Contradictory Allies

A look at Relation R and Embodied/Extended/Embedded Cognition

Student: Jurriën de Vries

Nummer: S4208137

Begeleider: Leon de Bruin

Woordenaantal: 14180

Datum: 29-5-2019

Scriptie ter verkrijging van de graad "Master of Arts" in de filosofie
Radboud Universiteit Nijmegen

Hierbij verklaar en verzeker ik, Jurriën de Vries, dat deze scriptie zelfstanding door mij is opgesteld, dat geen andere bronnen en hulpmiddelen zijn gebruikt dan die door mij zijn vermeld en dat de passages in het werk waarvan de woordelijke inhoud of betekenis uit andere werken - ook elektronische media- is genomen door bronvermelding als ontlening kenbaar gemaakt worden.
Plaats: Nijmegen Datum: 31-05-2019

Abstract

This thesis is an examination of Derek Parfit's 'relation R' theory and the embodied/extended/embedded cognition(EC) theory. It argues that a synthesis between these two theories is desirable, despite the incompatible elements that exist between them. In service to this goal both theories are examined, primarily in the light of the counterarguments their proponents provide against their ideological adversaries. The paper argues that relation R and EC are natural allies opposed to dualism and that a synthesis would create a stronger anti-dualist position. It outlines boundary conditions for such a synthesized position and finally proposes one example of how relation R could be modified into a synthesis.

Introduction

In everyday life, we constantly use the concept of identity, often without examining it very closely. Sometimes different conceptions of identity may conflict like in the famous ship of Theseus example, where a ship is continually repaired with new planks, nails, ropes and sails. Most of us would say that these replacements do not create a different ship. Even when every part is replaced, Theseus' ship remains. But theoretically, one could collect all the old parts and reconstruct the original ship once all parts have been replaced, and doing so would raise a number of questions. Do both ships belong to Theseus? Are they one and the same ship? If so, can one ship be in two places at once? If not, which ship is the one that was present in Theseus' great battle? Did the identity of the ship change when the original parts were put together again?

Of course, most of these questions are easy to brush aside when they are about ships. It may be a pain for Theseus if his ownership of his ship is questioned, but that can be settled in a court. But the same questions can also arise when it comes to persons and that can make them much more important. Take an individual with multiple personalities, for example: should we hold all of the personalities responsible for the crimes of one? If one conjoined twin wants to get married, but the other does not, can they? Is a patient with brain trauma still the same person if they've lost all their memories and their personality has changed? In philosophy, these questions are usually grouped together under the topic of Personal Identity (PI). Philosophers concerned with PI try to provide answers to these questions by giving definitions and requirements that determine whether two individuals are the same person or not.

One such answer has been given by Derek Parfit, in the form of his theory of relation R. At its heart, this theory is a reductionist redefinition of what identity means. It is intended to solve the problems that arise when one tries to fit a dualist notion of identity into a materialist worldview. However, it is my view that Parfit does not entirely succeed in this redefinition, because he remains focused on a purely psychological, brain-centric notion of identity and neglects the role the body and environment play in making people what they are.

I am not the first to make such criticisms of contemporary philosophers concerned with the mind. The embodied/embedded/extended cognition (hereafter referred to as EC) movement accuses much of traditional cognitive science (TCS) and philosophy of an unjustified focus on the brain. These EC theorists reject what they call a computationalist view of cognition, and propose their own view which incorporates the body and environment in its explanations of how the human mind works.

In this paper, I do not intend to prove or disprove either relation R or EC. I am sympathetic towards both, and will provide arguments in their favor when appropriate, but do not intend to give a thorough overview of all the arguments for these positions nor refute all their objections. Instead, I wish to establish that relation R and EC are well-suited to each other, despite the fact that the two theories are incompatible in their base forms.

Parfit's relation R dismisses the importance of the body out of hand, which is a major weakness of the theory. However, this weak point is not a central feature, and could be accounted for without tarnishing the revolutionary new insights the theory can provide. Simply put, it could use a bit of EC to shore up its defenses and doesn't need to compromise its key observations to do so.

On the other hand, EC lacks a coherent theory of identity to explain what makes one embodied person different from the next, which means it currently can't take part in many of the influential debates within philosophy of mind. Because of EC's anti-dualist nature, it requires a reductionist theory of identity to explain what a person fundamentally is and I will argue that Parfit's relation R is well-suited to the task.

The thesis will be structured as follows:
Chapter 1 and 2 will explain the theory of relation R and EC respectively, providing the relevant context for each. In chapter 3, after briefly noting why the two theories are incompatible in their base forms, I will argue why they nevertheless form a complementary pair. In particular, I will outline under which conditions the two theories could be unified without compromising the core value each adds to to philosophy as a whole. And finally, I will propose some preliminary alterations to relation R that can allow the two theories to form an alliance which will strengthen both.

Chapter 1: Parfit's Relation R

When we say that Parfit's theory is one of identity, what is it that we actually mean? In philosophy, a distinction is often made between numerical and qualitative identity.To illustrate the difference, imagine a pair of pencils, of the exact same make and model, from the same box. If neither has been used and if they are very carefully crafted, they will be almost entirely qualitatively identical. That is to say, they will have almost exactly the same qualities. They will have the same colour paint, the same length, the same thickness of pencil lead, and so on. However, they will not be numerically identical. One can be used, broken, or sharpened, without affecting the other. They can be in different places at the same time. They are, in short, not the same object. The question of personal identity concerns this last type of identity, numerical identity. Persons cannot be in two places at the same time, and none of their qualities may be contradictory, which means that they cannot be dead and not dead or present and not present at the same time.

Philosophical theories attempting to explain our identity can be very broadly categorized as reductionist or non-reductionist. The same distinction can also be made with the terms dualist or non-dualist. Dualist or non-reductionist theories of identity assume that there is some nonmaterial thing or property, like the soul, that is responsible for our identity. Reductionist or non-dualist theories assume that personal identity consist merely in some particular set of facts about material reality, like a particular configuration of the brain and body. In the following part, we will examine philosophical theories in both categories, in order to illustrate why Parfit rejects non-reductionism and which problems with reductionism he attempts to solve.

1.1: Why other theories of identity don't work

Let us begin with Parfit's objections to dualist theories. Dualism, exemplified most famously by Descartes, is the position that humans have both a material body and an immaterial soul. According to Descartes and most other dualists, the soul is the essence of a person, leading to the logical conclusion that PI is also tied to the soul. A dualist definition of PI would be something like the following:

Two individuals are the same person iff they have the same soul.

Further, dualists often define the soul as an unsplittable, singular entity with no separate parts. In this way, if one could observe souls, it would always be obvious which soul is which and therefore which person is which. However, souls are also postulated to be immaterial and imperceptible through material means. No device or organ can detect the presence of a soul, let alone determine which soul an individual houses, nor do we have any idea what exactly is responsible for the connection of a particular soul to a particular body. So, the dualist position is posed with a dilemma: either the irreducible fact of our identity corresponds to some of the elements that we consider typical for individual people (like their personality, memories, physical appearance etc.) or it does not.

Parfit argues that, if dualist identity were to correlate with some or all of these traits, we would see very different results than we truly do, when it comes to, for example, patients with brain damage (Parfit 1984, 227-228). Suppose that the soul is the carrier of our memories, and that it is attached to our bodies through a particular neuron in the brain. What we would see in experiments is then that we can destroy large parts of the brain without having any impact on memory, but if we carefully remove or alter that neuron, the memories of the person would completely change. Their brain would now be connected to a new soul, or to no soul at all. While this is perfectly possible, it is not actually what we see when people's brains are damaged or altered through surgery. So, according to this argument, this version of dualism can be empirically tested and such tests have already disproved the hypothesis. This argument can be reformulated with minor changes to work against any dualist view that ties the soul to any observable quality we typically use to identify persons.

So if there is a soul, it cannot be responsible for any of those things that we typically use to identify our friends, family or even ourselves. But if this is the case, then we have no way of knowing that we are associated with one particular soul or one particular point of view. If all or our memories are simply stored in the brain, then our soul could be replaced every second of our lives, each receiving the memories of all the previous souls without realizing that there has ever been a break. In other words: a soul that does not correlate with any of the observable traits we think of as typical for a person is an unmeasurable and pointless hypothesis that adds no explanatory power.

