5 Mind

Having discussed in Chapter 4 what causation is, I can continue in this chapter describing how the mind can be understood causally. There are different philosophers, for example agent causationists or substance dualists, who will reject that we can understand the mind as a normal causal process. In order to argue that it is superfluous to include irreducible agents or souls in one's ontology, I must present a detailed causal understanding of the mind. I must present in detail the topics of mind, thinking, consciousness and free will (each in an independent chapter) in order to show how it is possible that persons thinking and making free choices can be ontologically reduced to causal processes between values actualized in fields.

In this chapter, I will argue that our mind functions like a causal process in the same way as other causal processes work in nature. It may sound very strange to believe that the mind functions as a causal process, and many will argue that when we are acting for reasons we are doing something very different from a causal process (McDowell, 1996). But as I shall show, many philosophers also believe that our mind is a causal process, and it is very difficult to understand how the mind works if it is *not* a causal process. When a person considers alternatives and end up choosing one, which she then acts upon, how does this happen if it is *not* a causal process?

I start in Section 5.1 with some biological background to describe how mind has evolved. Such an evolutionary explanation of mind supports a causal understanding of the mind, since it is then a product of natural causal processes.

Sections 5.2 and 5.3 deal with the relation between brain, mind and consciousness. Mind is understood as brain activity that can become conscious. In Section 5.2, we look at different arguments supporting the belief that the brain causes the content of consciousness. In Section 5.3, we look at many different examples of how almost anything we do consciously can also be done non-consciously by the brain. That mind can occur non-consciously in the brain also supports the view that mind is a causal process.

Sections 5.4 to 5.7 look at different components that are part of a choice in order to start understanding how a choice can be a causal process (and later how it can be said to be free). The components are emotion (5.4), memory (5.5.), the self (5.6), and desire (5.7). Especially important is the self, where I use neuroscientist Antonio Damasio to distinguish between the autobiographical self – which is a storage of memories in the brain – and the core self – which is a stream of conscious impulses.

https://doi.org/10.1515/9783110770957-007

In this chapter, I will be leaning quite much on the book *Self Comes to Mind* by the neuroscientist Antonio Damasio. It is a good book and allows for a concise presentation with much support from neuroscience. I shall start with some biological background which will show itself relevant later. I believe that showing how the mind could arise through evolution also helps to support the view that the mind is a causal process.

5.1 Biological background

Although many questions remain unanswered, scientists have quite detailed theories which give a good explanation of how biological life could arise from chemical components. I have in mind especially the RNA world hypothesis developed by, among others, Jack Szostak, Leslie Orgel and Gerald Joyce. The general idea is that simple reactions can cause stable copies of itself.[89] Gradually they could form into RNA molecules, and scientists have been able to replicate RNA from the same ingredients and conditions assumed to have been present on earth 3.5 billion years ago (Lincoln and Joyce, 2009). The exciting potential in RNA is how it can store information, make copies of itself and catalyze certain chemical reactions to make, for example, proteins. RNA, DNA and ribosomes do crucial work in cells, and they are all made of RNA, which suggests a gradual natural process from chemistry to biology.[90] One possible way to go from strings of RNA molecules to cells is that they could have merged with fat bubbles to become simple cells, which can also divide under certain conditions (Budin and Szostak, 2011). This would be some of the important steps towards cells dividing and containing DNA.[91]

Already at the level of a cell, one can see how it resembles a simple version of a larger organism: the cytoskeleton is like the skeleton, the nucleus is like the brain, the cytoplasm is like the tissue and organs, the cilia are like limbs. Like organisms, cells need to get nutrition in and waste out, and turn the nutrition into energy that can be used for reproduction and getting more nutrition in and waste out (Damasio, 2010, pp. 33, 41).

---

89 As suggested for example in Dawkins (1976, pp. 13 – 20), and illustrated well by Terrence Deacon, in Deacon (2006).

90 That ribosomes are ribozymes made of RNA is a recent discovery supporting the RNA world hypothesis (Cech, 2000).

91 There is a nice series of videos on YouTube from ibiology.org where Jack Szostak explains all this in detail.

The evolution of the senses is also well understood. I mention this also briefly to support the general evolutionary approach to our mental life as a causal process. Cells sensitive to light evolved into cells in a cavity which could then register the direction of the light. A lens sharpened the signal, and cells reacting to different wavelengths made further discrimination possible. Eyes have evolved independently many times in evolution. The nose started as cells sensitive to chemical stimuli and pheromones and, since it is sensible not to eat everything that comes into the mouth, taste developed in a similar way to smell. Hearing started as reactions to vibrations in the jaw, which evolved into the middle ear, and of course nerve cells in the skin are sensitive to touch.[92]

When cells start to cooperate, everybody can enjoy the benefits of specialization. In organisms, different cells do different kinds of work and get nutrition via the blood system (Damasio, 2010, pp. 33–34). Neurons are a kind of cells which have the specialized ability to change other cells by sending an electrochemical signal (called "firing"), by which they can move the muscles of a whole organism and help the organism survive by moving around (Damasio, 2010, pp. 37–38, 50).[93]

Two more very important things that neurons started doing were to make simple representations and simple dispositions. For example, some neurons could react to something poisonous by firing and activating some other neurons, making the organism move away from the poison or spit it out if it was an organism with a mouth. If this happens regularly, we could say that the first neurons firing in response to the poison represent the poison, and the second set of neurons firing could be said to actualize a disposition to move. "Represent" should here be taken in a minimal sense to just mean a consistent relation between the poison and the neurons that fire, and "disposition" means an "If presence of A, then act by doing B" mechanism, which is triggered by a specific stimulus and gives a particular response (Damasio, 1999, p. 320; Damasio, 2010, p. 134).

Through evolution, representations and dispositions would become more and more advanced. This would then help the organism to maintain body functions and to reproduce. We shall now look more into how the brain could make such representations, and see that we have good reason to think that it is causal processes in the brain that gives us the content of our conscious experiences. The description of the evolution of mind in this and the following sections fits very

92 For details, see Joseph (1996, pp. 7–16).

93 This is the classical description of neuronal interaction, but such interaction might also happen in other ways, such as through coherent oscillation (Fries, 2005).

well with the evolution of more and more advanced computers and robots (Nilsson, 1998).[94]

5.2 Does the brain cause conscious experiences?

It would be very helpful to survival and reproduction if neural patterns could represent states of affairs in the world, like food, drink, attacking lion, or potential partner interested in sex. Damasio argues that neural patterns represent the body and the world, and we shall look at how he argues that there is a lot of thinking, feeling, remembering, and desiring that happens non-consciously. We shall look at some examples of this, since it supports the idea of mind as a causal process. In the next section, I shall also discuss whether terms like thinking and feeling should be reserved for conscious thinking and feeling only, as some argue.

Now follow some examples to support the view that the brain causes the content of our conscious experiences. Every time you have a conscious experience of seeing red, the same area of your brain is active. If that area is destroyed, you will not experience red anymore, and if that area is stimulated, you will experience seeing red even if there are no red objects in front of you (Hadjikhani, Liu, Dale, Cavanagh, and Tootell, 1998; Wandell, 2008). If it is destroyed you will even have problems imagining something red (Damasio, 1994, p. 101). Damage to an area of the brain called the fusiform gyrus of the temporal lobe causes face blindness, and stimulation of this same area causes people to see faces spontaneously (Shermer, 2012).

In an experiment carried out by Damasio and colleagues, they found a pattern in the brain which consistently correlated with the conscious experience of a certain sound. What is interesting about this experiment is that the same pattern was active even when the person was just imagining hearing the sound in his mind, even if no sound was actually made in the world outside his mind (Damasio, 2010, p. 134).

The correlations are quite exact, indicating a close connection between the brain activity and the conscious experience. However, correlation does not necessarily imply causation, as two things can be correlated without one being the cause of the other, like the correlation between night and day. But correlation can also indicate causation. Do we have any reason to believe that it is the neural pattern which causes the conscious experience? Yes, because we can stimulate

94 See especially page xix.

neural patterns in the brain and achieve an effect in consciousness from it, which indicates that the effect in consciousness is causally dependent on what happens in the neural patterns.

Many problems are thus solved if there is a neural pattern underlying conscious experiences. Hallucinations are explained: they occur since a neural pattern is activated even if what it represents is not present. For example, you can think that you see a red tomato even if no such tomato is present because you have a neural pattern representing a red tomato in your brain, making it conscious to you. Phantom pains are also explained, since neural patterns representing (now lost) parts of your body can become conscious to you even if you have lost the actual limb. Certain phantom pain phenomena are very well explained by this model. For instance, there was a person who lost his arm but kept feeling it. Then he lost the feeling in his arm, but retained the feeling of his hand, now just feeling that the hand was sticking out from his shoulder. Finally, he lost that feeling as well. This is explained by the fact that the area representing our hand in the brain is much larger than the area representing the arm, so the neuron connections representing the arm faded away before the neurons representing the hand (Joseph, 2006, 18:13 – 19:34).

