", a psychological continuity theory would claim that person A and person B are the same if there is *sufficient* psychological continuity between these people, and that this relation *only* holds between person A and B. It posits

that an individual at one time is identical to an individual at a later time if there is continuity of psychology between them, *and* this relation only holds between two people (Kind, 2015, p.43).

Why is it important to claim that for person A at one time and person B at a later time to be identical there must be sufficient continuity of psychology between them, *and* this specific relation must only hold between these two people. This latter point is added to the theory in order to solve any potential issues regarding address cases of duplication or reduplication where psychological continuity might branch into multiple future individuals. This is often called the *non-branching* psychological theory, which qualifies that a "Person A at time T1 is identical to an individual B at a later time T2 if there is non-branching psychological continuity between them" meaning that there is no other person who also stands in a relation of psychological continuity to the first individual (Kind, 2015, p.58).

This variation of the psychological continuity theory might not at first glance appear as an important qualification to the standard theory. In basically all real-life cases regarding real people and their identities, this relation always holds one to one. In order to see why this modification is important we need to look at hypothetical thought experiments, where the necessity quickly becomes apparent.

We can start by returning to the idea of the clone. As we explored previously in our section on numerical and qualitative identity, we could imagine a situation where a person is *perfectly* cloned, that is to say that at the moment of the cloning there is absolutely no qualitative difference between the two persons. Further, we can imagine that after the cloning has been completed, both the original person and the clone wake up in the same room, and we have no idea which one is which. In this case, we see why the non-branching criteria becomes necessary. The two persons (B1 and B2) both share the exact same level of psychological continuity with the "original" person (A). As such, we have no reason to believe that person B1 has any "greater" claim to the identity of persons A over B2, or vice versa, unless we wish to invoke a further fact or physical continuity. As we have already discussed, identity must be a transitive relation, and as persons B1 and B2 are different people, but with the same relation to person A, this relationship clearly cannot stand as identity. What if the cloning wasn't perfect? What if we could claim that while both resulting people share psychological continuity with person A, person B1 holds a strong degree of continuity with A than B2 does? This would not solve the problem as there still would be two persons with *sufficient* degrees of psychological continuity to person A. If there was only one such person after the cloning, we would claim that this person was the same person as A. How could it be the case that this person is no longer sufficiently continuous just because a totally different person shows *more* continuity. This is why the non-branching requirement is added, and we will discuss this more

in detail when looking at Parfit's version of the theory.

Part of what makes a person psychologically continuous is their memory. As we saw while discussing the memory theory, memory plays a big part in the experience of our actual lives, as well as our idea of our own personhood. However, there is one big issue with memory, and that is that, as Butler claims, it presupposes personal identity (Parfit, 1984, p.196). In order to get around this, psychological continuity theories often use memory in terms of quasi-memory, which includes apparent memories that are causally dependent on a past experience in the right kind of way (Kind, 2015, pp.43-44).

This is one of the ways theories of psychological continuity has evolved beyond the memory theory of Locke, using quasi-memory as a way of solving the circularity concern, which we looked at in the last chapter.

The psychological continuity theory is especially highlighted when considering cases where psychological continuity and physical continuity diverge. For example, in hypothetical scenarios such as brain transplants, avatars, or teleportation, these cases suggest that our intuitions about personal identity align more with psychological continuity than with physical continuity (Kind, 2015, pp.44-45)

However, psychological continuity theory also faces some challenges. One major challenge is the question of what constitutes a sufficient degree of psychological continuity to maintain personal identity over time. For example, if someone undergoes a radical change in personality or values, does that mean they are no longer the same person? Additionally, psychological continuity theory struggles to account for personal identity over very long periods of time, such as centuries or millennia. As Derek Parfit's theory of personal identity falls within the category of psychological continuity, we will be exploring these and other challenges in more detail further on in the text, as we take an in-depth look at Parfit's own version of a psychological continuity theory.

## 2.5 Animalism

Animalism is a type of theory that while being essentially reductionist and somewhat similar to the bodily theory, still rejects both the physicalist and psychological approach to personal identity. According to animalism, human beings are fundamentally biological organisms (animals), and personal identity over time is a matter of being the same living organism.

Given a human animal, X, existing at one time, T1, and something, Y, existing at a later time, T2, Y is identical with X if and only if Y continues the life processes previously undergone by X. (Blatti, 2014)

Animalism, closely related to the bodily theory, is a perspective in the philosophy of personal identity that emphasizes the importance of biological continuity. Unlike psychological theories that focus on mental states, memories, or consciousness, animalism centers on the idea that human identity is fundamentally tied to being a biological animal.

The core principle of animalism is that personal identity over time is maintained through biological continuity. Animalism argues that humans, like all animals, possess certain persistence conditions that are not defined in psychological terms. This is because not all animals, such as oysters or cockroaches, have developed psychological capacities. Therefore, according to animalism, the identity of an animal over time consists of a continuity of life-sustaining functions or biological continuity. In essence, the animalist perspective does not focus solely on a theory of personal identity over time but rather on a theory of animal identity over time, particularly concerning animals like humans. (Blatti, 2014)

## 2.6 Narrative Identity

Diverging from traditional reductionist and non-reductionist approaches, narrative identity theory posits that our identities are fundamentally shaped and defined by the stories we tell about ourselves.

Narrative identity is distinguished by its focus on the subjective construction of self-identity through storytelling. It proposes that individuals craft their identities by integrating their past experiences, present realities, and future aspirations into a cohesive narrative. This narrative is not static but evolves over time, influenced by new experiences, relationships, and reflections. Olson & Witt (2019)

One major criticism of narrative identity theory points out the lack of reliability and accuracy of personal narratives. Critics argue that narratives can be self-deceptive, or overly simplistic, potentially leading to a distorted sense of self. Furthermore, the theory faces challenges in addressing cases where narrative coherence is disrupted, such as in severe amnesia or dementia. Olson & Witt (2019)

## 2.7 Hybrid Theories

**Finally, hybrid theories** consist in various positions that combine or incorporate elements from multiple philosophical perspectives. These theories aim to overcome the limitations and challenges associated with purely psychological, bodily, or other singular approaches to personal identity.