Another view which Parfit also sees as non-reductionist though it does not propose a Cartesian soul is the ‘point of view’ position. According to this idea, all people have their experience from a certain perspective, namely their first-person point of view (Parfit 1984, 210). This point of view cannot be reduced to impersonal facts about material reality, because a point of view is inherently something that belongs to a person. A point of view definition of identity would be something like:

Two individuals are the same person iff their experiences belong to the same point of view, which persists over time.

This argument is subject to the same dilemma that faces the dualist position: either consistency of point of view is responsible for consistency in some other (empirically verifiable) quality, or it is not. If it is, then we should find some sharp division in this correlated property before and after we sever the connection to this point of view. If it is not related, the point of view is a postulated entity which provides no predictive or explanatory power and should therefore be dismissed with Occam’s razor.

Parfit classifies the two theories above as non-reductionist, and concerns himself mostly with reductionist theories. The key feature of reductionist PI theories is that they believe identity consists just in the holding of certain more particular facts. Formulated broadly to encompass all the different reductionist theories, reductionists believe:

“A person’s existence just consists in the existence of a brain and body, and the occurrence of a certain series of interrelated physical and mental events.” (Parfit 1984, 211)

The ‘brain and body’ of that definition is important, as it allows us to make a rough categorization of reductionist accounts of PI, based on what they believe is key to identity. There are physical theories which tie identity to consistency of the body, psychological theories which tie identity to consistency of the mind, and combined theories which consider both important. A physical definition of identity would be something like:

Two individuals are the same person iff they have the same (crucial parts of the) body, which has persisted over time.

A psychological definition would be very similar, but swap out the body for the mind, and a combined definition would incorporate both. There are a great many different variations on these nonreductionist theories, but most can be incorporated into this schema by more precisely specifying what must be shared between the individuals, and in which way that shared property can persist over time.

I have collected these reductionist theories together, because just like non-reductionist theories, they mostly face the same objections with minor variations. The primary objection is the ‘Ship of Theseus’ problem, or the problem of a spectrum of continuity. Let’s take the physical view as an example, which holds that personal identity over time is explained by the persistence of the body. Now, the body does not actually persist during a whole life: every molecule in a human body is replaced several times over the course of an average lifespan. In addition to this, surgeons already regularly replace body parts with metal or plastic prosthetic. In the future we will likely be able to replace more and more organs and body parts with equally functional prosthetics. So if PI depends on the continued existence of the body, how much of the body can we replace before the individual coming out of the operating room is not the same person as the individual who went in?

One possible solution which Parfit discusses is to draw a line, for example by specifying that at least 50% of the body must remain over a period of a year in order to maintain personal identity (Parfit 1984, 234-236). Another option is to assign PI to a certain critical cell or molecule, such that if this single part is replaced, a new person is created. Neither of these options are very convincing, however. Both require us to believe that we can replace huge portions of grandma’s body with artificial hearts, bones and neurons without destroying her identity. But then if we change one particular cell or cross the 50% threshold by one cell, grandma is suddenly gone and replaced by a new individual who is almost perfectly identical to grandma as she was before that single cell was replaced.

Though replacing mental properties through surgery is more difficult than replacing organs, psychological theories of identity suffer from the same conceptual problem. To see this, we need only imagine a hyper advanced kind of neurosurgery which is capable of changing a patient's personality, memories, etc to those of another person, like Napoleon (Parfit 1984, 231-233). We will once again have to draw an arbitrary line somewhere or designate a specific mental quality as essential to a person's identity. Even a combination of psychological and physical qualities cannot avoid this problem of the spectrum. It seems perfectly reasonable to say that a new person is created if an individual undergoes a procedure that alters or replaces them by 100%, and that this is not true for a minor procedure that only changes 1%. Yet, if we draw any line between these extremes we are led to the absurd conclusion that, for example, the 49% individual is identical to the 1%er, but not to the 51% individual, despite that fact that he is much more like the latter than the former. If we do not draw the line, the only other option seems to be to designate a single undividable particle or quality as the sole carrier of our identity. Beyond the lack of evidence that such singular particles or qualities even exist, this too leads to the absurd conclusion that vast swathes of our physical and/or mental being can change without destroying our identity, but one tiny change can end our existence.

1.2 The problem of duplication and relation R

Besides the problem of drawing lines in spectra, reductionist theories of identity also face the problem of duplication. The duplication problem can most easily be explained through a short thought experiment. Conveniently for our purposes, the following thought experiment around duplication also provides a good starting point to explain what relation R entails.

For simplicity's sake, the thought experiment only directly addresses physicalist criteria for identity, but it can be adjusted to cover any of the standard reductionist positions discussed in chapter 1.1.

> *Suppose that you have a deadly disease. It affects exactly half the cells in your body, in a randomly distributed way, such that all of your organs and body parts are made up of equal parts healthy and diseased cells. A scientist approaches you with a device that he thinks will cure you. The device will scan your entire body, identify all of the diseased cells, and destroy them. Then it will extrapolate from the 50 percent of the body that is left to perfectly determine which healthy cells belong where. From tanks of raw biological sludge, it constructs all the healthy cells needed and places them in the correct position. All this will happen within a fraction of a second, so quickly that removing half your biomass has no adverse effects. After this procedure your body will consist of 50 percent new material, including your brain, but all your mental qualities will remain completely the same. After explaining the procedure, the scientist asks you if you would like to step into his device. Do you think you should, because it would cure you, or would you refuse because all the machine does is kill you and create a new healthy person?*

For simplicity's sake, the thought experiment assumes that survival is your only consideration. Most people, I think, would take this deal and step into the device. After all, the alternative is to die, and once the procedure is over, the resulting individual would continue your life content in the knowledge that they were cured. For those who would not take this deal, believing that such a procedure would result in immediate death and replacement, there are two possibilities. Either the thought experiment can simply be adjusted slightly to accommodate their view, in which case they must face the duplication problem. If it cannot, the thought experiment is brought back to the spectrum problem discussed in chapter 1, with the question being where we draw the line between conventional surgery like replacing a heart valve and this hypothetical procedure. Assuming that we accept the deal and agree that the devices cures us, the thought experiment continues as follows:

> *Now imagine that your diagnosis was a lie. You were never going to die. But you still stepped into the device because you have been tricked by the scientist. The device still scans your body, but instead of destroying half of your cells it removes them from your body. In fact it moves both halves of your body. To illustrate, before the procedure your body is made of two equally big sets of cells: a set of cells called A and a set of cells called B. Both cells from A and from B are present in every part and organ of your body making up roughly half of everything. Nothing in the cells actually distinguishes these two sets from one another, cells are assigned randomly to each set by the machine during the scanning process. After scanning, the device takes all of the A cells and moves them a meter to the left. At the same time, it also moves all of the B cells a meter to the right. Then, before either resulting body has a chance to suffer from the billions of tiny wounds, it fills in both the A body and the B body with the necessary cells to be complete again. As with the previous example, your mental qualities remain the same except that they are now present in two bodies, A and B. Both individuals, A and B, exit the machine in such a way that they cannot see each other, and through careful planning and manipulation, the scientist makes sure that neither knows of the existence of the other. A and B live out the rest of their lives, content that they were cured of their disease, none the wiser about the deception.*

This is where traditional conceptions of PI run into trouble. Because such theories, like all of the ones discussed in chapter 1, consider identity to always be a definite matter, they must choose from three equally unsatisfying answers:

1) Neither A nor B is you. The second part of the thought experiment destroys personal identity, even though the first part did not.
2) Only one of the pair is you, though both are exactly equally like you and they both think they are you.
3) Both A and B are you, though they go on to lead separate lives.

The first option is unsatisfying, and contradicts a key feature of the concept of personal identity: that it is a quality possessed by a person or a connection between two individuals.

As far as A or B are concerned, the first part of the thought experiment is what happened to them. We would have to accept that an individual's personal identity can be destroyed or altered due to something that happens elsewhere in the world and doesn't affect them at all.