The process of how the brain can represent the world is well understood, especially vision. Humans have feature-detecting neurons, which fire in response to a certain feature being seen in the world. These neurons fire for at least thirty different types of features, like angle, size, movement, contour, color, distance from observer, etc. (Imbert, 2004, p. 39; Pinker, 1997, p. 20). Wolf Singer has shown that when neurons that fire in response to certain objects fire synchronously, the object is consciously experienced. If two completely different visual inputs are given to each eye, only one of them becomes conscious to the observer at a time, and it switches back and forth which image is conscious to the observer. When the first object is consciously seen, then the neurons detecting the features of that object fire in synchrony. When the other object is consciously seen, then the neurons detecting the features of that object fire in synchrony (W. Singer, 2004a, p. 25).

The brain puts together information from the feature-detecting neurons to make a unified picture. We know this because sometimes some types of neurons do not function, yet the brain creates a unified picture of the rest of the information. One case is color blindness. In another example, a person (known as D. F.) was almost blind, but she did receive information from the neurons detecting color and texture. Everything she saw was blurry, but she could see a banana

and guess that it was a banana because of its distinctive color and texture, but could not say what position the banana was in or what shape it had.[95]

Note that when there is a lack of visual input it does not create black holes in the visual field, but rather the rest is turned into a unified picture since the brain creates a unified impression based on the input it gets. Some people with brain damage in one hemisphere do not lose half of the visual field, but rather create a whole visual field from the input sent to the one hemisphere. The point is that the input is spread out to create a unified impression. If the brain is given contradictory sensory input it will merge it together in a kind of compromise (Gazzaniga, 2009b, 50:08 – 51:55 and 51:00:40 – 51:01.40). Various tests have been performed whereby people are given one visual input but feel something else. For example, they sit on a chair rotating 120 degrees while being shown a film indicating that they rotate less; or a big object is placed before their eyes while they feel a similar, but smaller, object with their hand. Their brains then merge the information together in a compromise. When they do not receive a disturbing visual input, they guess quite well how much they have rotated or how big the object is. But when they have disturbing visual input, the brain mixes the information together to form one impression so that they estimate the size differently (Viaud-Delmon and Jouvent, 2004, p. 73).

None of these examples show that the physical side of reality is ontologically more basic than the non-physical conscious side of reality, which idealists hold to be ontologically basic. For example, it could be that a non-physical conscious mind requires a very complex physical structure to interact with before it can become active. But consciousness does not seem to be an independent or soul-like entity as envisioned by most substance dualists or idealists. Rather, it seems very dependent on the brain, thus at least slightly indicating an ontological priority to the physical. Many experiences are better explained as made by the brain for the sake of survival than as an accurate depiction of the world by the mind. Things do not have colors independent of light setting or of someone watching them. But adding colors to objects makes them easier to see and distinguish, so it makes evolutionary sense that the brain should add colors. The spectrum of light that we can see is the spectrum that most of the

95 This phenomenon is further complicated by the fact that we seem to have two distinct visual systems in the brain, where one allows conscious seeing and the other is concentrated on adjusting body movement. Therefore, although D. F. was unable to see a letter box, she had no problem sticking a letter into the slot. The converse is Balint Holmes syndrome, where a patient can see the letter box clearly but would be unable to stick the letter in, see Carruthers (2006, pp. 88 – 89).

radiation from the sun and stars come in (Fernald, 2001),[96] so again it makes sense that evolution used this spectrum of light as a basis to make conscious experiences of color.

Sometimes the process goes wrong. For example, there are reports that some people see colors when they hear sounds (Gray, 2004). Neurologist V. S. Ramachandran and philosopher W. Hirstein suspect that such stories may be more metaphorical than real color experiences, but they cite an even better example. A person lost his sight; he became totally blind. But after a while he could start to see clearly the objects he was feeling in his hand – not just imagining them in his inner eye, but having an experience which was like seeing the ruler he was holding in his hand. This happened with all kinds of objects (Ramachandran and Hirstein, 1999, pp. 96 – 97).

Another fascinating experiment shows something similar with blind and blindfolded people: cameras sent output stimuli on either their back or their tongue, and these stimuli followed patterns consistent with what the camera was filming. After amazingly little training, both blindfolded and blind people reported that they started seeing images, and they were able to recognize faces, describe objects, read, manipulate objects and much more. Several times when the camera suddenly zoomed in on something, the blind(folded) people ducked because they felt that something was being thrown at them (Sampaio, Maris, and Bach-Y-Rita, 2001; B. W. White, Saunders, Scadden, Bach-Y-Rita, and Collins, 1970).

There are also many things we see that are not a correct picture of what we see, and it is useful for us to see things like this. Many optical illusions are based on the fact that the brain distorts what we see to make it fit what it should look like.[97] One example is the moon, which looks a lot bigger if it is close to buildings or mountains than when up in the sky, but there is nothing physical that makes it look bigger. The illusion happens because we know that the moon is much bigger than houses, so the brain makes it look relatively bigger.[98] The illusions are generally useful clues for survival but not correct depictions of the world. If our conscious minds were ontologically unique non-physical entities with capacities for grasping the world, one would not expect these distortions caused by what is accessible to the brain and its survival value.

---

96 More specifically, it is what the first animals in the sea would have been most exposed to in the water.

97 Michael Gazzaniga shows some very convincing examples in Gazzaniga (2009c, 29:30 – 30:17).

98 At least that seems to me to be the best theory (Kaufman and Rock, 1962). See also the similar evolutionary explanation of the Müller-Lyer illusion, in Sternberg and Mio (2006, p. 117).

To sum up this first point, it seems clear that the brain creates patterns that can become conscious, and that it is the physical side of it that determines how the conscious experience comes to be. The next question is whether or not there can be mind without phenomenal consciousness – that is, non-conscious thinking, remembering, feeling, desiring, etc.[99]

5.3 Non-conscious mind

Damasio argues that the brain constantly makes many neural patterns that represent states of affairs in the world or in the body, and these neural patterns can be consciously experienced as images, but he also says that most of them are never experienced consciously. But even if they are not conscious, neural patterns and dispositions seem able to perform their work as a causal chain and do the same work as a conscious person does (sense, think, remember, feel, desire) – often even better. Numerous experiments support this, and I do think that much philosophical confusion could be avoided if more philosophers accepted that there can be conscious and non-conscious thinking.[100]

Let us take an example from daily life. Many people have found that they can sometimes drive a car "on automatic pilot" while thinking about something other than driving. But even if they are not conscious of seeing signs, red lights and so on, their driving indicates that the signs and lights have been registered as such and acted upon (Armstrong, 1981, p. 59). A more astounding example is people who are blindsighted or deafhearing. They have no conscious experience of ever seeing or hearing anything, and yet they can move well through a labyrinth, catch what is thrown to them, move towards where a sound comes from and so on (Carruthers, 2006, pp. 87–88).

Priming is another good example. People can be shown words or pictures on a screen so quickly that they have no conscious idea what the picture or the word was, but testing afterwards shows that they must have registered and understood

99 When I use the term "non-conscious" I do not mean "not awake" but "not phenomenally conscious". I give a fuller description of what this means in the chapter on consciousness.

100 For example, Mark Rowlands argues against thinking as consisting of images in the mind since one can say that the glasses are in the drawer without thinking about the glasses in the drawer, but rather thinking about the game on TV (Rowlands, 2003, pp. 78–79). In this case I think it is obvious that the person is thinking at least non-consciously about the glasses in the drawer, and cannot see any other plausible explanation of how he is able to answer the question of where the glasses are.

the words or pictures (Sternberg and Mio, 2006, pp. 64–65).[101] Damasio uses the example of a cocktail party: You are listening to your conversation, but the brain registers other conversations as well. Suddenly you hear your name or something else in another conversation which is marked as important so that you become conscious of it, and you start listening to that other conversation (Damasio, 2010, p. 173).

These examples are mostly concerned with non-conscious *sensing*, although some interpretation is also involved. Below I look at emotions and the self, thinking and desiring, and give examples of non-conscious feeling, thinking and desiring, but here is an example which shows complex non-conscious reasoning. Damasio and colleagues performed an experiment where people were asked to draw cards from various decks: some decks were good, i.e. leading to a reward, and some were bad, i.e. leading to a punishment. There was also a system determining which decks were good and which were bad, so that if you cracked the code you could just draw good cards. The subjects played the game while their skin conductance was measured. The interesting thing was that it seemed that the code was cracked non-consciously several minutes before the players understood it consciously and before they started drawing only winning cards. After a while they would get one type of skin response just before drawing from every bad deck, and another type of skin response just before drawing from every good deck. This was so consistent that somehow some part of the brain must have cracked the code, but the person could not consciously tell this and would keep drawing bad cards (Damasio, 2010, p. 276; Bechara, Damasio, Damasio, and Anderson, 1993).