One example of a hybrid theory is the psychological continuity theory combined

with bodily continuity. It suggests that personal identity is maintained through both psychological and bodily aspects. Psychological continuity, which involves the continuity of memories, beliefs, and personality traits, is seen as essential for personal identity. However, bodily continuity, such as physical characteristics and biological processes, is also considered significant for personal identity. In this view, personal identity is not solely dependent on one factor but emerges from the combination of psychological and bodily continuity. (Kind, 2015, pp.98-100)

An advantage of this type of approach is the obvious increase in nuance, allowing the theory to incorporate more elements that we find intuitively plausible. This can however be a double-edged sword. The fact that such a position relies on two or more distinct approaches to personal identity means that it also potentially carries twice as many problems and inconsistencies. As such, hybrid theories are usually seen in a less favorable way than the more fundamental theories they are based on.

# Part III

# Reasons, but mostly Persons

In this part we will be looking at the main theoretical work up for discussion in this thesis: *Reasons and Persons* by Derek Parfit. This part is divided into two sections, the first of which is a fairly simple and straightforward overview of the thoughts presented by Parfit, primarily in part 3 of the book (Parfit, 1984, pp.178-311), whereas the second part will be dedicated to analysis of various challenges to Parfit's theory. In the first section I will go through Parfit's main ideas, arguments, and thought experiments, discussing and explain what I believe to be the most important points he raises. The order in which I examine these topics will not always coincide with the order Parfit has chosen to explore in his book, and I think it is useful to quickly explore the way in which Parfit approaches the subject matter.

Parfit approaches the topic of personal identity in this book with one particular notion in his mind; namely that his theory, his way of looking at the subject matter, is highly unintuitive to most people. The purpose of his book is not simply to present an alternative view in the field of other metaphysicians, but also to *convince* a somewhat uninformed reader that this is in fact the correct view. With this task in mind, I think the way Parfit has chosen to explore the subject makes more sense. His approach mirrors a more natural way of jumping from idea, to thought experiment, to concept, to a counterexample, and back to another idea. This kind of jumping back and forth works well, in my opinion, as a way of guiding someone unfamiliar with the field through the various concepts. I also think it serves as a way of opening up a new door, so to speak, for someone that maybe be familiar with the subject matter, but that may be entrenched in another camp, so that step-by-step they are naturally led to Parfit's conclusions. As Parfit himself writes, while he is absolutely convinced that his Reductionist account must be true, he still expects that he "would never completely lose my intuitive belief in the Non-Reductionist view." (Parfit, 1984, p.251)

While I think this style of reflection works well in achieving Parfit's goals of essentially bypassing some of our ingrained assumptions, it does have some drawbacks. First, the way Parfit chooses to jump between ideas, thoughts, and concepts can make the overall thrust of the arguments and reasoning hard to follow. Whereas each individual leap is perfectly reasonable and even natural to follow, when a number of these leaps have been made, I think looking back can cause the reader difficulty in retracing these steps (personally, even after multiple re-readings I still felt this difficulty, although I'm sure others might disagree). I assume Parfit was aware of this effect because it seems one way of addressing these challenges is by use of multiple rounds of repetitions. After having explored some concept for a while, he will keep making various of these leaps touching on other subjects that connect in various ways, before then returning to the original concept, at which point he will often repeat much of what he has already stated.

While this style of writing serves as a convincing way of guiding the reader to the conclusion that Parfit has drawn, for the purpose of further work and analysis this is somewhat cumbersome. Therefore, in the following chapter I will attempt to break down what I think the main points are, but I will not be structuring my overview in the same manner as Parfit does. Instead, I will attempt to structure this section in the way that I think provides the easiest way to break down each argument and idea and how they correlate with one another.

Firstly, we will begin by looking at all the main claims that Parfit makes regarding Personal Identity. I think this makes it far simpler for the reader to see which arguments and ideas are connected, and for what purpose. With these conclusions in mind, we can then begin looking at how he builds up his case. This second step will be to examine Parfit's introduction and discussion of various key concepts for his theory. In a similar fashion to the first step, and for the same reasons, I will focus on simply addressing what Parfit claims regarding each concept and how they correlate. These two steps will establish a general overview of Parfit's case, which will then make it easier to discuss and reflect on specific arguments and concepts in more detail.

This leads to the second section of this part, where I will discuss the concepts and arguments that were introduced in the previous section. In this discussion, we will examine various other alternatives to Parfit's view, theories that either directly or indirectly challenge his position, as well as more general weaknesses and strength of the various claims he makes. It is in this section that I will provide the overall argument in favor of Parfit's theory, by explaining how all the various objections and alternatives are best explain by the reductionist view on personal identity over time, and more specifically why I think Parfit's own interpretation seems necessarily to be the most plausible candidate.

# Chapter 3

# Parfit's Theory of Personal Identity

In contemporary philosophy, few works have been as influential and provocative as Derek Parfit's *Reasons and Persons.* It is to this date one of the most cited philosophical works of the 20th century (Parfit, 1984). A cornerstone of the philosophy of personal identity, the book challenges traditional views about personal identity and its persistence over time.

Parfit holds a Reductionist view, arguing that personal identity over time can be explained solely through relations of psychological continuity and connectedness. This theory contests the notion of *further facts* about personal identity over and above the physical and psychological facts about a person and that there is no fundamental, deep fact of personal identity that persists over time.

However, whereas Parfit does define a version of the psychological theory of personal identity in his work, he also makes a larger claim. Central to Parfit' s thesis is the concept of *Relation R*, which stands for psychological connectedness and continuity with the right kind of cause. Whereas Parfit argues for a theory of personal identity based on psychological continuity, he also claims that it is Relation R, *not* personal identity, that matters for survival and moral responsibility. This controversial claim has been the subject of intense scrutiny, prompting debates about the implications of this view for ethics, accountability, and our sense of self.

Parfit' s philosophy also gives rise to an *impersonal* description of reality, where the boundaries between individual persons become blurred.

## 3.1 A Note on Thought Experiments

Some of these hypotheticals can be closely related to events and occurrences we might actually experience in our own lives, but many of them will be far outside the realm of normal experiences. We might imagine scenarios that range from that which is theoretically possible but not capable of occurring due to limitations such as current technological capabilities etc. all the way to scenarios which we can imagine, but we might reasonably assume that are not only practically impossible but in fact *theoretically* impossible as well.