The second answer is also unsatisfying, because it again requires us to assume that identity is some sort of unmeasurable entity that we are unaware of, falling into the same trap that Cartesian dualism did in chapter 1. In this scenario, no outside observer nor A or B themselves can tell whether they are the same person as you. To insist that PI exists in one of the pair but not the other, despite there being no differences between them, is to divorce the concept from all empirical evidence and even internal experience.

The final option seems the most satisfying, and comes closest to Parfit's theory, but still falls short. PI simply does not allow for two individuals to be the same person at the same time. That would allow for the possibility of contradictory propositions to both be true: if you are both A and B, you could be pregnant and not pregnant at the same time. You could be in pain and not in pain, dead and not dead, etc.

This duplication issue, as well as the spectrum problem, seem to point toward a fundamental flaw in reductionist theories of PI that wish to provide definitive answers for all cases. Whatever criterion for identity is suggested, edge cases can be imagined where the answer is unclear or the criterion lead to absurd conclusions. Philosophers keep trying to find new criteria or refine old ones to get around this, but new problem cases are always thought up shortly afterwards.

Rather than trying to propose yet another set of criteria that has no edge cases, Parfit starts by simply accepting the observation that the question of identity can sometimes be 'empty'. That is to say, there may simply not be an answer. He compares personal identity to the identity of a club (Parfit 1984, 260). If a group of people come together once a week to play a game, they may choose to call that a club. Now, if these people don't meet for several years, and then decide to play together again, is that the same club, or a new club with the same members? What about if some of the original people left and new ones joined? These questions don't seem to have a clear answer, because they are empty questions. It's not that we don't have enough information about the group to know whether it's the same club, it's that we simply haven't decided. If the original group calls themselves "Aeneas", and picks up that same name when they come back together years later, we may say it's the same club. If they choose a different name, we may say it's a new club. People may even disagree on whether or not it's the same club, even with all the same information. If they do, that's not because they disagree on what is the case, but on what should be the case. "Is this new group of people Aeneas?" is an empty question, the proper question would be "Should this new group of people be considered Aeneas?".

For clubs, this can be easy to accept, but when it comes to the identity of people, it's a bit harder. We place tremendous importance and value on our identities, and to say that they are essentially arbitrary and can be empty seems to rob them of that importance. In order to illustrate why this is not the case for Parfit's theory, we must return to the idea of the spectrum. Instead of starting with the spectrum of identity, however, we will first look at an analogy involving the spectrum of age.

In most countries, there is a legally and culturally accepted threshold of adulthood called the age of majority. Generally, this threshold is set at 18 years of age, though it can be a bit higher or lower in some countries. At the age of majority, a person is considered an adult and gains the legal rights and responsibilities associated with control over one's own body, property, actions and decisions. We consider the age of majority to be an important concept, because adults are wise and responsible enough to make their own decisions and be held accountable for those decisions, while children are not. But at the same time, we also know that the night of one's 18th birthday is not really a special magical moment that grants that wisdom and responsibility. There is a much bigger difference between the wisdom of a 12 year old and that of a 15 year old, compared to a 17 year old one day before and one day after his/her birthday. If we were to view the age of majority like we often do with PI, we would be led to the absurd conclusion that one day of age difference can grant the experience and knowledge necessary to buy a house. Generally, we don't do this, instead accepting that the age of majority is essentially arbitrary and that, outside the context of a legal system with defined boundaries, it can be an empty question.

Yet somehow, admitting that the age of majority is arbitrary does not really impact our view that it is important. Even if we disagree on whether the threshold should be 17, 18 or 19, most people believe that an 8 year old should be protected from their own irresponsibility, while a 34 year old should be held accountable for their actions. This is because we recognize that the boundary of adulthood may be arbitrary, but the spectrum that boundary is drawn on is not. That spectrum, describing the growth of our wisdom and responsibility with age, stretches from a child who would run into traffic to grab a ball, to the judge who must carefully weigh evidence in a case, is very real. Most of us progress along it over time, though not always at the same rate and not always equally far. The legal system must draw an arbitrary boundary somewhere in this spectrum for the sake of convenience, and that boundary informs how we think of age and responsibility. But it does not encompass or determine all of our thoughts and decisions. We treat a 6 year old differently than an 11 year old, even though they are both minors. The same goes for a 20 year old and a 60 year old, and the reverse can apply to a teen a day before and after his/her birthday. In summary, age-related wisdom is a non-arbitrary spectrum, on which we may draw arbitrary boundaries. These boundaries can be important and useful, but they derive their worth from the importance of the underlying spectrum, which is what truly matters. Parfit views personal identity much in the same way.

The underlying spectrum is what he calls relation R. It describes the amount of psychological connectedness and continuity between two individuals. On one end of this spectrum is the relationship between the individual who writes this sentence and the one who wrote the previous sentence. These two individuals are virtually identical and therefore very strongly

R-related. Towards the other end is the relationship between the individual called Jurriën at 24 years of age, and the individual called Jurriën at 8 years old. They are not very strongly connected, as they do not share many thoughts, intentions, beliefs, tastes, etc. However, they do have psychological continuity, because there exist a chain of overlapping relations of strong connectedness between them. At the far end of the spectrum is the relationship between the individual called Jurriën writing this sentence and any individual living in Indonesia in 4000 BCE. These two are about as weakly R-related as two human individuals can be, with only the most basic psychological connectedness and no psychological continuity.

In Parfit's view, this spectrum of R-relatedness is what we truly care about when we talk about personal identity, just as the spectrum of age-related wisdom and responsibility is what we truly care about when we talk about the age of majority. This can be difficult to see, because our first instinct when faced with questions or thought experiments like the one above is to ask: "Are these individuals the same person?". But, so Parfit argues, this question cannot give us much insight into what truly matters about identity, and certainly cannot settle what the criterion of identity should be. Just like the question: "Is that individual a minor?" cannot settle what the age of majority should be or provide much insight into age, wisdom or responsibility. This does not mean that we should not have a criterion for identity, Parfit admits that having a clear definition of identity (or at least one that covers all actual cases), even if it's arbitrary, is important for a functioning legal system. But just as we make distinctions between minors of different ages or adults of different ages, Parfit argues we should make distinctions between different levels of R-relatedness, even if they are on the same side of the PI threshold. In fact, he claims that outside of a legal context, the absolute position on the R-relatedness spectrum should guide our views more than where the arbitrary threshold is placed or what the arbitrary definition says.

So when we return to the problem of duplication, relation R provides a simple set of answers. A and B are equally R-related to the individual who stepped into the device, and at the outset are also strongly R-related to each other. As their lives go on and they have different experiences, their R-relatedness will decrease. After a certain point, they may only be slightly more R-related than a set of identical twins might be, especially as their memories and other mental qualities from before the procedure begin to fade. As for their identity, Parfit leaves that question up to the courts or other authorities (Parfit 1984, 325-329). Such an organisation may decide not to allow branching paths of identity (situations where one person splits into two individuals who are still the same person), which would mean that A, B, and the original individual are all different persons. This does not pose a problem for the theory of relation R like it does for traditional PI theories, because identity is allowed to be arbitrary when relation R takes on the role of fundamental connection between individuals.

1.3 Relation R and the body

By shifting our focus from discrete notions of identity to the non-discrete notion of relation R, Parfit can maintain many important features of identity, without giving up very much at all. Relation R is truly a property of the two individuals involved, independent of the features of other individuals. It also does not depend on trivial externalities, like whether a machine only creates one copy or two copies of the individual in question. Both these features are important parts of the concept of personal identity, and Parfit works hard to preserve them.