Is it right to use words like "see" and "think" in contexts other than conscious seeing or thinking? This raises the question of first-person and third-person descriptions of events. Although some kind of identity theory about mind and body might be right, the most common view is that images in the mind and patterns in the brain are not identical (Kim, 2006, pp. 112–113). They seem to have many different properties, so the prima facie view should be that they are not identical. Even if they are identical, it is in any case helpful to distinguish between the first- and the third-person perspective. For this reason, I shall specify whether I mean conscious images or not when I write about images.

What about "seeing" and "thinking" and words like that, which usually presuppose a first-person perspective? There has been much philosophical critique

101 This is a well-established fact, and the reason why many commercials use subliminal stimulation.

of neurologists who use first-person language to describe what happens in the brain. Many neurologists speak of the brain as a person and say that the brain does things we normally just say that people do, like "seeing", "remembering", "interpreting", and "mapping".[102] It is important to be clear about the distinction between first-person and third-person perspectives when one is writing about the brain and mind. But it can also be very difficult. The reason is that humans can do so many things non-consciously in the same way as when they do them consciously. We can see, hear, smell and so on without ever being conscious about what we see, hear and smell, yet we act as if we have seen, heard or smelled it.

Many examples were given above, like blindsight or driving without paying attention. A telling example is that of split-brain patients, where the connection between the two hemispheres of the brain has been cut so that there is no interaction. If you flash the word "spoon" to the left eye only, so that the visual impression is sent only to the right hemisphere, then ask the person, "What did you see?", she will answer "nothing", since the left hemisphere is where most people have their language modules, and obviously nothing was registered in the left hemisphere. But if the person is allowed to put her left hand (which is controlled by the right hemisphere) in a box, she will feel around and pick up the spoon, indicating that the right hemisphere did see the word spoon and understood it.[103] In examples like this it is very tempting so say that one hemisphere saw the word and the other did not and that the left hemisphere said that it did not see it, whereas the right hemisphere cannot speak. But usually we just use words like "see" and "speak" about people, not about cerebral hemispheres.

John Searle criticizes Damasio for saying that these non-conscious representations are part of the mind. "What is the fact that makes them mental?", Searle asks, suggesting instead that they could be understood as a non-mental step on the way to consciousness (Searle, 2011a). I would answer that the fact that makes them mental is that they are causally related to the objects in the world that they represent and internally related to other representations in a structurally similar way to how objects in the world relate to each other. Several reasons have been suggested for viewing these non-conscious representations as parts of the mind, largely because they perform functions we usually call mental. When a brain process can go on non-consciously and have all the same effects as a conscious process, that makes it reasonable to think of it as part of the mind.

102 A list of examples can be seen in M. R. Bennett (2007, pp. 154 – 156).

103 A film of this experiment can be seen in Gazzaniga (2009a, 25:30 – 28:00).

People may act as if they were consciously sensing and thinking and yet they are not doing it consciously. The reason they can is that the brain creates neural patterns that represent what happens in the world, and this triggers dispositions that the brain works according to. This happens all the time to everyone, and we are conscious only of a few of the things that go on in the brain. It is possible to describe it in uncontroversial third-person language, but it is much more efficient to say simply that a person "sees" the object instead of saying that the object "activates a representation of the object".[104] For instance, in the card game test described above I said that the brain cracked the code before the players did, since the skin conductance always matched the good and bad decks even if the players drew the wrong cards. The expression "crack a code" is usually used for something a conscious mind does, but in this case it seems a very appropriate and efficient description of what has happened.

It is also difficult to distinguish clearly between first-person and third-person language since there are so many words that have been used as metaphors so frequently that they acquire a literal meaning. For example, it is common practice to write that "an argument shows", but one could complain that only people can show something, depending on how the word is defined. Above I wrote that "the brain creates neural patterns", and again one could complain that only people can create something in one definition of "create". The term can also be used, however, to mean something like "to cause".

So, on the one hand it is difficult to separate first-person and third-person language, and of course I want also to show how similar conscious and non-conscious brain processes are, since I think that what we usually call sensing, thinking, etc., while implying a conscious first-person perspective, may well happen non-consciously and be correctly described in a third-person perspective. It is my aim to show the close connection between a phenomenological description from a first-person perspective and a neurological description from a third-person perspective. On the other hand, I do want to keep the distinction since I believe that consciousness and the subjective perspective make a difference, and we shall come to a discussion of what difference consciousness makes. My solution to the problem of first-person and third-person language is to use terms like thinking, feeling and so on in third-person perspective description but specify whether I refer to conscious or non-conscious activity in the brain. Hence I shall distinguish between non-conscious thinking and conscious thinking,

104 A neural pattern being *activated* means that the neurons that actualize the pattern are firing, and then the representation they constitute *occurs* in the mind. Activation need not mean that it is conscious, however, as I argue with different examples of conscious and non-conscious mental events.

non-conscious feeling and conscious feeling, and so on. To those who think that these terms are meaningless when referring to non-conscious events I would say that the examples show that the descriptions make sense after all.

Some brain activity never becomes conscious, like for example autonomous processes in the body, and as far as we know they are *not* consciously experience*able*. Some brain activity is not conscious, but can become conscious (*is* consciously experience*able*, but non-conscious), and some activity becomes conscious (*is* consciously experien*ced*), and we know quite a lot about where the different things happen. It is quite specific, so that certain areas of the brain can create conscious experiences and activity in other parts of the brain never becomes conscious. Common to the areas that create conscious experiences is that they are complex clusters with massive interconnectivity organized around a gate of input from the world outside or the body, and this fits well with the view that the brain at a certain level of complexity creates conscious experiences out of its input (Damasio, 2010, pp. 86 – 87).

5.4 Emotion

In the previous section, we considered arguments in favor of thinking that there can be non-conscious sensing and thinking, which again supports the view that they can be understood as a causal process. In the next four sections (5.4 – 5.7), we shall take a closer look at emotion, memory, the self and desire, and see how they can work non-consciously and consider reasons to think of them as causal processes. An extra reason why it is important to understand how these work is that they all play an important role in the process of deliberation. Understanding how they work will help us understand how free will can be a causal process. We start here with emotion.

We have already seen how small organisms can have life-preserving reaction patterns, like spitting out something poisonous. Another example is how baby birds react to a large shadow flying over them, which makes them huddle and sit still (Damasio, 1999, p. 69). There are many examples where animals seem to act without conscious thinking or feeling and just react with simple dispositions resulting from causal stimulus and response.

The automatic response to stimuli has evolved into the advanced responses that we call emotions. In advanced organisms the brain contains a representation of the body in homeostasis (functional balance) which can be compared with a representation of the body as it is now. Those brains which could detect a difference and make something happen to restore the body to homeostasis would be selected by evolution (Damasio, 2010, pp. 48 – 49). An example

would be to detect low blood sugar and create hunger to make an animal eat. This could happen without consciousness, since all that is required are some "If presence of A, then act by doing B" dispositions in the brain (Damasio, 2010, p. 52).[105]

Another advance made by organisms was the evolved ability to detect likely threats (e. g. animal with sharp teeth approaching) or likely delivery of goods (e. g. partner preparing to have sex). It is important for the organism to have rules on when to move and some motivation that actually makes the organism move. This is achieved by the brain sending molecules through the blood vessels and signals through the nerves to warn the organism and prepare the right response. This is an important part of what emotions are about, and fear is a good example: something is registered as threatening, and the brain sends out molecules that prepare the body for fight or flight. In this way, emotions allow for more differentiated and optimized responses to stimuli than automatic action-responses (Damasio, 2010, p. 54).

Damasio distinguishes between emotions and feelings. An emotion is a series of events happening in the body. A neural pattern representing something in the body or the world outside activates trigger regions of the brain, which sends out various chemical molecules in the blood and different signals through the nerves. This prepares the body for certain actions and usually also triggers certain kinds of thoughts. As Damasio says, running from a gunman, you do not think about what to make for dinner tonight (Damasio, 2010, p. 144). An *emotion* is therefore a series of events which places the body in a certain state. A *feeling*, on the other hand, is a term Damasio uses to denote a neural pattern in the brain representing the body's being in that state of emotion. It is a neural pattern representing the body's being, for example, in a state of fear or happiness or anger. This neural pattern may become conscious or not, which means that we need to distinguish between emotions (a series of events in the body), non-conscious feelings (a neural pattern in the brain) and conscious feelings (a consciously experienced feeling/neural pattern) (Damasio, 2010, pp. 109 – 110, 114).

Only conscious feelings are experienced from a first-person perspective. It may seem strange to speak about non-conscious feelings, but again there are good reasons to distinguish between them. Non-conscious feelings can become activated and change our body state before we become consciously aware of what we are feeling. Men can be shown pictures of naked women so quickly that they are unable to consciously experience it, and yet their bodily reactions

105 For examples, see how Peter Carruthers explains very many workings of our mind by employing such "If A, then B" dispositions in Carruthers (2006).

are as if they had seen them consciously (Koch and Tononi, 2014, 10:40 – 11:32). Several other examples could be given.