Here we might choose to be skeptical regarding the value of scenarios that are outside not only the realm of what we personally might experience, but in fact outside that which we believe to be possible in accord with the laws of nature. After all, how could that which we *know* (at least "know" as far as reason can provide true knowledge) to be outside of reality tell us anything true or useful regarding the nature of our actual world and reality. And it is, I think, a reasonable skepticism to have, one which we will explore further later in the text, as we address specific examples of these kinds in more detail. However, I will here briefly provide a basis for why I believe hypothetical examples of this kind can still be of use to us.

Ever since childhood I've always been interested in astrophysics. I remember at one time I was watching a scientist speak about the physics of black holes, and he asked the audience to imagine what it would be like to pass over the event horizon[1]. Now, I do not remember the exact details of this thought experiment, but it illustrated various elements of astrophysics and how amazingly bizarre the nature of black holes are, and it did so in a way that made it somewhat understandable for an audience that did not have a PhD in physics. However, the hypothetical itself is deeply impossible. "What would it be like to pass the event horizon?" is like asking what it would be like to be a quark[2] - the gravitational pull would break even your atoms down, not to speak of the many other more technical reasons for which it would be impossible, which I will leave for the physicists to explain. Nonetheless, the thought experiment still helps to illustrate that which otherwise only be understandable through pure mathematics.

A more philosophically and down-to-earth (pardon the pun) example could be asking "If you could have whatever you wanted at the snap of your fingers, what would your life be like?". In this example you are given out right God-like powers of complete control over the universe, you can get anything, anyhow, anywhere at any time. In

[1]The event horizon is the border of the black hole from which the velocity required to escape its gravitational pull exceeds the speed of light i.e., nothing, not even light, can escape from within.

[2]A quark is an elementary particle, combining to form protons and neutrons which in turn form atoms.

doing so you can start to imagine the mountains of cash, the huge mansions, the luxurious yachts, the extravagant parties with beautiful celebrities and every other desire you have ever dreamt about. Then you imagine living this life for a year, 10 years, a hundred years, a thousand, and so on. The point of this thought experiment is to explore what it is you truly want. The fact that it is in fact impossible to have this kind of power does not change the fact that it provides a technique with which to reflect on one's own desires.

## 3.2 Not What We Believe Ourselves To Be

Parfit starts his discussion by presenting the view that he claims most people, whether they are in fact aware of it or not, believe about themselves. This view is not necessarily the kind of view that a philosopher might hold in developing his or her theory, but rather a mix of various intuitive beliefs, some of which might contradict each other.

This is the framing Derek Parfit chooses when he famously[3] presents his version of the transporter case in Parfit (1984).

"I press the button. As predicted, I lose and seem at once to regain consciousness, but in a different cubicle. Examining my new body, I find no change at all. Even the cut on my upper lip, from this morning' s shave, is still there. " (Parfit, 1984, p. 178-179)

Now, at this point in, as I will describe what's going on in Parfit's example, it can at times be deceptively simple to speak, write, and think in the way that is common to us in normal situations such as we encounter daily. I might be tempted to state that in this example "you" experience what Parfit describes, or I could say that the "person" who steps onto the pad has this experience. Would it actually be *you* that regains consciousness at the other cubicle? This is the core question we are trying to find an answer to, and so Parfit proceeds by altering the thought experiment in a way that probes our intuitions further, giving us more ideas to grapple with.

The experiment is rather long and detailed, so I will here quickly outline that which is most significant: You step onto the pad as you have done many times before. This time, however, after having pressed the button, nothing seems to happen. You are then informed that this is a new experimental type of transporter, one which rather than dematerializing you just simply creates the copy at your target destination (which in Parfit's example happens to be Mars). You are then allowed to speak to "yourself" vis-à-vis on a monitor by using satellite connection (like Zoom[4]).

[3] Famously for those who engage in scholarly work regarding this topic, at least.

[4] Which probably does a fair job of dating when this paper was written.

By changing the thought experiment in this manner Parfit has allowed us to temporarily suspend the question of whether it is you that appears at the other end. From "your" perspective, or in other words, the perspective of the person who stepped onto the pad in the first place, we can clearly see that it is not "me" who has travelled to Mars. By virtue of standing here on Earth this could clearly not be the case. With our own eyes, we see what is in every way an identical copy of ourselves standing on Mars, talking to us through the monitor in much the same way that we would expect a conversation with ourselves to go based on the experience we have of our own thought processes. Essentially what Parfit has done is to turn the transporter case into a cloning case – there is really no doubt, at least in terms of ordinary common sense and normal intuitions, that this has simply been a way of creating a clone. This allows us to the scenario in more detail before being forced to make a final conclusion.

What observations can we at this point make? First of all, our own situation appears to be unchanged, at least from our perspective as the person who stepped onto the pad in the first place. If we had not been told that the cloning had taken place, or if we had not been able to see our clone via the monitor, we might have assumed that the machine simply had no effect. This is what I think most of us would regard as "our" perspective in this case. There is however a second perspective, namely that of the "clone".

From the clone's perspective, the whole exercise would seem like everything that happened was simply a case of instantaneous travel. In fact, "the clone" in this case, would not think of himself as "the clone", but rather as "you", the person who stepped onto the pad in the first place. The experience would be that you stepped onto the pad, lost consciousness before immediately regaining it on a different pad on Mars. You could then speak to yourself on Earth. What would you think in this scenario? Well, you might have exactly the same thought as "you" on Earth did: I'm here, so the person I see via the monitor can't be me. Furthermore, from your personal perspective in this case, it would seem just as plausible that the person on Earth is the "clone" in this scenario. Keeping these observations in mind, we will now be looking at some of the most important claims Parfit makes in his theory. When relevant, we will return to this though experiment in order to illustrate the arguments he makes in this regard.

## 3.3 An Impersonal Description

One of Parfit's claims in his version of reductionism is that even "though persons exist, we could give a *complete* description of reality *without* claiming that persons exist." (Parfit, 1984, p.190). That is to say, if we imagine a scenario where we have

the god-like ability to know everything there is in the entire universe, down to the tiniest particles and every single possible phenomenon, and we were to write down a *total* description of the state of the universe (whether it be at a single point in time or over eons) then this description would not need the term *person* in order to achieve a full description.