But strangely, he almost completely ignores the importance and value of the body, though this too is an important part of most people's conception of their identity. The only strong philosophical arguments presented against physical theories of identity (the spectrum and duplication problems) apply equally to psychological ones. So why then does Parfit explicitly deny the relevance bodily continuity could have to identity, as illustrated by this quote:

"What we value, in ourselves and others, is not the continued existence of the same particular brains and bodies." (Parfit 1984, 284)

Though he does admit that physical similarity could be important to allow for psychological connectedness, the conception of this importance is very limited. In Parfit's view, only very large differences like a body of a different sex could be a limiting factor on psychological connectedness. Only a select set of people, like those who are very beautiful, might require more precise physical similarity in order to maintain their relation R. No true arguments are provided in favor of this view, only the insistence that all we truly care about is psychological connectedness and therefore that the only body part we care about is the brain. In making this argument, Parfit makes an illuminating comparison, saying that we should care about our own physical bodies like we do about wedding rings:

"This could be like one's wish to keep the same wedding ring, rather than a new ring that is exactly similar. We understand the sentimental wish to keep the very ring that was involved in the wedding ceremony. In the same way, it may not be irrational to have a mild preference that the person on Mars have my present brain and body." (Parfit 1984, 286)

In this example of the wedding ring, as when he speaks about how all we value is psychological connectedness, Parfit seems to overgeneralize his own preferences. While to some, the continuity of a wedding ring may only be a mild sentimental preference, others place far greater value on it. Some people, upon the loss of a wedding ring, spend months searching for it, hire divers with metal detectors if it fell in the water, and generally spend large amounts of time and money worrying about this small piece of metal. Of course, most of us would be less distraught at the loss of a wedding ring than the complete loss of our identity, but not necessarily the loss of part of our identity. One could very reasonably prefer to lose all of one's preferences in food, movies and music over losing a wedding ring. The same thing might well apply to the body, if it ever became possible to 'lose' one's body.

In the following part, when we discuss EC, the importance of the body will become clearer still, especially with regards to the possibility of psychological connectedness and its relation to physical similarity. The main project of EC could be reasonably described as a critique of brain-centrism, which is a charge that may be fairly applied to Parfit's relation R, after all[1]. For now, we may simply (with equal evidence as provided for Parfit's dismissal of the importance of bodily similarity) claim that bodily similarity between two individuals is required for the existence of personal identity, separately from relation R, and that we hold the two relations to be of similar value.

[1] While Parfit's notion of relation R is theoretically not restricted to human brains (and therefore not technically brain-centric), he does restrict his view of cognition to just those processes that are commonly thought to take place entirely in the brain in real human beings. I will continue to use the term brain-centrism in reference to Parfit's theory as well as traditional cognitive science, but note that the charge applies to Parfit only in a more abstract sense, not in the direct sense that applies to TCS.

Chapter 2: Embodied/embedded/extended cognition

EC, unlike relation R, is not a theory from a single author. Instead, it is a philosophical and scientific movement that recently rose to prominence in many different fields, all concerned with how the human mind works. In philosophy, the main question that EC provides an answer to can be formulated as: "How should we conceptualize cognition?", where cognition is the term for basically everything a human mind is capable of doing, from complex mathematics to perception of the world around us to walking.

Before exploring EC itself, I would like to give a quick overview of the history of the cognition debate, viewing it through the lens of what I will call the historical movement away from dualism, or the anti-dualist steps of history. Partly, this is because EC is best understood in terms of the previous movement it rejects (namely computationalism or TCS), like so many philosophical movements. This rejection centers mostly on the remaining dualist aspects within computationalism, and the historical anti-dualism lens allows us to more clearly identify these aspects.

The other reason this perspective is valuable for our purposes is that it explains why EC positions itself as the natural successor to computationalism. Since Parfit's relation R assumes a computationalist view of cognition, EC's promise to replace TCS will be very relevant in chapter 3, when we discuss why EC and relation R would both benefit from being made compatible.

2.1: The historical movement away from dualism

Of course, a description of an anti-dualist process will have to start by describing dualism. The quintessential example, closely linked to modern Christian beliefs, is Cartesian dualism. Descartes holds that persons are an immaterial, abstract and unobservable soul, which constitutes the mind. All cognition occurs in the soul, and all mental qualities of a person are stored there. The soul is connected to the body through the brain, and controls our body from there (Ravenscroft 2005, 9-24).

From the Cartesian example, we can extract a few dualist properties or tendencies that come up in non-dualist theories of cognition. The most obvious is a disconnection from the material world: Descartes literally equates "thinking substance" with "non-material substance". This disconnect also portrays abstract cognition as typical cognition. Descartes' view that activities like math and reasoning are cognitive while soccer and painting aren't is still a very common one, even among non-dualists. Cognition is thought of as something that happens 'inside' and is empirically unobservable, while events that we can see happening can't be part of cognition. A final interesting quirk is that dualism is forced to designate some matter as special, to explain the connection between immaterial mental events and their material consequences/causes.

With these things in mind, let's quickly go through the history of the cognition debate to see where these dualist properties pop up and how they're eliminated.

Behaviourism is the first major anti-Cartesian movement, centering around the problem of empirical observation. According to behaviourists, we should not try to establish the inner workings of the mind, since it's a 'black box' that we cannot observe. Instead, we should restrict ourselves to what we can see, describing the mind only in terms of the inputs it receives and the outputs that result from those inputs. In this behaviourist model, all mental events and qualities like thoughts are simply tendencies toward a particular mode of behaviour. So, something like anger should not be viewed as an internal emotional experience, but merely as an inclination towards violent actions, generally caused by stressful types of input (Ravenscroft 2005, 25-38).

While behaviourism provides an interesting method to get around the presupposed opaqueness of cognition, the 'black box' itself is even more closed to observation than the soul, as even introspection is written off as unreliable. Cognition is a lot less abstract than in dualism, as an inclination towards math is treated the same as one towards running. Behaviourism loses many of the hallmarks of dualism by refusing to engage with the 'black box', focusing only on the material realities of input and output. But in doing so, it copies dualism's biggest problem: the inability to explain how cognition really works with empirical evidence.

Directly opposing behaviourism was identity theory, which claimed that not only was the mind not a 'black box', it was directly observable inside our heads. New technological advancements (mainly cerebral angiography) allowed brain activity to be measured, which identity theorists believed gave us direct access to the workings of the mind. We could induce anger in a subject, scan their brain and know exactly which brain state is identical to the internal experience of anger. With large numbers of scans and more advanced technology, identity theorists were confident that we could eventually fully explain every mental process, state and property in terms of brain states (Ravenscroft 2005, 39-49).

Identity theory may be the most materialist theory of cognition discussed in this entire paper. It directly equates cognitive states with material states and makes no distinction based on the abstraction level of different cognitive tasks. However, it does still designate 'special' matter by equating mental states solely with brain states and ignoring all other objects and organs. In doing so, identity theory still creates a duality between thinking and non-thinking things, only making the change that both are material objects. The direct link to the material also presents another issue, since it turns out that brains differ greatly between individuals and species. Identity theory is forced to either make separate mental concepts for each species or individual, or accept that the general mental states like anger still can't be observed empirically as they differ in each of us.

On the basis of this last objection, functionalism was born. Though functionalism does draw the same connection between brain states and mental states, it does not define them as the same thing. Instead, mental states can be realized in multiple ways and are defined by their

function. Fear, for example, is something that occurs in response to danger, and drives the individual experiencing it to distance themselves from the danger or eliminate it. In humans, this mental state is connected to a particular brain state, but other animals may instantiate fear differently. In this way, functionalism is very flexible. It can even allow for mental states to be instantiated differently between humans, or for them to occur in non-biological entities like computers (Ravenscroft 2005, 50-63).
Functionalism sacrifices some of the direct connection to the material that identity theory has, in exchange for making common-sense notions like 'fear' applicable to all thinking creatures. As a consequence, all cognition is folded into the abstract notion of functions, meaning that clearly goal-oriented cognition is seen as the prime example while more passive mental properties like preferences are ignored. Functions themselves are also non-observable, although most functionalists accept that they are arbitrarily defined and don't really exist. Their causes and effects, which are far more important to the theory, are empirically accessible and real.

Finally, there is the theory against which EC positions itself: computationalism, also called TCS. The key to computationalism is the analogy to a computer and the emphasis on symbolic representation. In this view, all cognition consists of the manipulation of symbols according to rules. These symbols represent things in the outside world, but also abstract concepts and internal states like hunger. The rules can be acquired over time, but some have to be present at birth, to allow new symbolic information to be transformed into rules. As with computers, cognition is a kind of software that can be run on multiple different kinds of hardware (though as with earlier theories, computationalism does locate cognition entirely within the brain in real humans) (Ravenscroft 2005, 81-96). In this sense, computationalism is similar to functionalism, allowing for the possibility of non-neuronal cognition. However, computationalism is more strongly committed to the idea that its abstract entities (the internal symbols and rules) are not arbitrary and really exist, which entails a dualistic separation of the world into physical objects and non-physical symbols and rules. This 'hardware-software' distinction again removes cognition from the material world into an empirically inaccessible realm, even if that realm is more strongly connected to the material world than Descartes' immaterial soul.