The distinction made here is that between emotions as body states and feelings as neural representations of that body state, which may or may not become conscious. There are some universally recognized emotions, namely happiness, sadness, anger, fear, surprise and disgust (Damasio, 1999, p. 50). Several of these emotions have their own distinct physical patterns in the body, meaning that a scientist with the right apparatus can to some degree know what you feel without your telling her (Damasio, 1999, p. 61). Again, there are many aspects of emotions and feelings which support the idea that they have evolved as survival-enhancing mechanisms rather than something ontologically unique in another dimension of mind. The basic emotions have distinct physical patterns, and areas of the brain can be stimulated electrically to make people feel extreme anger or fear.[106] The fact that the intensity of an emotion can be determined with electricity supports the understanding of it as a causal process. The fact that we have the basic emotions we do, with their clear survival value, also suggests their origin in evolution. Even the fact that feelings show in the face can be explained with evolutionary reasons (Pinker, 1997, pp. 414 – 415). These facts all fit well with an evolutionary account of our mental life.

A survey of feelings becomes much more complex as feelings combine with different thoughts, as then we can speak of many different feelings, although what is happening in the body may be very similar.[107] Damasio himself distinguishes between universal emotions, social emotions, background emotions, moods, drives and motivational states. These distinctions are not so important here, although we shall return to drives and motivations when we look at desire (Damasio, 2010, pp. 22 – 26). Pain and pleasure are important for our topic, however, so where do they fit into this picture? Many feelings have incentive and disincentive functions. Some are negative and can be experienced as punishment; they are meant to make the organism withdraw from something negative. Others

106 Electrical stimulation of the hypothalamus can cause extreme rage, and electrical stimulation of the amygdala can cause both fear and rage (Joseph, 1996, pp. 173 – 174, 182 – 183).

107 Damasio speaks of secondary emotions, which are combinations of cognitive states and basic emotions, and these can evoke numerous feelings with subtle variations, like the differences between euphoria and ecstasy based on happiness, or melancholy or wistfulness based on sadness, or panic and shyness based on fear, and so on (Damasio, 1994, pp. 134, 149 – 150). Feelings like jealousy, envy, *schadenfreude*, etc., may feel quite similar. In an experiment, people were given a drug which puts the body in a certain state. When different groups were in the same room with an actor behaving a certain way, the people interpreted their own feelings in the same way as the actor behaved (Schachter and Singer, 1962).

are positive and can be experienced as rewards; these are meant to make the organism approach something positive. Pain is clearly negative, but it is almost as basic as an automatic response to a stimulus.[108] It has the function of making the organism withdraw from that which creates the pain. Pleasure is a common name for different good feelings, and motivates the organism for doing things that are good for survival and a good life but also for sexual reproduction (Damasio, 2010, pp. 52–53).

So far I have said that feelings arise when the organism senses something in the outside world or its own body, and I shall argue that this is often an important influence when people make choices. An emotional influence of the body which is important for understanding free will is the fact that we can remember earlier emotional experiences we have had. When we are about to make an important choice we can remember earlier relevant events which evoke feelings that influence the choice we are about to make. Feelings are stored in memory and influence choices, and this is another reason to take a closer look at memory.

5.5 Memory

The brain can reproduce events in our mind regardless of our choosing. Memories are stored together with the feelings that we reacted to the situation with, and more emotional memories are better remembered and more easily recalled. This is generally an advantage for humans, since it often makes us recall important earlier events when we are in similar situations and can take advantage of what we learned the last time. But it can be a disadvantage for particular individuals, for example, people with post-traumatic stress disorder who constantly recall horrible events. It is not the event alone that is stored in the memory, but our relation and reaction to the event and our feeling at the time (Damasio, 1999, pp. 130–132).

108 The area which is hurt sends signals to the brain through special nerve cells called C-fibers and A-δ-fibers. The destroyed cells in the area release chemicals and these also send a signal to the brain. The input from these different fibers and nerve cells creates representations in the brain of the body being in pain in a certain area, and as these patterns become conscious we have a conscious experience of being in pain (Damasio, 1999, pp. 71–73). The same C-fibers also mediate itching, but they cannot be used for both itching and pain, so if you are itching in an area which is then hurt the itch will disappear and only the pain will be experienced (Gieler and Walter, 2008, pp. 54–56).

Several different functions of feelings have been mentioned, but feelings have another function as well, which has to do with memory and choice. Damasio has put forward a hypothesis known as the somatic marker hypothesis. The main point is that neural patterns representing events in the world are connected to feelings which give the different neural patterns different levels of importance. This explains what we become conscious of both in sensing and in remembering, for at any one time there are numerous neural patterns that could have become conscious: the ones that are more important because of their connected feeling are selected by the brain, which has a disposition for selecting the ones with the strongest feelings attached to them (Damasio, 2010, pp. 174 – 175).

Furthermore, the somatic marker hypothesis explains how people make efficient choices despite the so-called "frame problem", which means that in any situation there are numerous things that could be considered before one makes a choice; this would delay the choice for a long time. But when feelings are connected to the different images, the brain sorts them according to importance and generally makes the choice much easier. Damasio supports this hypothesis with findings from patients whose brain injuries left all knowledge and logical capabilities intact but whose feelings of emotions were lost. They then also lost their ability to make rational decisions (Damasio, 1999, pp. 40 – 42).[109]

All of this is important for the understanding of deliberation and free will since, when we are deciding what to do in important situations, memories and feelings will often be activated (without us choosing that such should happen) and influence what we think and feel about the different alternatives for action, thereby influencing or changing our initial desires. Which alternatives pop into mind may seem undetermined, but may nevertheless be a regular causal process with no indeterminism in it. Much more will be said about this later.

There are different kinds of memories which must be stored somewhat differently since it is possible to lose all memories of one kind but not those of another. Short-term and long-term memory is a distinction most people know of. It is also usual to distinguish between procedural memory (remembering how to play the guitar or ride a bicycle) and fact memory (remembering facts or events). There are some who cannot consciously remember what a thing is or does, but are still able to use it correctly because of their procedural memory. A more interesting distinction for the topic of choices and free will is that between general facts and episodic memories. One patient, E. D., could remember many facts

109 For example, Damasio gave such a person a choice of two dates for their next meeting, and he wavered between the two dates for almost half an hour before Damasio stopped him and decided the date for him (Damasio, 1994, p. 193).

about Kilimanjaro without remembering having been on top of it (Markowitsch, 2004, p. 52). Damasio mentions a person who looked at pictures and correctly identified them as portraying a wedding, but he did not remember that it was his own wedding (Damasio, 2010, pp. 138 – 140). Most important for the question of free will are fact memories and memories of past events. The reason is that fact memories activate images of alternative actions that can be chosen, which again activate autobiographical memories and feelings connected to each alternative possibility for action (Damasio, 1994, p. 196).

The important thing to note from this presentation of memory is that we have stored facts and memories that can be activated as alternatives for action in the future, and we have stored memories with feelings connected to them which can be activated and influence what we desire most strongly to do. Our memories and our experiences can be the cause of our choices, and some of these memories constitute our autobiographical self, which means that our autobiographical self can sometimes be the cause of our choices. The self is the next topic to be considered.

5.6 The self

Damasio distinguishes between wakefulness, mind and self. The mind is neural patterns that can function non-consciously but which require a self to become conscious (Damasio, 2010, pp. 159 – 166). Damasio has an influential theory about the self. He thinks that the brain can cause a conscious experience by adding a self process to wakefulness and mind. But how did the self develop? I will spend some time on this since it helps us understand who the agent is who has free will. Damasio argues that the self was built in three different stages. I will present these stages in more detail after a quick overview first to make it easier to follow. The first stage is the proto-self, which is a neural pattern representing the whole organism. This proto-self produces a primordial feeling, which is the feeling of my own body existing, but without any further connection to the world. In addition to the proto-self, the brain creates neural patterns representing objects and events in the world, but it also creates neural patterns representing the relationship between the organism and the outside world. From moment to moment there is a series of neural patterns representing how the organism changes in relation to the outside world. This creates changes in the primordial feeling which are consciously felt as an experience of changes in the world. The representations of change create pulses of core consciousness that together constitute the core self, which is the second stage. The core self is this series of consciously experienced changes that arise because of the neural patterns represent-

ing changes in the body in relation to the world outside. Finally, these conscious experiences can be held together in extended consciousness to create the autobiographical self, which is a neural pattern representing the life story of a person, created by memories and continuously reconstructed.