Well, how could this be possible? Part of this claim includes the fact that persons do *in fact* exist, and so it seems quite contradictory that one could claim that X exists without stating that X in fact exists. This relatively superficial confusing can quickly be dispelled by a simple example. Parfit himself illustrates by example of the planet *Venus*, which is also known by some as *the Evening Star* or the *Morning Star*, depending on where in the world you live. If a total description includes the claim that *Venus* exists, then our description could be complete, even though we do not claim that the Evening Star exists (Parfit, 1984, p.190). This claim must obviously be true, as the object that is being described is the same whether you use name X or name Y for said object.

This principle can then be used to justify the reductionist claim Parfit makes regarding a total description of the universe not mentioning *persons*. As we have seen, Parfit's theory is a reductionist theory, and as a reductionist he believes that "... a person just is a particular brain and body, and a series of interrelated physical and mental events." (Parfit, 1971, p.190-191) If this belief in correct, and our description of the universe claims that there exists a particular brain and body, and a particular series of interrelated physical and mental events, and how these elements stand in relation to one another, then we have claimed that the particular person exists in an impersonal way. As with Venus and the Evening Star, a particular person and a particular series of events and objects are simply two ways of describing the same thing.

## 3.4 Indeterminacy and Empty Questions

Most people, as Parfit claims quite reasonably, believe their continued existence to be all or nothing. You either continue to exist wholly, or you cease to exist completely; there are no borderline cases or other alternatives. As a result of this belief, we also believe our identity must be determinate. If we ask the question "Am I about to die?" there must always be an answer, and the answer must simply be "Yes" or "No". This belief is perhaps one of the most fundamental beliefs we have about ourselves that guide our lives and the societies we live in. In fact, it is such a core concept in our conceptualization of ourselves in relation to the world, that most people would not even consider their being any other possibility.

Parfit challenges this notion with the claim that questions such as "Am I about to

die?" might not have an answer. He posits that, in the same way that a complete description of the universe does not need to include a reference to a *person*, we could know everything there is to know about the circumstances of our past, present, and future, and yet we might not have any clear answer to the question of whether we will die or go on living. In these types of cases, where we know everything there is to know, and yet cannot provide a direct answer to the question, Parfit calls the question *empty* (Parfit, 1984, p.192). In other words, the question is based on the false assumption that there are different possible alternatives to the outcome. In an empty question this is not the case; there is only one outcome, and any difference in answer we might give is simply a different description of the same outcome.

### *Clubs and Countries*

To illustrate this point, we can imagine the case of a sports club. In our scenario, there was once a sports club that existed, but was then closed down and disbanded. Many years later, various people decided to form a club with the same name and club rules. We might then ask, is this the very same club? Or is it merely a new club which is qualitatively the same?

Of course, in this scenario there could be many legitimate reasons for us to choose one answer over the other. There could be some inherit rule within the club itself or some administrative organization that states what would happen in case of a club being disbanded and how it could be reformed. Likewise, there could be original club members involved in the founding of the new club who could make claims about whether this is just a continuance of the previous club or not. However, for the sake of our thought experiment, we will imagine that there exists no such rule and no such people – we have no direct ruling in the case of how this club is to be labeled. How should we approach this case?

Well, in this scenario, the only facts we have to go on are the ones relating to the existence of the current and previous club. We can look at who the members are, how they interact with each other, what the rules are, etc. If we know all these interactions, then we know all there is to know in the case. We know exactly what the existence of the club entails, and yet it would not give us any answer to the question we asked. Whether it is the same club that has been reinstated, or a new club that is exactly similar to the old one, is simply not included in the facts of its existence.

This is because, as Parfit claims, whether or not a club is the same as some previous club can not be due to any *further fact*. By further fact he means that once we know about the rules, the members, and how they interact with one another, there is nothing else to know about the club. "The club" is just these facts in relation to

one another, there is no further fact which would solve our question about the club's identity (Parfit, 1984, p.216).

So, how should we answer the question? Seeing as we can know everything there is to know about the constituent parts that make the club, and still not have any knowledge about the identity of the club, the question is empty, and as such we may choose to give it an answer. In this case, we could say that the club is the same club, or that it is a new club. The important thing here is that either way, our answer is arbitrary. Whether we call it a continuation of the old club or a new club, *all* the facts remain the same. Our answer should therefore be given based solely on non-metaphysical assumptions, such as what answer would be most practical for us to use.

This applies to all sorts of identity-related concepts we encounter in our lives. The countries we live in; is Italy a continuation of the Roman Empire? Again, we see that even though we may now all the facts about some entity, there is no "further fact" which would give us a definitive answer regarding the identity question. Should this conclusion surprise us? I don't think it should.

### Sorites Paradox

I think we might see the reason why this is the case even clearer when examining what is often referred to as *Sorites paradox* or the *Paradox of the Heap*. In this ancient thought experiment, we start with what is clearly a "heap" of sand. We are then asked to remove a single grain from the heap. Is it still a heap of sand? It seems like the removal of a single grain could not possibly turn a heap into something less than a heap. We might think that it would be ridiculous for such a tiny change to alter our perception of the heap. And yet, we continue to remove sand, one grain at a time, until we are left with perhaps ten grains of sand. At this point it is clear that if we continue to remove grains of sand, even just one, it will become increasingly ridiculous to call what remains a heap. So, it seems we are left with a paradox: Either we say that the removal of a single grain cannot change the nature of the heap, at which point the heap will gradually be reduced to a single grain - or we say that at some point it ceases to be a heap. But how could we possibly choose? Can we really say that ten grains constitute a heap while nine are not?

It is here we can resolve the paradox by addressing the underlying fact: "We do not believe that each of these questions must have an answer. We know that the concept of a heap is vague with vague borderlines." (Parfit, 1984, p.207) Know that the concept of a heap is vague and superficial, we can *choose* to set a concrete border, at which point the heap ceases to be. The choice of where this border should be is then completely arbitrary, but this is no issue, as long as we can agree on the definition it makes no difference.