Through the perspective of this historical anti-dualist movement, the purpose of the EC project becomes clearer: to take the next step away from dualism, hopefully without falling into the pitfalls mentioned above. In the next chapter, we will be exploring EC in more detail, but for now I wish to note the dualist tendencies it tries to avoid. Firstly, the exclusive focus on the brain common in most of the above theories (including relation R) is obviously absent from all forms of EC (embodied, embedded and extended). All forms regard cognitive processes as at least potentially involving the whole body, with no particular preference or bias towards the brain as the organ responsible for any particular act of cognition. Embedded and extended cognition push this boundary even further, regarding all material objects as potentially part of cognitive processes. EC in general can describe cognitive processes in terms of actions and reactions in the material world, without turning actual cognition into a 'black box'. Nor does it require the 'new dualism' of a hardware-software distinction. Whether EC achieves these lofty goals is left up to the reader to judge, but this elucidation of EC's aims will be helpful in the following part, when we discuss what EC theory actually consists of. As mentioned at the start of this chapter, the historical lens we just explored also provides one of the main reasons for making EC and relation R compatible if we can.

2.2: E. Cognition

As mentioned above, embodied cognition is more of a research programme than a single well-defined theory. Its proponents sometimes support different interpretations of what embodiment means, propose contradictory theories or explanations, and even use different terms like extended cognition or embedded cognition. I will use the convenient abbreviation EC for all these terms, and stick to the most common interpretations of what EC is. Mainly, I will make a distinction between the expanding and replacing variants of EC, but before that distinction can be understood, let us first discuss what EC is in general.

Embodied-, embedded- and extended cognition are all ways of describing how the human mind works. Proponents of the EC programme suggest that it represents the next step in the debate around the mind-body problem (Shapiro 2007), a step I would characterize as another move away from dualism. EC theorists stand in opposition to what they call traditional cognitive science: the view that cognition takes place entirely within the brain, and consist in manipulating symbolic representations according to algorithms. Many of the theories mentioned above (all those that came after behaviourism) are considered to be part of TCS. Despite their differences, they all share a core picture of cognition that EC wishes to abandon.

In a simplified version of this picture, cognition starts when the brain receives input from the senses. This input is considered to carry insufficient information about the environment. For example, the image a room projects on the retina underdetermines what objects are present in the room and in which configuration. Optical illusions show this very clearly, take for example a famous scene from the Lord of the Rings movies. In this scene, a human-sized character sits at a table with a Hobbit, which are creatures roughly the size of a child. In reality, both actors playing these characters are of regular human size. Yet, though carefully managed perspective and props of different sizes, our eyes can be tricked into seeing a man and a Hobbit sitting next to each other, rather than two normal actors sitting several meters apart. In this example, we are being deliberately deceived, but the same principle applies in normal situations too. In order to make sense of the inputs of our senses, the brain must make several assumptions, like that a table is a single continuous object, not two differently sized objects at different distances whose edges overlap perfectly.

The traditional cognitive science picture combines these sensory inputs and the assumptions about them into representations: mental pictures or descriptions of the world around us. These representations are then transformed through algorithmic processes. This does not mean they are necessarily mathematical in nature, but that the representations are used as symbols and manipulated according to rules. All cognitive activity is seen as this kind of abstract symbol manipulation, eventually resulting in action when the proper kind of symbol is produced (Shapiro 2011, 7-27). For a practical example, let's look at waiting for a traffic light. First, the image of a traffic light turning green is projected onto the retina. The brain has to assume that the light is indeed coming from the traffic indicator, and not some other source of light being reflected strangely or a much closer light that perfectly obscures the real traffic light. Then, this symbolic representation is combined with the algorithms that describe the laws of traffic and the operation of a car. If the drivers' algorithms and symbols regarding both are accurate, he will know that a green light means go, and that you press the gas pedal to go. This knowledge is then combined with another algorithm and the symbol representing the driver's destination, and the driver presses the gas and turns in the correct direction.

Key to this traditional picture is the symbolic and abstract nature of cognition, as well as the exclusive focus on the brain. EC rejects both these assumptions, assuming that cognition is neither the exclusive domain of the brain, nor purely symbolic (Shapiro 2007). Much of what the traditional cognitivists would attribute to abstract representations and symbol manipulation is instead explained through simpler interactive feedback systems between the brain, body and the environment. Take for example the earlier point about visual input containing insufficient information. The cognitivist solution is to say that the brain makes abstract assumptions based on rules to complete the input and form a representation. EC theorists point out that visual input is not normally stationary and 2-dimensional, which is what causes most of the confusion. If one were present on the set of the Lord of the Rings movies and able to walk around, it would be very obvious that the two actors were both of normal adult size, and sitting far away from each other.

The difference between the two approaches is perhaps best illustrated by the example of a slope-descending device. Imagine that a scientist is given the task of building a device that can walk down a large variety of sloped surfaces, one based on EC principles and one on TCS. The TCS device would end up looking a lot like the popular conception of a robot: it would be shaped roughly like a human or an animal, equipped with cameras and sensors to accurately measure things like the angle of the slope and irregularities. It would have a CPU to process this information, and then use servos to move its limbs carefully to the correct spots, based on an internal representation of the slope. An EC theorist would build something a lot more like a slinky, already a virtually perfect descending device. Of course, just copying a toy would be cheating, but the same principles that a slinky uses to 'walk' down stairs could be used to drive a human-shaped device that completes the same task. No internal representation or accurate measurements are needed, only an initial push to get down the first step and feedback loops that keep the balance of the device correct. In fact, when these kinds of robots have been built, they display a more human-like gait than the computational robots, suggesting that they perhaps resemble humans more closely in their approach (Collins et al. 2005).

There are several central differences between these two approaches. First and most obviously, the EC approach does not depend on representation and symbol manipulation. Of course, some cognition clearly does depend on symbol manipulation, like language, but an EC theorist is not obliged to conceptualize all cognition in this way. Secondly, computationalism often assumes cognition happens on a static picture of the world, while EC sees cognition as extended in time. In TCS, information comes in, is rendered into a representation, and then cognition is a matter of manipulating that representation. Because of the emphasis on feedback loops, EC has to consider cognition as part of a changing world. With the descending robots, the computationalist bot takes in all the information, computes a path and then executes that path. It may use sensors to check if everything is going as planned along the way, but the plan is devised up front (Chestnutt et al. 2005). A mechanical walker, on the other hand, is driven entirely by ongoing interaction with the environment, and any explanation of how it works has to work on the span of time it takes the device to walk down.

So what is the picture of the mind that EC is trying to portray? Why do I believe that (if this picture is correct) EC forms the next step in the anti-dualist journey? Let’s look at a few notable examples where EC steps outside the bounds set by previous theories of cognition. In recent years, there has been a lot of research into the influence of microorganisms on human psychology. Different profiles of bacteria in the gut show significant correlations with stress response, sociability, diet preference and mental disorders like depression (Clap et al. 2017). EC can seamlessly incorporate these findings, because it can freely attribute mental properties to non-neuronal objects, including non-human bacterial cells. Something like diet preference is conceptualized in EC as a system that can encompass the brain, gut and external objects, keeping itself in homeostasis through feedback loops. A person who eats a lot of fast food does not just have a preference for these foods in their brain, their gut microbiome is also better adapted to processing these foods than other foods. Their web browser might show results for fast food restaurants above healthier options, and their phone might have local fast food delivery places as contacts. In the EC picture, all of this is considered to be part of the mental apparatus responsible for this person’s preference for fast food. Of course, some parts may be more crucial to the preference than others, but it’s certainly not down to the brain alone.

The above mention of a phone and web browser might sound extreme, or more out of place in a picture of cognition than gut bacteria. While some EC theorists do restrict themselves to the body and don’t consider external objects part of cognition, I will go with the more radical notion of EC that does allow for ‘notepad’ cognition.