In more detail: Damasio defines the proto-self as "an integrated collection of separate neural patterns that map, moment by moment, the most stable aspects of the organism's physical structure" (Damasio, 2010, p. 190). The proto-self is constituted by three different kinds of neural patterns (which Damasio calls "maps"). The first kind represents the body – not the whole body, but the most stable parts of the body. Damasio argues that that is important, since it can explain the stability of the self process, and this is also the reason for the last part of the definition of the proto-self. It is this representation of the body which gives rise to the primordial feeling, which is the basis for all other feelings. The second kind of neural pattern which constitutes the proto-self is a general representation of how the main parts of the body relate to each other when the body is not moving, and it is then constantly compared with patterns representing how the body is moving right now. The third kind of neural pattern is representations of the sensory portals of the body (eyes, ears, nose, tongue, skin), and these have the function of locating where the body is relative to the sense impressions. Not only do we see and hear, but we also feel that we see with our eyes and hear with our ears. This creates an experience of having a certain perspective and particular location in the world (Damasio, 2010, pp. 190 – 198).

Damasio does not think that the proto-self and primordial feeling are enough to account for the phenomenon of self which we humans experience today. Had it only been a proto-self its experience would have consisted only of a primordial feeling of being a self from moment to moment with no connection to anything. What is needed is a clear experience of being a protagonist connected to events in the world, and Damasio thinks that this happens in the following way: When the organism encounters something, this changes the organism and makes the brain construct a new pattern representing the change. The change in the proto-self leads to a change in the primordial feeling, and this change is experienced as a conscious experience of the object. But it is also experienced as something happening to a protagonist, to a self. The narrative of objects encountering a body and giving rise to different feelings in the body also suggests that there is a protagonist to whom things happen, who feels and acts and has a sense of ownership of the experiences. The self is, so to speak, deduced from the narrative and experienced as such a self (Damasio, 2010, pp. 201– 204). Damasio has also described it by saying that our conscious

mind is like a movie, and when we ask who is watching the film the answer is that the watcher is a part of the movie (Damasio, 2004, p. 11).

I find it useful to distinguish between a minimal self or basic subjectivity and the more complete sense of self which the core self is. By "basic subjectivity" I mean the phenomenon that something is like something *for* something or someone, and I shall argue that this is presumably what Damasio means by his "primordial feeling". Something cannot be *like* something unless it is like something *for* something or someone. What is expressed in this "*for* something or someone" is what I mean by basic subjectivity. Even if the core self can be deduced from changes in primordial feeling, it requires a basic subjective element in the primordial feelings in the first place. What I want to do here is to add some support for the idea that the core self can be deduced from changes in primordial feelings.

A sense of self, by which I mean a sense of being a single conscious subject, may not need to imply a physical or non-physical continuous self which has this sense of self. Pete Mandik has argued that a self can be deduced from experiences in the same way that we can see a picture and deduce that there must have been a camera at such and such an angle and distance from what is seen in the picture. The perspective and distance suggest where the camera must have been to take the picture, but the picture may have been computer-generated and given a certain perspective. Mandik's point is that there are a lot of subjectively conscious experiences going on in the brain which may create a sense or thought that it must have been a self having these experiences, but in fact there is just a series of experiences *including* the experience that there is a self having the experiences. We understand temperature as hot or cold and we experience things as good or bad relative to ourselves. But there could be something in the nature of the experience that makes us assume a continuous self is having the experiences even if there is no such continuous self. Perhaps there is just a series of experiences, pulses of consciousness, which give the illusion of a self owning the experiences – and even the sense of a self as the agent of certain events even if they just happen as a causal chain of micro events (Mandik, 2001).

But must it not be *someone* who deduces that there is a self? No, the suggestion is that there are ontologically subjective experiences that can be configured in a unified way which then constitutes a self-experience. Ontological subjectivity is a feature of the world, as argued by John Searle (Searle, 2007, p. 327), which can create a self. This does not explain what basic subjectivity is, or how basic subjectivity is possible at all, but it is an explanation of how a full-blown self can arise from experiences that are subjective in a basic sense. Translated to the theoretical framework of this book, it means that the qualia field consists of values that have basic ontological subjectivity to them, which again means that a uni-

fied configuration of these can produce a conscious core-self-experience like the one humans have.

How are then all the different qualia produced by different parts of the brain combined into one coherent picture? What "glues" the parts together so that one subject can be the part of many pieces, sometimes referred to as the subject summing problem? Here I think similarly to Sam Coleman, who says that qualia are arranged by being located at different places on a phenomenal screen into a coherent whole. Being this coherent system is to be a conscious subject (Coleman, 2012, pp. 159 – 160). To this I would add that the phenomenal screen is a part of the qualia field with a unified set of actualized qualia values, and that the brain has evolved to create such unified parts. Above, we saw examples of how the brain merges contradictory input into a unified whole, and below I shall describe how and why this evolution took place. It is the unifying work of the brain that explains how the pieces combine to have a subject as their sum.

The third stage in the development of the self is the autobiographical self. The autobiographical self consists of our memories, including memories of our thoughts about the future and who we are. Whereas the core self is always present when a person is conscious, either in focus as self-awareness or in the background since attention is on something else, the autobiographical self is either dormant or active. It is important to understand that every time memories are recalled, they are modified a little, and the feelings they are connected with may change a little. This means that the autobiographical self is reconstructed all the time (Damasio, 2010, pp. 210 – 211).[110]

The autobiographical self can be constructed only because the consciousness is able to hold several elements present over time. Different memories can become conscious to a person and grouped together or seen in the light of each other, thus giving a coherent picture of who that person is. I have already mentioned Damasio's somatic marker hypothesis. It says that every time an image is recalled, it is automatically marked with a certain value in the way that a certain feeling expresses this value it is connected with, and this valuation is constantly revised. The image of who we are – our autobiographical self – is

110 Interestingly, this explains how (parts of) psychoanalysis work(s): non-conscious memories from childhood might be activated in the mind or in dreams, but without becoming conscious to the awake person. If there are bad feelings connected with the memories, these may create more negative feelings or problems like anxiety or depression, while the memory itself is still non-conscious. When the memory is recollected in consciousness, new feelings are connected to it in the same way as all recalled memories. But now one may see the incident in a whole new light, or relate to it in a safe context, and connect new feelings with the memory, which might stop or reduce the negative influence on the person.

also marked by such a feeling and constantly revised. Damasio argues that this marking depends partly on preset dispositions acquired through evolution so that things which are important for survival are considered important. But the marking also depends on values acquired through life – things we have come to see as important for us in the light of our individual experiences and reflections (Damasio, 2010; Damasio, 1994, pp. 177–180). This is important for the topic of free will, because it means that things we have come to see as important through our experiences and thoughts are stored in our autobiographical self and influence the later choices we make.[111]

It is useful at this point to sum up the terminology concerning the self that will be used in the rest of this book. The core self is the stream-like consciousness of what happens here and now together with a feeling of what that is like. It is a series of conscious events – the stream of consciousness consisting of experienced qualia. The core self is a process of conscious experiences, and the physical basis for the core self is patterns representing changes in the interaction between the proto-self and the environment. The autobiographical self is a person's understanding of herself based on memories and their connected feelings. The autobiographical self is a physical neural pattern which can become conscious. This image of oneself also influences how it feels to be that person in any moment of core self-consciousness.

Memories of what has happened – consciously or non-consciously – in the mind (sensing, thinking, feeling, desiring) constantly return to the mind and influence what happens in the mind the next time. A person has experiences and these lead to feelings and thoughts that are stored in the memory. The more memories of thoughts and feelings a person has connected to her experiences, the more these memories will influence what the person desires later, since desires also depend on what we feel about different alternatives (as I argue in the next section on desire). The autobiographical self is a collection of memories of thoughts and feelings which influences the desires and choices of every person. Our choices are not only influenced by the autobiographical self, since we are born with many dispositions and other non-chosen influences (e. g., a non-chosen acquired mental disorder) that can also come into play. But the autobiographical self becomes a larger and larger influence on most people's choices during

111 Let us say that future research in neuroscience rejects Damasio's distinction between the core self and the autobiographical self. There will nevertheless remain something structurally similar, where something corresponds to our conscious experience of here and now, while something corresponds to the important memories that shape how we experience ourselves as persons and what we desire. These parts of a better theory of the self in the future must then replace what I here call the core self and the autobiographical self.

their lives, since a larger and larger collection of thoughts and experiences can be remembered, and then habits and character traits are developed which influence future choices.

I use the terms "person" and "agent" interchangeably for a living human body with a mind and a core self. What happens in the core self is stored as memories and added to the autobiographical self, which in turn influences what happens in the core self. The term "autobiographical self" can be understood in several different ways. Every person has many experiences and these are stored as memories. A neural pattern is constructed in the brain representing the person who has had these experiences, and this is a coherent pattern of important memories. This pattern changes over time and can become conscious. This neural pattern is what I mean by "the autobiographical self", and although it can be seen as a process over time, I use it largely to mean the neural pattern as it is today or at the time of discussion.