The point of these thought experiments is to show that when it comes to concepts such as clubs and heaps, we are all (or at least the vast majority of us) reductionists. We do not believe there is any further fact to the existence of these entities, that goes beyond their parts. As such, we are not confused by the fact that in certain cases, the question of these entities' identity may be an empty question. In such cases, we may choose to give this question an answer, and as such choose whether the club or the country is new or the same, but in each of these cases, this choice will be arbitrary.

That is not to say that there can *never be* any answer, or that *all* questions regarding identity are meaningless. This would be going a step too far. In the case of clubs, Parfit points out that there are legitimate questions that could be asked regarding the identity of a club. For example, whereas two qualitatively identical clubs at different points in time might be regarded as the same or different based on arbitrary decisions, two qualitatively identical clubs at *the same point* in time must be numerically different. Therefore, asking a person whether they belong to one or the other of the two qualitatively identical clubs is *not* an empty question; we could know the answer to this question, and it would be a meaningful answer (Parfit, 1984, p.216). In this case, a person belonging to either club A or club B would be two different possibilities.

Here we can see part of Parfit's attack on non-reductionist theories of personal identity. If we believe in the reductionist account of personal identity, we should also accept the same reasoning we have seen with regards to clubs and apply it to our concept of persons. This means that there could be cases in which questions about our identity, whether or not we will go on living, are empty questions.

### How There Could Have Been a Further Fact

Even though I think the case so far against the idea of a Further Fact is strong enough, we might take a moment to hammer the final nail in the coffin. After all, the idea itself is quite intimately tied to our intuitions regarding the subject matter, so it might be useful to put out any doubts.

If we accept the idea that identity could be the result of a further fact, how might we go about proving this theory? Parfit provides a colorful example of the type of evidence which might have convinced us: "A Japanese woman might claim to remember living a life as a Celtic hunter and warrior in the Bronze Age. On the basis of her apparent memories she might make many predictions which could be checked by archaeologists." (Parfit, 1984, p.202) If all of these predictions then turned out to be correct, and the archaeologists could see no other way that one could have known about these artefacts, we would have to assume that these were accurate quasi-memories of another person. Then, we can imagine that these types

of memories became common place, and that virtually all people alive had memories such as these of past lives.

If this was the case, and we had no other possible explanation for how these memories occurred, we might start to consider the possibility that there is something more to our psychology than simply the brain. If we could not discover any way in which these people are physically connected to the persons who's memories they can recall, we would have to assume that there was some purely mental carrier of memory. This would significantly strength the plausibility of the non-reductionist view, and could be used in favor of the idea of a Cartesian Ego, Soul, or some such entity. (Parfit, 1984, p.203)

Another such piece of evidence could have been that when there was any damage to the brain, psychological continuity would either hold fully or break completely. This would imply that there was either some critical physical element that was wholly responsible for our identity, or that there existed some non-physical entity which was in some way connected to our bodies.

Needless to say, there is no such evidence. Everything we know about our psychology points to a direct relation between it and some part of our brains. Seeing as there could conceptually have been evidence in favor of the non-reductionist view, and yet every piece of verifiable evidence points in the opposite direction, I think that it is beyond implausible to hold such a view. Although it might be tempting to justify our intuitive beliefs, we have no rational or empirical justification in this case.

### Tele-transporter II

If there is no further fact which enables personal identity, Parfit argues that this is the conclusion that we must accept. When applied to persons, we might again return to the example of the tele-transporter. In this case, where we have two different individuals, one on Earth and one on Mars, we can apply the same logic. If it is the case that these two individuals are qualitatively identical, this means that we might be unable to answer a question regarding their identity in relation to the original person. How would we justify this kind of assumption? If all there is to a person is simply the various physical and psychological elements and their various interrelations with each other and the outside world, then we could know all there is to know. In this case, we could know all about both of the resulting people, and still we would have no reason to conclude that one or the other person has a better claim to the continuity of the identity of the original person. As Parfit states: "If personal identity does not involve a further fact, we should not believe that there are here two different possibilities: that my Replica will be me, or that he will be someone else who is merely like me." (Parfit, 1984, p.217)

So, how could we best describe the outcome of this hypothetical case? As we have seen, we have no reason to claim that person A has a better claim to be the original person over person B, or vice versa. We could also not plausibly claim that *both* people are numerically identical to the original person, as this would break the transitivity of identity. Could we say that neither of the resulting persons are the original person?

This might, at least on the surface seem more plausible. There is however one problem. In this thought experiment we have assumed that both resulting persons are equally continuous with the original person, and in this case this continuity is very strong. As this is the case, if either of these people had existed *without* the existence of the other, then we would have to claim that they were the same person as the original. If the Original Persons steps onto the pad, and after the process has been completed, there is only one Resulting Person (whether on the same pad or on Mars), then we would conclude that these are the same person. And here we arrive at the problem. If either resulting person contains what is necessary for personal identity, how can the fact that someone else holds the same relation plausibly break this relation? As Parfit puts it "How could a double success be a failure?" (Parfit, 1984, p.228)

## 3.5 William's two plausible requirements that no Criterion of Identity can meet

These questions play a large part in the criticism philosopher Bernard Williams posit against Parfit and his version of psychological reductionism, which Parfit addresses throughout *Reasons and Persons*. First, we should briefly establish William's position in general. Williams believes in a non-reductionist view on personal identity. He claims that, as we have discussed in the chapter regarding psychological theories in general, identity must logically be a one-one relation. Seeing as the standard version of psychological continuity allows for this relation to hold between multiple people, such as one person in the past and two people existing simultaneously today, William concludes that this relation cannot stand for identity. (Parfit, 1984, pp.238-240)

These objections should be nothing new to us. These kinds of challenges are the reason why psychological continuity often carries a *non-branching* stipulation. Seeing as the non-branching version of psychological continuity is logically a one-one relation, it answers this objection. This is where our questions regarding our previous thought example come into relevance, as Williams also rejects the non-branching version of the theory. He does so based on two requirements he claims have to be met in order for a criterion of identity to be valid. These are as follows:

*Requirement*(1): Whether a future person will be me must depend only on the intrinsic features of the relation between us. It cannot depend on what happens to other people. (Parfit, 1984, p.239)

*Requirement*(2): Since personal identity has great significance, whether identity holds cannot depend on a trivial fact. (Parfit, 1984, p.239)

**Requirement One**

Let's start by looking at requirement one. Is this view plausible? We can return to the orginal thought experiment of the teletransporter. In this case, once I have been scanned on the pad I am then instantly dematerialized, before my blueprint is used to create an exact copy of my entire body on Mars. For the person waking up on Mars, they will have my memory of stepping onto the pad, before then immediately appearing on the pad on Mars. This person will believe they are me, and they will be fully psychologically continuous with me as I was when my blueprint was taken.