This example, originating with Eric Bredo, is often used to illustrate how EC incorporates external objects into the schema of cognition. It takes drawing as the quintessential example of how cognition can incorporate external objects. When drawing a picture, one does not plan out every line in advance, and then execute on those planned lines. Instead, the first lines are drawn based on a rough mental picture, and the following lines are made in response to those first lines, and so on. When the picture deviates too much from the original intention, an eraser may be employed, but small mistakes are just as often incorporated into the final work. In this sense, the notepad, pencil and drawing itself are just as much part of the creation process as the hand, arm and brain of the individual doing the drawing (Bredo 1994, 29). There is also another example of notepad cognition, which perhaps may strike you as more explicitly cognitive. In this example, the task being performed on the notepad is mathematics, rather than drawing. The individual in question is trying to solve some equation. She writes down the equation at the top, and wracks her brain trying to think of the correct way to solve it. On the notepad, she jots down attempts at the solution, crossing out what she knows doesn't work. She reaches the answer only by virtue of being able to see the incorrect answers and where they went wrong. If she had attempted the same exercise in her head, she would have been unable to do it. In this scenario, the notepad is very explicitly a part of the individual's cognition. It allows her to keep the wrong answers clearly defined, rather than relying on her own fickle memory, and prevents her from thinking in circles.

This same process of feedback loops constructing our cognition can also be applied to the social context in which we are embedded. For example, the way in which a notepad can make certain types of cognition possible can also apply to social groups like sports teams. It is virtually impossible to imagine the emotional response people feel to their favorite sports team losing outside of the social context of sports. Most people normally find it difficult to deeply care about the death and suffering of complete strangers, yet many feel very intense grief when another group of complete strangers suffers a loss in a much more trivial competition. According to some EC theorists (Huebner 2013), this discrepancy is best explained through the same kinds of mechanisms used above. A fan cheers on a team, buys their merchandise, and the team, along with all the other fans, encourages and strengthens the fan's emotional investment. In the notepad example, memory and calculation is shared between the human and the piece of paper. In this example, emotion and the sense of community is shared between all the fans and the team.

All of the above examples must be 'explained away' by computationalism. Because computationalism is committed to the brain-centric symbol-manipulation picture of cognition, it cannot allow for external objects like notepads or bacteria to be part of the cognitive apparatus. They can provide inputs for the brain to work on, but cannot themselves constitute any mental process like calculation or any mental property like diet preference. This is also why I believe, on the assumption that EC is correct, EC forms the next step away from dualism. Unlike computationalism, it allows any material object to be endowed with mental properties or to be part of a cognitive system. Under EC, cognition can be a far simpler matter than ever before, as we saw with the descending machines, allowing us to incorporate basic cognitive tasks into the explanatory framework of physical reality. Of course, complex cognition still eludes us, meaning that EC will also require a more abstract kind of explanation for certain cognitive tasks, but it is not limited to that kind of explanation. With a lot more research and knowledge, EC at least theoretically allows for cognition to be explained entirely through physical laws and material objects, rather than having to make the ‘hardware’ and ‘software’ distinction that is virtually universal among other materialist theories of cognition. In other words, it may solve the mind-body problem by eliminating the distinction entirely.

Chapter 3: Synthesis

Now that both relation R and EC have been explored, we can clearly see why they are incompatible theories. Relation R relies on the brain-centric psychological model of computationalism, while EC is explicitly a movement against that model. Relation R also starts from a static 'snap-shot' picture of the mind, while EC can only provide explanations of the mind over a certain minimum amount of time.

Of course, it's not exactly surprising that the two theories make different assumptions, they are meant as answers to different questions. On the other hand, the two questions are closely related: one can't fully explain human cognition without explaining our identity, and any theory of our identity must include some concept of how we think. In other words, relation R needs a theory of cognition, and EC needs a theory of identity.

3.1: Why EC and relation R make a good team

So, cognition and identity are closely interwoven, and any theory about one needs to incorporate an explanation of the other. Why should these two particular theories join forces with each other in this way, rather than merging with some other explanation?

First, let's explore why relation R would be strengthened if it were made compatible with EC. At its core, the theory of relation R is an attempt to redefine or replace the concept of identity for a reductionist view of the world. As I've argued above, EC (or at least its most mainstream forms) is in many ways more reductionist than computationalism. It relies less on abstract concepts like representation and does not require the new kind of dualism that is the hardware-software distinction. It also allows far more parts of cognition to be explained in terms of relatively simple mechanical interactions between the body and environment (including the social environment). Finally, EC also does not arbitrarily restrict the attribution of mental properties to particular pieces of matter, like the brain or a computer. Instead, the concept of cognition is truly reduced to the holding of a particular set of facts, namely being part of the kinds of feedback loops that compose what we call cognitive acts. These facts or properties are not restricted to things made of neurons or transistors: any kind of object could conceivably be part of a cognitive process.

However, even if moving towards a more reductionist view of cognition is not a desirable goal, or if EC simply turns out to be wrong, it can still be desirable for relation R theory to become compatible with EC. This is simply because EC's main objection against computationalism, that of unjustified brain-centrism, applies equally to relation R as Parfit presents it. Not only does Parfit explicitly deny the importance of the body (and implicitly, the environment), the psychological properties which he claims constitute connectedness are described as abstract and representational. Preferences and plans for the future are not described in terms of the acts in which they are expressed, but in computationalist terms of internal representation (Parfit 1971). Since reductionism does not require this brain-centrism in any way, eliminating this weakness can only make the theory of relation R more robust, provided that becoming compatible with EC does not mean other strengths are eliminated or new weaknesses are introduced.

Put simply: it is desirable for there to be a version of relation R that is compatible with EC. Primarily so that, if EC's ambition is fulfilled and it becomes the dominant theory of cognition, the valuable insights of relation R theory need not be lost. Additionally, creating such a compatible version will require us to differentiate between the essential insights of relation R and the superfluous elements, which will make the core of the theory clearer and therefore less likely to be undermined by a critique that actually targets a non-essential part of the theory. This consideration has already been made relevant for relation R by the accusation of brain-centrism mentioned above.

EC's position in regards to relation R, and indeed the PI debate as a whole, is somewhat different. Currently, there is no dominant philosophical theory of numerical identity within the EC movement. This restricts the movement's potential to explain human cognition, as so much of our cognitive activity is in some way rooted in our personal identity and those of others. But this restriction can be lifted by virtually any theory of identity, so long as that theory is consistent with EC theory as a whole. Unlike relation R, EC does not have any particular weaknesses or critiques that would be addressed by the merging or compatibility of the two theories. However, relation R does have a number of properties that make it a promising candidate for an EC theory of identity, provided that the two can be made compatible.

The first property is perhaps somewhat obvious: relation R theory is non-dualist (or in Parfit's terms, reductionist). It shares this quality with many other theories of identity, of course, but dualist conceptions of identity are still quite popular and eminently unsuitable for the EC movement.

More uniquely, relation R provides a non-discrete theory of identity. This is, in my view, the key insight of Parfit's theory: that the question of identity need not always have a definite answer. As we've seen in chapter 1, the assumption that identity must always be a discrete fact is a major obstacle for reductionist theories of identity. Where dualists can simply assume the existence of an indivisible soul or point of view, those who are committed to the material world are forced to reckon with the fact that matter can be divided or destroyed, and information can be duplicated and partially erased. This creates problems for reductionist theories like those of duplication and the spectrum, problems for which the solutions discussed by Parfit are unsatisfactory. From the EC perspective, however, these solutions are unsatisfactory for an additional reason: they all impose an additional level of abstraction onto material reality. When we draw a line to solve the spectrum problem, or devise a set of requirements for PI to solve the duplication problem, identity becomes exactly the kind of abstraction imposed on material reality that the EC movement is trying to get away from. In contrast to these solutions, relation R simply describes a comparison between specific material objects (that is, brains) and their physical properties at different times. No non-physical property like "is identical with some brain/individual X at some time Y" needs to be attributed to any brain or individual in order for the theory of relation R to work. This quality of the theory is very similar to EC's approach: attributing abstract properties to individuals is avoided in favor of sticking to purely materialist explanations.