The autobiographical self is not all my memories, but a selection, and although it can become conscious, not all details of this neural pattern can be conscious at the same time. For some people there are probably parts of their autobiographical self that never become conscious. This means that they may have a conscious understanding of themselves that does not match their autobiographical self as neural pattern at all points. For example, they may have experienced something disgraceful which they non-consciously deny ever happened as a survival technique. They may think that they handle certain situations well, but in these situations the non-conscious and denied memory may still be non-consciously activated so that they behave strangely in these situations – which they may also fail to realize.

I will end this subchapter by noting some meanings of the terms "self" and "sense of self" on which I do not focus in the rest of the book. I have already mentioned the primordial feeling based on the proto-self, which is the subjective experience of being present, whereas the core self is the experience of being an agent owning the experiences, and this experience is based on representations of changes happening to the proto-self. This experience explains why agents feel that they own their actions even if there is no homunculus inside our body but rather a series of causal processes happening in the mind. The sense of being someone who acts is based on the fact that actions follow intentions.[112]

The term "sense of self" can include many different experiences. There is the sense of *presence*, which is the primordial feeling. There is the sense of being *one*, which is based on our unified conscious experience. There is the sense of

112 Similar distinctions are made in Gallagher (2000).

being *one with your body*, which is based on the core self experience of being a protagonist to whom changes happen, and this is the same as the sense of *ownership* of your own experiences. The sense of *where your body is* and *how it is positioned* is based on a constant comparison between a master map of the body in repose and the position of the limbs relative to this.[113] The sense of where you are located in the world is based on the perspective of your conscious images.

What about the sense of identity over time? Since a person can consciously remember what has happened to this body with this core self, the person will have a sense of persisting identity over time, and identify with that same body and its core self of several years ago. The memories are what give the *sense* of persistent identity, but what *constitutes* personal identity over time, metaphysically speaking? I can only deal very briefly with this here, but I think that at this level of detail, it is possible to describe quite accurately what happens in different difficult cases.

For example, we can ask: if each half of your body was united with another half or each particle in your body was removed and replaced one at a time while you are anesthetized, so that two identical replicas are created, or a teleportation going wrong suddenly caused two replicas, which of them is you? "You" is ambiguous, since it can refer to the person, the autobiographical self or the core self. What happens is that the physical structure of the body and the autobiographical self are made into two versions of the original. These two bodies and autobiographical selves now give rise to one core self each, which in their first waking moment have exactly the same memories, but from the next moment on have distinct experiences, thoughts and feelings and start to change their autobiographical selves into two different autobiographical selves.

One person will then have turned into two, just as it sometimes happens that an embryo splits into two and gives rise to twins. Jack in 2020 becomes Jack 1 and Jack 2 in 2021. Jack 1 and Jack 2 were Jack in 2020, but neither of them is Jack of 2020 anymore. Against this, one can then object that two different persons (Jack 1 and Jack 2 in 2021) cannot be identical to one (Jack in 2020). But that presupposes the classical understanding of identity where all properties are undistinguishable, and while this applies to synchronic identity, it is not given that identity over time or persistence should be defined the same way.

---

113 This sense can be disturbed, and scientists have managed to stimulate an area of the brain which allows them to create out-of-body experiences for the person so stimulated (Blanke, Ortigue, Landis, and Seeck, 2002). There are also many simple experiments that one can perform on oneself to disturb this sense; see for example Ramachandran and Hirstein (1999, pp. 105–107).

Derek Parfit argues that what should matter to people when it comes to such cases is not identity in the sense of all properties being undistinguishable, but rather survival in the sense of psychological connectedness and/or continuity for any causal reason. What that means is that the mental states of a person must be causally related through overlapping series even though there may be gaps in the series (Parfit, 1984, pp. 205 – 207).

Parfit's argument that this is what matters is as follows: Imagine that half your brain had survived an operation in a new body, and you were still conscious and had many memories intact. Then you would and should clearly think that you had survived. If the doctors then came and said that the other half of the brain had also been saved, and now one more body was conscious with many of your memories intact, then you should not think that you have now died, but instead celebrate it as a double success (Parfit, 1984, p. 254). Asking which person is the original is like asking which party is the original if a political party splits up: It is an empty question – we know all there is to know, and there is nothing more to say or know which makes one of them the original (Parfit, 1984, pp. 260 – 261).

I agree with Parfit's reasoning and think that persistence or personal identity over time is just that a person has a quite stable internal structure that only gradually changes over time, without there being an exact border for when something persists. This is the same kind of reasoning that I presented with identity over time for any object in Chapter 2. You can raise a version of the Theseus' ship paradox against this view, since Theseus' ship would remain the same ship if it was gradually rebuilt, but not if everything was suddenly replaced. Likewise, we say that a human being is the same when cells are gradually replaced, but we would not say so if all parts of human body were suddenly replaced except the feet.[114] Somewhere between these extremes there is a diffuse border, and the coarse concept of personal identity over time is useful even if not exact.

When we use the concepts of autobiographical self and core self we can explain everything that happens in an imagined fission-case where Jack turns into Jack 1 and Jack 2 and there is nothing more to know and no reason to insist that one must be the original. Jack's wife may have a pragmatic reason to see it differently, but will have to find a pragmatic solution, like making a copy of herself as well, and the two new couples must share house, money and responsibility for their children.

114 This example and argument is from Rescher (2001, p. 86). I recommend this book by Rescher for an overview of how to think of some typical classical paradoxes.

Returning to the self, most important for the question of free will is the autobiographical self, the neural pattern which represents a person's life right now and is a product of earlier thoughts, feelings and actions influencing future choices. Whereas the core self is a series of conscious experiences, the proto-self and the autobiographical self are physical neural patterns, and the person is the combination of a physical substrate including the physical mind, representations in the brain, and the conscious experience of the core self.

What about the concept of "the self"? (I will not use this term in isolation after this paragraph, but specify just which self I am referring to.) It is normal to think of the self/agent/person as a continuous entity which is the cause of actions. But the core self is not one entity with a continuous existence; rather it is a continuous series of experiences only interrupted by certain forms of sleep and anesthesia or coma. As I shall argue in the chapter on free will, actions are caused by desires triggering motor neurons, and these desires can be caused by the autobiographical self or other causes. It is the physical substrates which give the experience of continuity, the sense of identity with the body and memories, even if both the memories and the body change gradually. Much of what happens in our conscious mind depends on non-conscious physical activity in the brain and body. So those who wish to speak of the self as continuous or a cause of actions should include a physical aspect of the self. I find it better to leave out the concept of "the self" altogether in favor of more precise terms.

As regards these concepts – agent, person, core self, autobiographical self – what does the term "I" refer to? What do I identify with when referring to myself? When I want to be the cause of my choices, what is the cause that is *I?* I certainly identify with a body with a conscious mind which can act now. But I also identify with the same body with a conscious mind at earlier stages which made choices that shaped the autobiographical self I have today. When the term "I" is used in daily life, it can refer to several different things – a body, a core self, an autobiographical self – now and over time, but even if the term is ambiguous in daily speech, it should be precise here. When I, the author of this book, use the terms "I" or "my" about myself, it refers to a particular body with a mind, a core self, and an autobiographical self from its beginning and development until today. In other words, when people say "I", in my definition they refer to that which I define as a person, and the autobiographical self that persons usually develop over time.

There are so many philosophical problems that it is difficult to give a coherent definition of what a person is and what the conditions are for (personal) identity over time, but the most important question is what actually exists in the world, and that is bodies with minds and core selves which usually develop autobiographical selves over time. I suggest that the term "I" should be thought

of as referring to this whole configuration of structures (a body with a mind and core self and usually an autobiographical self), but in daily speech it is often ambiguous what "I" refers to since it refers to different parts of this configuration. Thus, if I say that I am six feet tall, "I" refers to the body which is so tall, but if I say that I am thinking, "I" refers to the conscious brain activity that occurs in the same body. Or I can say that I was 30 years old ten years ago and refer to the body, the memories of which are stored in my autobiographical self now.

So far we have looked at how the memories, emotion and the self can influence what we desire when making choices, but time has now come to take a closer look at desire.

5.7 Desire

The concepts of desire and of desire strength are important for my causal understanding of the mind and free will. I will unpack this later, but start with untangling different distinctions when it comes to desire. Damasio does not have much to say about desires, but he does distinguish between emotions on the one hand and drives and motivations on the other, which he says are simpler constituents of emotion (Damasio, 2010, pp. 109, 111). I interpret Damasio's terms "drives" and "motivations" as having roughly the same meaning as "desires", since he exemplifies them with hunger and thirst, which seem like obvious examples of desires (Damasio, 1999, p. 77; Damasio, 1994, p. 116). But just what are desires? Damasio seems to think of them as dispositions in the brain that are meant to help the organism achieve homeostasis (Damasio, 2010, p. 55). That is a third-person-perspective description of the physical realizer of desires and their function. I shall focus mostly on the first-person-perspective experience of desire and return to the physical side later.