If we believe in the psychological criterion of personal identity, we should accept that this person is me. However, if the scientists in charge were to also create a second replica, we would have to apply the non-branching requirement, and thus claim that neither person is me. This breaks with requirement one. How could it be, Williams might ask, that whether or not you wake up on Mars could depend upon what happens to someone else? In Parfit's illustration, he imagines that the blueprint is sent to another group of scientist, this time on the moon Io. (Parfit, 1984, p.240) If the scientists ignore this blueprint, it will be me waking up on Mars; if they do not, I do not.

Does this example convincingly show that Requirement One is plausible? If it does not, we can add another layer to the thought experiment. Let us imagine that the blueprint is sent both to Mars and Io, but do to a malfunction in the machine on Io, that only the scientists on Mars make the replica. In this case, it would be me waking up on Mars. However, what if, one hour later, the scientists on Io finally fix their machine, and they produce another replica of my blueprint. Well, according to the non-branching requirement, I do not fulfill this requirement any longer, and as such I must cease to exist. "Though the people around me on Mars will not notice any change, at that moment a new person will come into existence in my brain and body." (Parfit, 1984, p.239)

At this point, if we accept the implications of this thought experiment, I think we must also accept that William's first requirement carries some significant weight. It seems completely implausible that questions about our identity could rely on facts about other people, potentially on the other side of the universe. We could all,

technically, without any awareness of the fact or any noticeable changes, cease to exist in this manner due to an alien species of mad scientists in a galaxy far, far away.

Before discussing the implications of William's second requirement, we should now look at Parfit's Relation R in more detail. Once we have done this, we can fully answer William's objections.

## 3.6 Relation R – Connectedness and Continuity

Out of all the various concepts Parfit uses and introduces in his text, *Relation R* is by far the most central to his theory. Before we can discuss *why* Relation R is so important, we should first briefly outline what is meant by this relation, and how it differs from the standard Psychological Criterion of Personal Identity.

### The Right Kind of Cause

Parfit defines Relation R as "psychological connectedness and/or continuity with the right kind of cause." (Parfit, 1984, p.193) As we have seen previously in the chapter on psychological continuity theories, this definition overlaps with Parfit's definition of Relations R. Parfit chooses to leave this definition vague enough so that the specific conditions of connectedness, continuity, and cause can all be discussed and modified as necessary. Whereas the theory relies on these elements being essential, it might not necessarily be possible for us to actually quantify each element and its exact value.

When discussing the "right kind of cause" for these relations to hold, Parfit defines it as one of three types: This cause can be *normal*, *reliable*, or *any cause*, depending on the version of the Psychological Criterion being considered (Parfit, 1984, pp.185-186). These correspond to what Parfit calls the *narrow* and two *wide* versions of the Psychological Criterion. We can quickly illustrate how these differ by some thought experiments. First, we can think of our ability to hear sound. For most of us, this ability comes as a result of various organic sensors and fluids in our ears being caused to vibrate by sound waves entering the inner ear. Then, the vibration of these sensors cause specific signals to be sent to certain parts of the brain, which then interprets these signals to create the experience of hearing particular sounds. What I have described in this case would the *normal cause* of hearing, and as such would correspond to the *narrow* version of the criterion.

For many people, whether due to age, injury, or sickness, their ability to hear is severely limited or even non-existent in the normal way. Luckily for many of these people, they have the opportunity to install various kinds of technological hearing

aides, such as cochlear implants, which could allow them to hear sounds again. In the case of the cochlear implant, their is a piece of the device that is surgically implanted into the brain of the person, while there is a special type of microphone that is placed in or around the outer ear. In this case, the microphone picks up the interference of the actual sound waves passing by, which are then converted into electrical signals which the implant directly sends into the part of the persons brain responsible for hearing. This is not the *normal* way of hearing, but it is for many people a *reliable* way of hearing sounds. Whereas part of this process is conducted by an artificial "organ", the process and effect is very much the same as the normal way of hearing.

Finally, on the *widest* version of the Psychological Criterion, we could have *any* cause. This is somewhat harder to describe in terms of sensory perceptions, so Parfit uses the example of an unreliable treatment for a disease. The specifics of the disease are irrelevant, but the functioning of the treatment is. In this case, when the treatment is applied, most of the time absolutely nothing changes - the disease remains. But, on a few sporadic occasions, the treatment fully cures the disease. (Parfit, 1984, pp.256-257) This treatment has the same effect as a reliable treatment, but only has this effect sporadically.

Does it matter which of these versions of the criterion we accept? In order to answer this questions, we must first explain the importance of Relation R in Parfit's theory.

### What Really Matters

As we have already discussed previously in this thesis, questions of personal identity are not merely interesting to us in an academic sense. Most people feel that their existence is a deep and fundamental fact and that the answer to questions like "Am I about to die or will I keep on living?" is extremely important. This view of our own existence is so fundamental in our normal way of thinking that it's hard to even realize that there are any possible alternatives. What Parfit does is challenge this notion – he claims that personal identity is *not* in fact what we should be concerned about in matters of survival. Instead, it is Relation R that is considered by Parfit to be what fundamentally matters, even when it does not provide personal identity.