Note also that relation R theory allows for 'grey areas' of identity, suggesting that it may be closer to physical reality than other theories of identity which do not. In nature, neat distinctions and separations are rare. Most often, like is the case with the distinction between species or between mountains and hills, they are purely a matter of human categorization for our own convenience. This does not mean that all theories which include grey areas are necessarily more true to nature than those that don't, but the acceptance of grey areas is a good sign if we are attempting to stay away from artificially imposed abstractions, like EC theory is trying to do.

Finally, relation R has another common property that is very desirable for EC: pure relationality. In relation R theory, the only thing that matters for identity is the relation between the two individuals in question. The strength of their R-relation does not depend on their relation to other individuals or trivial external events. Just like with the first property, this is not at all unique to relation R theory, but pure relationality is often abandoned by reductionist theories of identity. Often, like in Robert Nozick's theory of the closest continuer (Nozick 2013), these external requirements are added to identity in order to keep identity a 1-on-1 relation and avoid the problem of duplication. Because relation R sidesteps the problem of duplication, it does not need to give up pure relationality, meaning that identity can truly remain a property of the two individuals involved, rather than having extra requirements that are wholly external to these individuals. Again, this avoids unnecessary abstraction like the whole EC project tries to do. In addition, it also fits well with our common-sense view of personal identity as a property inherent to ourselves.

Aside from these technical properties, EC and relation R also share a more fuzzy or unclear quality: both embed us as human beings more strongly in our social environments. In EC, the people around us can constitute our minds by making certain kinds of cognition possible or more likely, like with the example of the sports fan and the community around their team. Relation R theory does something similar (Parfit 1984, 318-320), in proposing that our identity fundamentally just comes down to psychological similarity or connectedness. We are of course strongly connected to ourselves in this way, but also share weaker connections of the same kind with the people around us. Under relation R theory, our connection to other people can be the same as our connection to ourselves in the distant past or future. In other words, both theories allow for other people to be, in a sense, part of us.

3.2: Boundary conditions

Given the reasons mentioned in the previous section why combining EC and relation R is desirable, we must accept certain limits in how we can go about this combination. After all, the intention of revising the theories into a mutually compatible form is to preserve the strengths of both theories while resolving their conflicts, not to throw the baby out with the bathwater. These boundary conditions will underdetermine the form of the revised theory, potentially allowing for multiple competing revisions to exist.

Boundary condition 1: Degrees of identity
Core to Parfit's theory of relation R, and also its most novel insight, is the idea that what underlies identity is actually a relation between two individuals, which can hold to greater or lesser degree. Any revision that abandons this core insight may as well not be a revised version of relation R theory at all, but instead one of the other reductionist possibilities that Parfit undermines through the arguments of duplication and the spectrum. From the perspective of EC, this non-discrete view of identity is also exactly what makes relation R a promising theory, as I've argued above. Discrete identities are an unnecessary abstraction, and discrete categories in general raise suspicions of a disconnect from the material reality. Note that this boundary condition also means that another aspect will have to be shared between relation R and the revised form: the idea that identity can sometimes be an empty question.

Boundary condition 2: Temporally extended individuals
As explained in Chapter 2, most computationalist explanations of cognition can work on a 'snapshot' of the mind, a static picture or one that lasts a fraction of a second, where the brain processes information and manipulates symbols in the appropriate way. Relation R takes a similar picture of individuals, freezing them at one particular moment in time to compare their psychological properties to one another. EC fundamentally cannot conceive of identity in this way, because its model of cognition depends on the interaction with the environment that takes place over a longer span of time. Under any EC theory of identity, the individuals who are being compared must be examined over a certain minimum span of time. Put simply: our revised definition of identity will have to compare short videos of individuals, rather than still images.

Boundary condition 3: Multi-dimensionality
This condition applies to many, perhaps most reductionist theories of identity, though many compress the different dimensions or axes into a single measured quantity. Parfit's relation R, for example, makes comparisons between individuals on the dimensions of preferences, memories and intentions, and other psychological qualities are also occasionally mentioned. These different axes then get compressed into the single metric of relation R. Any EC theory of identity will have to incorporate more dimensions, even if it only intends to emulate

relation R and not expand upon it. Preferences, for example, can be conceptualized as a single axis in computationalism, with symbols for various activities and products spread out across this line. Under EC, any model of human preferences must also take into account gut flora, our belongings, our social circle, etc. This highly multidimensional nature is inherent in EC theory because it does not place restrictions on the types of objects that can constitute a cognitive process, and a single act of cognition can be spread out across many such objects. Of course, just like relation R does, these numerous dimensions can be compressed into one, or only a few. But regardless of the way this is done, many more dimensions will have to be considered at some point in the theory than is the case in relation R.

Boundary condition 4: Pure relationality
As explained above, pure relationality is one of the key properties of relation R that make it desirable for EC theorists. Maintaining this property while combining the two theories may seem difficult or even impossible. After all, EC theory centers around the idea that cognition is constituted in interaction with the environment, and allows any kind of object to carry mental properties. So, things that would constitute trivial externalities (like the destruction of an inanimate object) for relation R could be relevant to an EC picture of cognition, and therefore identity. However, this need not actually be a problem. Expanding the definition of cognition beyond the brain does not mean including everything else. The EC claim is that any object can be part of a cognitive process, not that every object is. An EC approach to the thought experiment with the duplication machine in 1.2 need not maintain that A plays any role in B's cognition, or B in A's. In fact, since the thought experiment requires that A and B never interact whatsoever, EC analysis of A's cognition would always ignore B entirely (and vice versa) as interaction is the very core of EC cognition. So, despite the fact that EC narrows the list of what can be considered trivial externalities by quite a bit, it can be compatible with pure relationality and avoid some of the nonsensical answers to the duplication problem.

3.3: A suggested definition

With these boundary conditions established, I wish to finally proceed to my own suggestion for an embodied/embedded/extended version of relation R, which I shall designate as ERR (Embodied Relation R).

Just like Parfit's relation R, ERR is fundamentally a comparison of a number of key properties that establishes the degree of similarity between two individuals. This is one of the great strengths of relation R, allowing it fulfill boundary condition 1 and deal easily with grey areas. However, unlike relation R, ERR cannot compare only abstract, internal symbols representing memories or preferences. Instead, it looks at actions and interactions between an individual's body, brain and environment (both material and social). For example, where the relation R approach might try to establish an individual's favored flavor of ice-cream by asking them or conducting brain scans, an ERR approach could observe them going through the process of choosing a flavor at an ice-cream shop. Of course, answering questions about food preference or entering a brain scanning device are also interactions that could be observed, but these are not interpreted as special interactions capable of directly representing internal symbols, merely as another set of interactions that could be relevant to an individual's personal identity.

Relation R establishes the degree of similarity between the memories, preferences and other mental properties of two individuals. ERR starts with a very similar comparison between the behaviour of both individuals in similar scenarios. If we wish to establish if B is the same person as A, we would observe B interacting with A's spouse or friends, playing A's favorite sport, working at A's job, reminiscing about A's youth and so on. The degree of similarity between A's and B's behaviour could be expressed as a value on a scale.(Preferably, both the scale and observations would be made by experts in human behaviour.)

These various observations and the values given to them can be represented in a mathematical arrangement, a multidimensional topology I will call P-space. Simply put, P-space is a map with a dimension for every observation made, and values for each of these observations is a coordinate in the appropriate dimension. Theoretically, given enough observations of the kinds of actions that distinguish individuals from one another, a set of coordinates could be produced that uniquely describes every individual and exactly how they act differently from all other individuals.

Of course, not all differences matter to the same degree. An individual's identity is likely to be far more impacted by their preference in romantic partners than their preference in pizza toppings. P-space itself does not represent this difference, but we can assign weights to different dimensions to represent their varying importance (and if we assign a weight of 0 to any dimension, we can ignore that observation entirely).

Given this definition of P-space, we can define ERR itself simply as:

The weighted total distance between two individuals in P-space.