From a first-person perspective, "desire" is a concept that can be defined quite widely to include all sorts of wishes and preferences or even judgments that something is good. On the other hand, it can be defined more narrowly to include a feeling often related to pleasure or displeasure. I may judge that something is good (e. g. that I give money to aid projects in Africa), without having a desire that it should happen that I give money to Africa. Since Damasio uses the terms "drives" and "motivational states", he seems to focus on this *felt* desire which is supposed to lead to action. Hunger is a feeling meant to make the hungry act on their desire and eat, and the same goes for thirst or sexual desire. I thus understand desire as including both a thought that something is good and a feeling that makes the person want the desired state of affairs to happen.

Alfred Mele offers several helpful distinctions when it comes to desires. He defines a desire for A as an A-focused attitude which constitutes motivation for A (Mele, 2003, p. 170). Here one should distinguish between a desire to perform an action (go for a walk) and a desire for a state of affairs to be true (that my team will win the cup) (Mele, 2003, p. 16). Most desires are desires for a state of affairs to come true which include doing something in order for it to come true (I mow the lawn since I desire a nice garden), but not all (I may desire to be someone else). In order to avoid constantly interrupting the text by making reservations about counterexamples, I focus on action desires, which include thoughts about some states of affairs one wants to become true. This way desires are connected to alternatives for action, but the alternative for action is understood as including both the goal and the means, and thus desires are related to both goals and means. I also focus on proximal desires, which include the desire to do something now, although there are also distal desires, which are desires to do something later (Mele, 2003, p. 167).

Mele further distinguishes between occurring and standing desires: I may always desire that there should be peace on earth but that does not mean that that desire is always activated or felt (Mele, 2003, p. 30). Some desires are intrinsic, meaning that one desires something for its own sake (e. g. if one likes to whistle), whereas other desires are instrumental, which is when you desire to do something to achieve something else (go to the dentist in order to have good teeth). Desires can also be a mix of the intrinsic and instrumental; for example, you enjoy swimming for its own sake but also do it in order to get into better shape (Mele, 2003, pp. 33 – 34).[115]

One might think that there are just two desires – a desire for pleasure and a deterrent desire for pain – and that all other desires are sub-desires. But some of the most common desires seem to have physical realizers located at different places in the brain, and it is good if the vague concept of desire can be related to something empirical. There are different areas of the brain that can be stimulated electrically and the people stimulated will report that they feel a certain desire. Animal tests confirm the same. The lateral hypothalamus is active when we are hungry or thirsty and also when we see food and drink. If it is stimulated electrically, we will feel compelled to eat and drink, whereas when it has been destroyed in animals they stop eating and drinking and must be force-fed if they

115 Note the potential confusion that I follow Mele here, who uses the term "intrinsic" to describe a goal in itself, while in the chapter on ethics I describe how Christine Korsgaard says that "intrinsic" should not be conflated with "goal in itself", and so I avoid using "intrinsic" in the sense of "goal in itself" in the ethics chapter. Hopefully, confusion is avoided with this notification.

are not to die. Parts of the amygdala and hypothalamus can be stimulated and people will suddenly feel a strong sexual desire and perform explicit sexual acts even if these are inappropriate in the situation. There are also famous experiments that have discovered what seems to be a pleasure center in the brain. If they are stimulated electrically in humans, the latter will report that they feel great. When animals are allowed to press a button that stimulates them they will continue to do so ceaselessly until they are exhausted (Joseph, 1996, pp. 171–172, 187).[116]

Obviously, humans have many desires that are not innate. People can desire all sorts of things, like desiring their football team to win, and it is impossible that there should be an innate area in the brain responsible for all kinds of concrete desires. Timothy Schroeder, Adina Roskies and Shaun Nichols give an overview over standard neuroscience on desires. They say that standard neuroscience supports that actions are caused by a physical desire in the brain (Schroeder, Roskies, and Nichols, 2010, pp. 84–87). While we have innate desires, they are changed by beliefs and by emotions, which are parts of beliefs (Schroeder et al., 2010, pp. 87–89, 101). In the brain, different desires point in different directions, all wanting the body to move a certain way. Schroeder et al. compare it with a class of screaming children all wanting to take the others with them in a certain direction, and then the teacher picks out one pupil, who gets to decide. In the same way, the brain selects one desire to cause action (Schroeder et al., 2010, pp. 81–83).

This process of acquiring new desires explains how desires relate to reasons for action. Reasons for action are often divided into three:

1) normative reasons, which are objective reasons for acting a certain way,
2) motivating reasons, which are the reasons a particular individual actually had for acting, and
3) explanatory reasons, which are the reasons why an individual actually acted, which may be something different than the normative or motivating reasons (Alvarez, 2018).

It may be confusing to talk about reasons, since the term "reason" is ambiguous. It can mean either an epistemic reason (as in 1 and 2), which is a proposition that makes another proposition more likely to be true, or it can mean cause (as in 3). Furthermore, epistemic reasons can be moral reasons that everyone have for acting certain ways, or individual reasons why something is good for an individual.

116 These examples strongly suggest that there is a physical side to desire strength and that desire strength can cause action – a topic to which I shall return.

Epistemic reasons connect to both goals and means and thus to alternatives for actions as including both goals and means. For example, I can have reasons to desire a goal, and therefore also reason to desire the means to that goal, which means I have reason to desire the alternative for action that includes both the goal and the desire.

I may seem overly focused on desires as causal explanations for actions – how should we understand the fact that people can act for moral reasons? I believe it happens in accordance with the paragraph before the previous one. We start life with our innate desires, but we can acquire moral goals that we adopt as our own goals and that we desire as good. This happens through different processes as described by psychologists: through parental guidance, development of empathy, processes of recognition, experiencing things as good, developing a moral identity, acquiring habits through practice, imagining scenarios about a good and safe world, etc.[117] This all fits well with a process where the autobiographical self can change itself and acquire new desires that can cause actions.

Desires shaped by beliefs about how they are best fulfilled can make desires compete. A person may desire to live well and desire to avoid pain, and believe that if she abstains from sex God will let her live forever, but if she does not there is eternal pain waiting, so these desires and beliefs together make her act on the desires to live and avoid pain and not on the sexual desires. Or a sexual desire can make a person eat less and work out more to impress someone. Thoughts and feelings influence the strength of the different desires. This is important for the question of free will, and before I discuss it further I must discuss what it means that desires have different *strength*.

Desires as consciously experienced seem to have a variable degree of strength. For example, one can be more or less hungry or thirsty or experience stronger or weaker sexual desire. The strength of desire allows a person to have a preference when they have contradictory desires; the strongest desire is preferred the most. But what is the variable in virtue of which the desire is stronger or weaker? Here I suggest that it is the feeling of pleasure and displeasure which is the main variable, although it comes in different shades, in the same way as different feelings can give different kinds of pleasures or displeasures. The variable is in the amount (how much it occupies the conscious mind) and the felt intensity of the pleasure or displeasure. Pleasure and displeasure are felt in different ways by the one who desires something: she may desire something and feel displeasure since she does not have what she desires, or feel

117 For a detailed theory combined with neuroscience, see Narvaez (2013).

pleasure when she thinks about achieving what she desires and feel pleasure when she actually does achieve what she desires. Note that this description was concerned with different degrees of *consciously felt* strength of desire. I shall complicate the picture later by adding a non-conscious and a physical aspect to the strength of desires to explain why we do not always act on the desire that consciously feels the strongest.

So far I have argued that there are competing desires, and that these are influenced by thoughts and feelings. How does such influence happen? We have already seen briefly how Schroeder et al. describe standard neuroscience by saying that we have innate desires that can be changed by thoughts and emotions, but this will now be combined with more details from Damasio and others.

Different desires are triggered by stimuli either from the body or from the world outside. The situation triggers fact memories, both facts about the situation and possibilities in the situation, and autobiographical memories from similar situations and the feelings that the person had in those situations. Autobiographical memories are especially activated if there are strong feelings connected with them. The thoughts about possibilities for action in the situation may also activate autobiographical memories.

Consider this example: A tired man may see an empty chair and desire rest, but he may at the same time see a woman approaching the chair, desire to get to know her, and remember that offering a chair is a way to make positive contact with someone, while swiping the chair in front of them can be considered rude. Maybe he remembers earlier episodes of either offering a chair or swiping a chair, and more generally of getting to know women or being criticized for his behavior. The sight of the woman and the chair can activate thoughts about many possibilities for action and the possibilities that these actions can further lead to, and memories and feelings connected with all of these.

The feelings activated by earlier memories blend in with the feelings that the first desires give rise to. Maybe the man in the example above first had a strong desire for rest and so wanted to sit on the chair, but as he thought about the possibilities of getting to know a woman by offering her the chair and the possibilities of what would happen if he did not offer the chair, chair-offering was connected with good feelings and chair-swiping was connected with bad feelings, and so the felt desire to offer the chair was strengthened and the felt desire to take the chair was weakened.