Parfit states that Relation R is not the deep, separate fact of personal identity, but rather the relations that justify special concern for our own future or the future of those we care about. If we can be justified in regarding this special concern as having weight, then the rational reason for doing so would be based on Relation R, not Personal Identity. (Parfit, 1984, pp.275-279)

Now that we have established why Relation R is so important, we can look at what the fundamental difference between a Criterion of Personal Identity and Relation R is, and how both terms fit into Parfit's theory. Personal Identity according to the psychological continuity theory overlaps with Relation R on all points except one; that while Personal Identity is always a transitive relation, Relation R does not necessarily need to be (Parfit, 1984, p.185) What this means is that while a Criterion of Personal Identity needs to hold to a one-one form (as we have discussed in Part One), Relation R can hold in various ways, such as in cases where one person is R-related to two other people.

This is what we see in the case of the teletransporter, and various other similar thought experiments. In cases such as these, the Psychological Criterion of Personal Identity fails to hold due to the breach of the non-branching requirement. In spite of this, the original person is still R-related in exactly the same way to both people. It makes little to no difference whether you will be R-related to one or multiple future people, what matters is the intrinsic nature of the relation to each of these people. We can now return to William's second requirement for a criterion of personal identity.

### William's Second Requirement

*Requirement*(2): Since personal identity has great significance, whether identity holds cannot depend on a trivial fact. (Parfit, 1984, p.239)

When it comes to the Non-branching Psychological Criterion of Personal Identity, we can see that this criterion does not suffice to meet William's claim head-on. Whereas we can make the claim that the R-relation held between people is a significant fact, Parfit agrees that whether or not the same person *happens* to be R-related to *another* person as well must be a trivial fact at best. (Parfit, 1984, p.239) Therefore, if we believe in the assumption of the argument, we must deny that *any* replica that is made this way would in fact be the original person. If this is the case, I would not wake up on Mars, and the teletransporter would simply be a way of dying.

Williams claims that these two requirements cannot be met by any reductionist criterion, and as such we should abandon this position. How does Parfit answer this challenge? It is in this answer that we can see the purpose of Relation R in his theory. As we I have already explained, Parfit agrees that William's two requirements carry weight, and cannot simply be dismissed. In order to make the Psychological Criterion consistent with the transitivity of Identity, we needed to add the Non-Branching stipulation. This made the criterion logically a one-one relation, meaning that it could still serve as a criterion of personal identity. However, as Williams shows, this leads to the theory having to make conclusions based on non-intrinsic relations

and trivial facts. Both of these features result in what would intuitively seem like a reduction in the plausibility of the theory.

### Identity is Not What Matters

This is where Relation R comes into the picture. The reason the Psychological Criterion seems implausible is *not* because we are trying to force a reductionist theory onto a non-reductionist concept, but rather because we are intuitively biased towards giving the concept of personal identity *too much weight*. In the case of the teletransporter and the multiple replicas, our own intuitive notion of the importance of identity forces us to make a criterion that encompasses more than what is useful. That is to say, the reason our criterion of identity cannot meet William's two requirements is not due to a failing in the criterion, but rather that we think the concept of identity carries with it a level of importance that it does not warrant.

As we have already discussed, in cases such as these, if there is no further fact that establishes identity, a question regarding the persistence of a specific person might simply be *empty*. In cases such as this, we might choose to give an answer to the questions, but this would be based on arbitrary considerations, and so the answer would hold little value. What we *should* do, is accept that Personal Identity, in fact, is not what matters in these cases. This is how Parfit addresses William's Second Requirement; by attacking his premise. Williams presumes that personal identity as *great significance*, whereas Parfit does not. Instead, in matters of survival, Parfit claims that *Relation R* is what fundamentally matters. "Unlike identity, this relation cannot fail to hold because of a trivial difference in the facts. If this fails to hold, there is a deep difference in the facts." (Parfit, 1984, pp.242-243)

Parfit has here sidestepped the main issue plaguing the Psychological Continuity theory; the fact that it struggles to keep consistency with its plausible Criterion of Identity when also having to maintain transitivity. As Parfit says: "Now that we have seen that identity is not what matters, we need not try to revise or extend our criterion of identity, so that it coincides more often with what matters." (Parfit, 1984, p.243) Relation R does not suffer from this issue, as it is does not make any claim about identity. Instead, it simply describes the way various actual relations come together to constitute that which we actually are. This is a big part of what makes Parfit's theory such an interesting contribution to the discussion; while it does incorporate a theoretical position that establishes a criterion of personal identity, its primary goal is to show that this identity is not what matters.

When we look at the case of transportation, we can see that whether or not "you" wake up on Mars does not carry the importance we intuitively think it should. We feel that this concept of the *self* must be of great importance, but when we examine these cases in depth, we do not find anything that resembles this notion. Instead,

what we find are the various connections and interrelations that constitute us, and therefore can be the only thing that matters.

# Part IV

# The Aftermath

Now that I have explained Parfit's position in detail and provided what I believe to be good enough reasons to accept that this theory is the best theory we have, it seems proper to discuss what the implications of this theory are. Some of these implications might be fairly obvious based on the discussions that were had in the previous part, whereas others might be less intuitive (which, as we have seen, does not mean less plausible).

I will divide this part into three sections. First, I will look at how this worldview, if it is new to us, ought to change how we approach ethics and how we judge our own and others conduct. Secondly, I will discuss whether or not we can *truly* believe that identity does not matter, and what the purely practical effects of an actual *belief* in this worldview will do to most people. By this I mean to reflect on the type of changes that are not mandated by a change in our view of persons and personal identity, but rather the various other beliefs, emotions, and habits that would simply change as a natural reaction to this new world view. Finally, in the third and last section of this part, I will provide an overall summary of the thesis and what conclusions I have found.

# Chapter 4

# Implications

## 4.1 Changes

There are two questions that need to be asked regarding the implications of believing this view. First, if we have been convinced that this view is the truth about our nature, should this change our behavior? And secondly, *does* this change our behavior? The first questions is a matter of ethics and rationality. The second question is one of emotions and instincts.

#### Should the Truth Change How We Act?

I think there are a few important implications of this view regarding how we should act. When dealing with questions of morality and justice, we feel that we have some special reason to value our own futures. That is to say, although we may care about the well being of people in general, we are especially concerned about our own well being. I think much of this belief comes from the idea that we are separate people. This idea stands strongest when we hold a non-reductionist view. If we accept the reductionist account, the distinct borders between us become less sharp.