In other words, the sum of the differences between two individuals in a dimension weighted by the relevance of that dimension. Expressed mathematically as:

$$ERR(x,y) \ = \ \sum_{d \in D} (x_d - y_d) \ \times w_d$$

Where x and y are individuals, D is the set of all dimensions, w_d denotes the weight of dimension d and x_d and y_d denote the value of x and y, respectively, in dimension d.

With this mathematical expression, the result is that ERR is described by a number that is somewhere between 0 (the distance between perfectly identical individuals) and infinity. In practice, there would likely be some upper limit, as human beings can only differ so much from one another. The limit could even be lower than the maximum human divergence, depending on the scale used during the observations.

So why do I propose this model as the definition for ERR, rather than some other model that also satisfies the boundary conditions mentioned above?

Firstly, the precise numerical value that can be calculated for an ERR distance is very useful in the context of a courtroom or policy-making. Just as with relation R, ERR itself allows for grey areas and can comfortably accept the existence of a spectrum of identity without having to draw lines anywhere. But ERR has the major advantage that it can more easily produce actual numbers for its spectra, which we can use to draw hard lines when we need them.To return to the age-related wisdom analogy: it's a lot easier to establish an age of majority if there's a calendar that can assign a number to someone's age.

Secondly, exactly which observations are incorporated into the ERR distance calculation is not set in stone. I am not going to attempt to provide an exhaustive list here, preferring instead to leave this concern to future empirical research by those with more expertise in the behavioural sciences. This not only means that courts and similar institutions can set their own policies, but also that ERR can easily incorporate future shifts in EC theory and possibly even survive paradigm shifts, if the right adjustments are made to the observations that make up P-space. It even allows for different philosophers to create competing lists, or for the relevant observations and weights to be tailored to specific individuals.

Finally, while ERR can be applied rigorously as described above, with expert observation and hard values, it can also capture the practical way in which we approach personal identity much more easily than computationalist models do. To illustrate this, let's turn to one of the rare real-life cases where personal identity is disputed: that of Phineas Gage.

Phineas Gage was a railroad construction foreman in the US. On September 13 1848, an accident occurred that shot a tamping iron (a round bar 3.2 cm in diameter and 1.1m in length) entirely through his head. The metal bar pierced his mouth and went through his brain and skull behind the eyes. Despite the traumatic injury, Gage lived and made a remarkable recovery. Less than 6 months after his accident he could travel, perform light labor and talk. In the long-term, the only injuries he had left were blindness in the left eye and partial paralysis of the left side of his face.

However, people in Gage's life have testified that he was "no longer Gage"(Harlow 1993, 277). His memories were only very mildly impaired, but his personality had changed radically. His doctor describes him as a responsible, polite, hard-working man who is popular with his crew, with a 'well-balanced mind' and diligent in pursuing his plans. After his accident, though the same doctor notes that his memory and general intelligence have remained unimpaired, he is described as rude, engaging in profanity, unable to control his passions or see any future plans through.

The prominence of the brain and mental properties in Gage's case should make it an excellent case-study for relation R analysis. Of course, Parfit's theory can be applied here, noting that Gage's preferences have changed a lot, while his memory seems to have changed little. But such an analysis does have a problem: it relies on proxies. The degree to which relation R holds between pre-accident Gage and post-accident Gage depends on his psychological qualities, the symbols in his mind that make up his cognitive processes. Determining these qualities would have to be done with a brain scan or the closest equivalent available in 1848: a long list of questions to be asked by a doctor. But of course, unless this same list of questions was asked both before and after the accident, comparison of the answers is impossible. So, a relation R analysis (or really, any computationalist analysis of Gage's identity) depends on a proxy (eyewitness accounts and speculation) of a proxy (Gage's answers to the questions) of the true measurement (Gage's internal psychological qualities). Not only does this introduce a lot of margin for error, it's also clearly not how Gage's friends and relatives actually came to their conclusion that he was no longer himself.

ERR, on the other hand, can be quite easily applied in these kinds of practical cases. Gage's friends and family did not ask him a long series of questions, but they did observe his behaviour in lots of different circumstances throughout his daily life. They could see and remember that he never cursed and made intelligent plans which he saw through with discipline before his accident. As such, they could compare his behaviour after the accident with what they observed before, and note the changes. If a list of dimensions to consider and guidelines for assigning the values were provided, with proper guidance from experts, it seems likely that his friends and family could have provided us with a rough ERR-distance value between Phineas Gage before the accident and the Phineas Gage they saw afterwards. While their values may be a little less accurate than those of an expert in human behaviour, they could still directly report on the relevant unit of measurement for ERR: interaction with the environment. And though they may not have assigned values to their observations, this is exactly what historical sources tell us the people around him did. They saw Gage's behaviour change so much that 'his society was intolerable to decent people', and decided based on this observation that he was no longer the same person.[2]

3.4: Summary

[2] Note that there are some doubts about the accuracy of the commonly cited Gage story, but also that these do not undermine the point illustrated by the story. Even if the story were entirely fictional, it still illustrates that we assess the identity of our friends by observing their behaviour directly, not by filling out a mental questionnaire regarding their internal mental states before and after the change.

In chapters 1 and 2 we have described relation R and EC theory and established that the two theories carry incompatible assumptions. I have argued that, through the lens of the historical anti-dualist movement, the two theories nevertheless make good allies. For relation R, incorporating EC elements can be a way of future proofing in case cognition turns out to be less symbolic and computational than Parfit believed. For EC, relation R provides a theory of identity that fits well within the practices and framework of the EC movement.

From there, we argued that a successful merging of the two theories is possible in spite of the incompatibility that arises when they are compared in their base forms. In order to do this, we first singled out the key features of each theory that must be preserved to make the effort worthwhile. Those key features provided certain boundary conditions that any proposal for a merged theory would have to abide by. Within these boundary conditions, there is the possibility for numerous theories which combine the valuable features of both EC and relation R.

Finally, we created a proposal for such a theory: ERR. ERR accommodates all the boundary conditions, without being subject to any of Parfit's or the EC movement's major criticisms of their dialectical opponents. It allows for hard lines to be drawn when they are desirable, but happily accepts the existence of grey areas in the spectrum of identity. Which observations should be includes in P-space remains somewhat vague (though no vaguer than Parfit's collection of psychological properties), which allows ERR to flexibly accommodate different interpretations of what matters for identity. Perhaps most importantly, it captures how people actually tend to make judgments about personal identity in real-life cases, while not simply describing our (often flawed) everyday reasoning about this matter.

Bibliography

Bredo, E. “Reconstructing educational psychology: Situated cognition and Deweyian pragmatism” *Educational Psychologist* 29 no. 1 (1994): 23-35. doi:10.1207/s15326985ep2901_3

Chestnutt, J., Lau, M., Cheung, G., Kuffner, J., Hodgins, J., and Kanade, T. “Footstep planning for the honda asimo humanoid.”*Proceedings of the 2005 IEEE international conference on robotics and automation* (April 2005): 629-634.

Clapp, M., Aurora, N., Herrera, L., Bhatia, M., Wilen, E., and Wakefield, S. "Gut Microbiota’s Effect on Mental Health: The Gut-brain Axis." *Clinics and Practice* 7, no. 4 (2017). doi:10.4081/cp.2017.987.

Collins, S., Ruina, A., Tedrake, R., and Wisse, M. "Efficient bipedal robots based on passive-dynamic walkers." *Science* 307, no. 5712 (2005): 1082-1085.

Harlow, J. “Recovery from the passage of an iron bar through the head.” *History of Psychiatry* (1993): 274–281. doi:10.1177/0957154X9300401407

Huebner, B.. "Socially embedded cognition." *Cognitive systems research* 25 (2013): 13-18.

Nozick, R. "Personal identity through time." *Personal Identity.* Oxford: Blackwell (2003): 92-115.

Parfit, D. “Personal Identity.” *Philosophical Review* 80, no. 1 (Jan 1971): 3-27.

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

Ravenscroft, I. *Philosophy of Mind.* Oxford: Oxford University Press (2005).

Shapiro, L. *Embodied Cognition*. New York: Routledge, 2011.

Shapiro, L. “The Embodied Cognition Research Programme.” *Philosophy Compass* 2 (2007): 338-346. doi:10.1111/j.1747-9991.2007.00064.x