My wife is a good example of how recalled feelings blend with desires. In the first stage of her pregnancy, the black coffee she usually drank and the pizza I usually cooked made her sick. When her appetite returned, she still did not desire black coffee or my pizza at all, but she did desire different variants of coffee and different variants of pizza. These different kinds of coffee and pizza tasted

quite similar to regular black coffee and the pizza I made, but whereas she would desire some kinds of coffee and pizza very much, she would not at all desire something which tasted almost the same, namely regular black coffee and my pizza. The most likely reason for this is that the memory of that special coffee and that special pizza had a bad feeling of sickness connected with it, which reduced her desire for it strongly.

The process I have here described is in essence the following: a situation triggers alternatives for action that a person desires with different degrees of strength, and activated memories mark the alternatives with feelings which change the strength of the desire for each alternative. In this process, every event was causal, because it was all about neural patterns firing and thereby activating other neural patterns with which they were connected.

Damasio describes this process similarly, although quite briefly. In a given situation different alternatives for action are activated, and Damasio's somatic marker hypothesis holds that the different images are connected with feelings. This somatic marking influences which alternatives become conscious at all, and it influences how long we consider something and what we end up choosing, since feelings influence the strength of desire. Parts of this process are conscious and parts of it are non-conscious, according to Damasio (Damasio, 1994, pp. 174, 184–187, 196).

Peter Carruthers also understands the process of practical reasoning like this, and he refers to Damasio when he describes it: we envision different alternatives and register our emotional response to the different alternatives (Carruthers, 2006, pp. 137–138). Also Chandra Sripada thinks similarly when he describes different ways that what he calls the deep self can influence choices (Sripada, 2016). This is important for the question of free will, since it means that the experiences a person has had in her life (including thoughts and feelings) and stored in her autobiographical memory will influence the future choices she makes. Consequently, her autobiographical self will play a causal role in some of the choices she makes.

What happens between the process where, after some deliberation, one desire becomes the strongest and the point when a person acts on her strongest desire? How does the strongest desire lead to action? Damasio does not have much to say about that process. He says that the somatic markings provide incentives for a person to act and that we have an innate preference system for what we like and do not like and dispositions to act in order to achieve pleasure and avoid pain (Damasio, 1994, pp. 174, 179).

Peter Carruthers has developed in much more detail what happens between when a person considers her desires and when she acts. Since his theory is good and fits well with what Damasio says elsewhere, I add it here. Carruthers argues

that we have different desire modules, which correspond to the different innate desires that I mentioned above (Carruthers, 2006, pp. 113–114). There is also a different module which Carruthers calls the practical reasoning module. This module selects as its input the desire it registers as the strongest, then starts to follow a set of heuristic rules. A basic sketch of these rules is as follows: The input is a desire for P. The practical reasoning module scans autobiographical and fact memories[118] to find a memory of the "If I do Q, then P will occur" sort. If it finds such a memory, it will search through a set of action schemata that are stored in memory and check if such an action scheme is doable here and now.[119] If not, then the module will search for memories of the "If I do R, then Q will occur" type in order to make the action scheme doable. If this process goes on for a while without success, the module will turn to the second strongest desire and start the process over again (Carruthers, 2006, pp. 57–58, 131).[120]

The details may be different in real life, but this is a scenario which describes the reasoning process as a causal process. The description above was serial, but the brain probably processes different desires in parallel and then the practical reasoning module selects according to certain rules; for instance, that it is worth considering a very strong desire for a certain period of time before

118 Carruthers uses the term "belief", but I understand beliefs as fact memories and their implications.

119 The theory that there are motor schemata guiding our actions has been developed by Richard Schmidt, among others (R. A. Schmidt, 1975; R. A. Schmidt, 2003). The theory explains why many of our actions are not caused directly by desires. Rather, we desire some overarching goal and then motor schemata actualize the desire. For example, when we write something, we do not make a decision to press every letter; rather there are motor schemata for the different words. There is constant feedback from the world interacting with our desires to find out whether a program should continue or be substituted with another. This happens on small scales and larger scales. An example of a small-scale motor scheme is the fact that when I want to type the word "Damasio" I always write "Damasion", which may well be because I have stored a small motor program for the ending "-sion" on English words. An example on a larger scale is driving familiar routes. When driving from my house I almost always take the same route into a roundabout and further into a tunnel. Sometimes I need to go another way and have to change lanes before the tunnel, but I have often ended up in the tunnel. That is probably because the normal route is a stored program and then I have to think consciously about changing lanes when it is time to do so; otherwise I end in the tunnel. The existence of such motor programs is my answer to the charge that event-causal theories of the mind cannot explain how we do many actions that are not directly caused by desires (Steward, 2012, pp. 164–165).

120 Whereas Carruthers uses this module to describe the whole deliberation process, I employ it to describe the final part of the process: from something becoming the strongest desire to the desire being executed in action. I write more about the physical side of this in the chapter on weakness of will (especially how the basal ganglia get a function from being formed by habits).

choosing to act on a less strong desire instead (Carruthers, 2006, pp. 131–132). The main point is that it is a mechanism which decides the strength of desire and follows a set of heuristic rules to turn the desire into action by selecting an action plan which is sent to the motor neurons for execution (Carruthers, 2006, pp. 131, 138).

Although Carruthers thinks that much of our practical reasoning happens like this, he follows Daniel Kahneman in distinguishing between two systems of reasoning, known as system 1 and system 2. System 1 works fast and quite reliably according to the general rules described above. System 2 is a conscious and slow-working reasoning process, but it is more reliable than system 1 and can trump system 1 decisions (Carruthers, 2006, p. 254; Kahneman, 2011, pp. 20–22). System 2's reasoning processes are typically concerned with important and difficult choices, and since I argued that consciousness plays a causal role this allows for consciousness to have a causal role in important choices.

I have so far argued that the strongest desire leads to action via the workings of a practical reasoning module that translates desires into actions. But there are two objections against this that should be considered, and both have been raised by Alfred Mele. He argues that intentions play an important role in reasoning which cannot be reduced to a combination of desires and beliefs. It is not good enough to say that the strongest desire leads to action, for there are many examples where the strongest desire is not the same as a person's intention. For example, the strongest desire of both Alan and Bob may be to insult Carl at a party, and they do, but still it may be correct that only Bob had the intention of insulting Carl at the party. What the concept of intention adds to the reasoning process, according to Mele, is that intentions settle which desire a person wants to act upon. A person has thoughts and can assess different desires in order to identify the best one and intend to act accordingly. What a person judges to be best can be different from his strongest desire. So in Mele's understanding of the deliberation process there are desires and one of the desires is the strongest, but it is a process of assessment producing an intention which settles which desire will lead to action, and that need not be the strongest desire (Mele, 1992, chapters 8 and 9).

Carruthers also discusses how we should understand the relation between what we judge to be best and what we desire the most. How can the judgment of something as best lead to action? Carruthers answers that we probably have an innate desire to do that which we judge to be best. How then is weakness of the will possible? Weakness of the will is not to do that which you judge to be best. It is easy for Carruthers to answer: even if we have a desire to do that which we judge best, we also have other desires, and sometimes these are stronger. So, even if Bob judged it best not to insult Carl at the party, he did it never-

theless, since his desire to insult Carl was stronger than his desire to do what he judged best (to not insult Carl) (Carruthers, 2006, pp. 391–392). I shall say much more about weakness of the will in the chapter on free will.

What about intentions? How can a person intend something which is not her strongest desire? Both Carruthers and Damasio think that desires can be at work non-consciously, and it seems that Mele does so as well (Carruthers, 2006, p. 400; Damasio, 1994, p. 185; Mele, 2003, pp. 30, 164). The answer is then close at hand: even if a desire is not felt consciously as the strongest, it is still the strongest in virtue of the non-conscious feelings connected with it, and selected by the practical reasoning module as the strongest. The module selects the strongest desire, and this is what settles the reasoning process. One could use the term "intention", but as long as there are thoughts, desires and a disposition in the brain for transforming the strongest desires into action there is no need for intention as referring to a separate entity in the brain.[121] I shall explain later in more detail what desire strength amounts to.

Using non-conscious desires in this way may seem like a non-falsifiable theory, but I shall defend it later when discussing weakness of the will, and answer more critiques from Alfred Mele. As mentioned above, however, both Damasio and Mele also believe that non-conscious desires influence our choices. Carruthers defends this belief by arguing that it works that way in animals, and that an account of the human mind must be evolutionary in the sense that it can show how the mind has been gradually built from animals to humans (Carruthers, 2006, p. 403). Much more could be said about desire, of course, but here I have said what I need to say about the self and the mind in order to lay the foundation for my solution to the problem of free will, and how to understand thinking and consciousness.

121 Mele has many suggestions echoing Carruthers, since Mele thinks that there are mechanisms in the brain which by default select what is assessed as best as the intention of the person and activate motor schemata to execute the intention (Mele, 1992, pp. 167, 221, 130–137).