As a result, I think it is plausible that we should care more about the well-being of others, and less about our own. This is not to say that there is no reason for prioritizing ourselves first. There are still practical reasons for why we should do so, such as the simple facts that in order to provide for others, we must first be in a stable positions. A second such reason is the fact that we *do* have some sort of intimate knowledge about our own lives, and a special ability to affect it, so that in most cases it would have the greatest positive effect if we took care of ourselves. But what changing our view about the nature of ourselves should do, is make the well-being of ourselves as a *goal* more on level with the well-being of anyone else.

On a personal level, I also think accepting this theory provides an important change. In viewing our future selves, not as some "carriers of the Further Fact of our identity", but rather as another person who stands in a very intimate relation to ourselves, we ought to rationally let go of certain fears. Whereas we still have good reason to fear the suffering of this person, as suffering would be bad regardless of what our relation would be, we might not have reason to fear death. At least not to the same degree that most of us do. Death simply means that no future person will be psychologically continuous with me. This might be regarded as worse than the alternative, but I do not think it warrants the level of dread and existential anxiety that our normal conception of death often produces.

**Does the Truth Change How We Act?**

When I read through the arguments provided by Parfit throughout this book, I find myself utterly convinced. It seems to me that there simply is no logically sound alternative to the reductionist position. However, as Parfit points out, no matter how strongly convinced we may be by the rationality of the arguments, there will always be some part of us where the doubt lives on. No matter how sure I could be of the rationale, should I be placed on the transporter pad and asked whether I wanted to hit the button, I might hesitate.

This, I think, would not be due to any rational doubt of the arguments we have been discussing throughout this thesis. Instead, this would be due to this ingrained intuitive part of my mammalian brain; a brain which has evolved through millions of years with the goal of keeping the DNA replication process going. Although developing faculties capable of logic and abstract reasoning have clearly been important in this evolutionary journey, there are limits to our biological thinking machines. Partly, in keeping this process going, there has clearly been ingrained in all of us a very strong urge to *survive*. I think that this instinct may be perhaps most responsible for the development of our concept of our *selves* as a further fact. This would explain why most of us find the non-reductionist account much more intuitively appealing, despite our logic pointing in the other direction.

All this is to say, that even if we can accept this theory intellectually, it may be of limited impact on our actual emotions and actions. This effect will undoubtedly vary from person to person. In my own case, I can report that after realizing the truth of these claims, I have become less worried about my own future. There does not seem to be quite as stark a contrast between my own life and those of others, and I find myself caring more about the overall impact an action has on *all* the people affected, and less on myself specifically.

However, the longer I go without reflecting on these ideas, the further these changes fade, and I gradually transition closer to what my emotional state was before this

realization. I recognize myself in Parfit's description: "Such concern [for our own lives] would remain, as a natural fact, even if we decided that it was not justified. By thinking hard about the arguments, we might be able to briefly stun this natural concern. But it would soon revive." (Parfit, 1984, p.275)

These observations lead me to believe that due to our inherent limitations, we might only be able to realize a tiny portion of the effects that would be realized if we were completely rational beings. Seeing as we are not such beings, I think the best we can do is to return to reflections such as these, so that we might be remined of what really matters.

## 4.2 Conclusions

Throughout this thesis we have explored various concepts and challenges related to the notion of personal identity. If any one thing is certain, it is that this is a vast and complicated subject, with a myriad of different interrelated elements. I do think though, that it is possible to reason our way through this metaphorical jungle, and arrive at rationally justified conclusions.

As I have argued, I think that Parfit provides *by far* the best alternative view on this subject matter. I've argued throughout this thesis that Parfit has shown undeniably that these five claims must be correct:

1. Personal Identity (PI) can be plausibly explained only by Reductionist accounts.
2. The best account of PI is based on psychological Connectedness and Continuity.
3. You can fully describe a Person by using completely impersonal language – i.e., an objective account.
4. The nature of PI cannot be determinate in all cases.
5. PI is not in fact what matters in matters where we are concerned about survival or morality – instead it is Relation R that we should care about.

If we accept this view, not only will the debate within the metaphysical field of Personal Identity make more sense, but it might change the way we view fundamental aspects of our own and others lives. As I have argued, the distinction between "self" and "other" becomes less clear, and there is more cause for empathy, and less for egoism.

We might not be able to fully believe these conclusions, at least not at all times, but in those moments we do; we might end up making the world a little better.

# Bibliography


Blatti, S. (2014). Animalism. URL: https://plato.stanford.edu/entries/animalism [Online; accessed 18. Nov. 2023].

Kind, A. (2015). *Persons and personal identity*. John Wiley & Sons.

Korsgaard, C. M. (1989). Personal identity and the unity of agency: A kantian response to parfit. *Philosophy & Public Affairs*, (pp. 101–132).

Levin, N. (2019). 1.1: Introduction to Philosophy and the Ship of Theseus. *Humanities LibreTexts*, . URL: https://human.libretexts.org/Bookshelves/Philosophy/Ancient_Philosophy_Reader_(Levin)/01%3A_The_Start_of_Western_Philosophy_and_the_Pre-Socratics/1.01%3A_Introduction_to_Philosophy_and_the_Ship_of_Theseus.

Noonan, H., & Curtis, B. L. (2018). The simple and complex views of personal identity distinguished. In V. Buonomo (Ed.), *The Persistence of Persons Studies in the Metaphysics of Personal Identity Over Time* (pp. 21–40). Editiones Scholasticae.

Olson, E. T. (2002). Personal Identity. URL: https://plato.stanford.edu/entries/identity-personal [Online; accessed 20. Nov. 2023].

Olson, E. T., & Witt, K. (2019). Narrative and persistence. *Canadian Journal of Philosophy*, *49*, 419 – 434. URL: https://api.semanticscholar.org/CorpusID:150202927.

Parfit, D. (1971). Personal identity. *The philosophical review*, *80*, 3–27.

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

Piccirillo, R. A. (2010). The lockean memory theory of personal identity: Definition, objection, response. *Inquiries Journal*, *2*. URL: https://api.semanticscholar.org/CorpusID:169430218.

Shoemaker, D. (2005). Personal Identity and Ethics. URL: https://plato.stanford.edu/entries/identity-ethics [Online; accessed 19. Nov. 2023].