8 Free Will

8.1 Introduction to the problem of free will – main positions and problems

To show how it deals with relevant problems, a new theory of free will should relate to typical challenges that other theories of free will face. I will now present the most common theories and the problems they face, and use this presentation as a structure for presenting my own theory of free will. There are three main questions that a theory of free will should answer: What does "free will" mean? Do we have free will? Is the proposed theory coherent, given that free will seems incompatible with both determinism and indeterminism?

What does the term "free will" refer to? There are many different definitions of free will, and one should ask which understanding of free will is being considered, whether it is being affirmed or rejected. It may well be that a strong form of free will is rejected while a weaker form is affirmed. In ordinary language, a minimum requirement for what it means to have free will is to say that it is the freedom persons must have in order for it to be meaningful to hold them responsible for their actions. (Like free will, though, responsibility is understood in many different ways.)

A common definition of free will is to say that

1) it is "up to us" what we choose between several alternatives, and
2) the source of the choice is *in* us, not outside of us or in something else that we cannot control (Kane, 2011, p. 5).

Even if this description does not apply to every choice, most people will say that they experience such free will at least in some of their choices. Still, both parts of even this definition are contested and can be further defined in various ways: What does it mean to be the source of a choice? Are alternative possibilities necessary for free will? How should such alternative possibilities be understood? Some philosophers say you have no free will, others say you have free will in a weak sense of the term, and still others say you have free will in a strong sense of the term. A way of formulating the problem of free will is thus to ask how strong a degree of free will humans have.

Among those who affirm free will, compatibilists and libertarians disagree on whether free will is compatible with the idea that all events are determined. Determinism is here defined as physical determinism, which is the view that previous physical causes plus the laws of nature determine one future with physical necessity. At any point of time, the rest of the content of history is then implied

https://doi.org/10.1515/9783110770957-010

by the state of the world at that time, which means that there is only one physically possible content of the future. Such physical determinism will be a focus in this book, since I believe that that is the strongest and most common challenge to the question of free will. This is a metaphysical position, and this is what I mean by "determinism" in this book.[164]

In the discussion on determinism, compatibilists believe that determinism is compatible with having free will. For a long time, the most common critique of compatibilism was the consequence argument. Roughly, this argument says that if determinism is true, then what happens in the future is determined by laws of nature and events that took place previously, even in the distant past. Even before any humans existed it was determined what the content of the future would be for all humans. Since the future was thus determined before our birth, it cannot be up to us what happens among different alternatives, and thus we cannot have free will (Van Inwagen, 1983).

Despite this widely debated argument, compatibilists believe determinism is compatible with having free will. Compatibilists today will usually say that it does not matter that only one specific future is physically possible. Rather, they will focus instead on what the inner mental life of an agent must be like in order for the agent to be free. One strand of contemporary compatibilism is the so-called *mesh theories*, which hold that a person is free when she has the right connections or "mesh" between internal parts of her mental life.[165] Another strand is *reasons-responsive theories*. According to these theories a person is free when her actions are based on a rational response to reasons for action.[166] These different compatibilist understandings of free will do not require indeterminism, so free will is argued to be compatible with determinism and in no need of alternative possibilities.

However, an argument other than the consequence argument has been in focus lately against compatibilism, and that is Derk Pereboom's four-case manipulation argument. This argument presents four cases, from a clear manipulation case to a deterministic world, where the point is to show that there are no rele-

164 Note that determinism does not necessarily mean that we will ever be able to predict the future.

165 For example, there are hierarchical mesh theories, such as Harry Frankfurt's theory. Frankfurt argues that we have several desires whose object is an action or a state, which he calls first-order desires. But we also have desires whose object is a first-order desire, and these he calls second-order desires. The second-order desires are internal responses to the first-order desires, which one may like or dislike. According to Frankfurt, we are free when our second-order desires approve our first-order desires, because only then do we have the will we want (Frankfurt, 1971).

166 See for example Wolf (1990), or Fischer and Ravizza (1998).

vant differences between the cases. Since the first case clearly seems to be a case of no free will, the charge is to explain the relevant difference between case 1 and case 4.[167]

167 Here are the four cases, quoted from Pereboom (2014, pp. 76–79):

Case 1: A team of neuroscientists has the ability to manipulate Plum's neural states at any time by radio-like technology. In this particular case, they do so by pressing a button just before he begins to reason about his situation, which they know will produce in him a neural state that realizes a strongly egoistic reasoning process, which the neuroscientists know will deterministically result in his decision to kill White. Plum would not have killed White had the neuroscientists not intervened, since his reasoning would then not have been sufficiently egoistic to produce this decision. But at the same time, Plum's effective first-order desire to kill White conforms to his second-order desires. In addition, his process of deliberation from which the decision results is reasons-responsive; in particular, this type of process would have resulted in Plum's refraining from deciding to kill White in certain situations in which his reasons were different. His reasoning is consistent with his character because it is frequently egoistic and sometimes strongly so. Still, it is not in general exclusively egoistic, because he sometimes successfully regulates his behavior by moral reasons, especially when the egoistic reasons are relatively weak. Plum is also not constrained to act as he does, for he does not act because of an irresistible desire – the neuroscientists do not induce a desire of this sort.

Case 2: Plum is just like an ordinary human being, except that a team of neuroscientists programmed him at the beginning of his life so that his reasoning is often but not always egoistic (as in Case 1), and at times strongly so, with the intended consequence that in his current circumstances he is causally determined to engage in the egoistic reasons-responsive process of deliberation and to have the set of first- and second-order desires that result in his decision to kill White. Plum has the general ability to regulate his actions by moral reasons, but in his circumstances, due to the strongly egoistic nature of his deliberative reasoning, he is causally determined to make the decision to kill. Yet he does not decide as he does because of an irresistible desire. The neural realization of his reasoning process and of his decision is exactly the same as it is in Case 1 (although their causal histories are different).

Case 3: Plum is an ordinary human being except that the training practices of his community causally determined the nature of his deliberative reasoning processes so that they are frequently but not exclusively rationally egoistic (the resulting nature of his deliberative reasoning processes are exactly as they are in Cases 1 and 2). This training was completed before he developed the ability to prevent or alter these practices. Due to the aspect of his character produced by this training, in his present circumstances he is causally determined to engage in the strongly egoistic reasons-responsive process of deliberation that issue in his decision to kill White. While Plum does have the general ability to regulate his behavior with moral reasoning, in virtue of this aspect of his character and his circumstances he is causally determined to make his immoral decision, although he does not decide as he does due to an irresistible desire. The neural realization of his deliberative reasoning process and of the decision is just as it is in Cases 1 and 2.

Case 4: Everything that happens in our universe is causally determined by virtue of its past states together with the laws of nature. Plum is an ordinary human being, raised in normal circumstances, and again his reasoning processes are frequently but not exclusively egoistic, and sometimes strongly so (as in Cases 1–3). His decision to kill White issues from his strongly ego-

Alfred Mele has a similar argument called the zygote argument: Imagine a goddess creating a zygote at exactly the right time and place with the exact right structure in a deterministic universe. She does this because she knows that the zygote will then become a man (Ernie) who at an exact point of time will do something the goddess wants done – for example, kill his grandmother. Ernie will be, like any other person in a deterministic universe, considered by compatibilists to be free, but many will have the intuition that he was not responsible for killing his grandmother since the goddess had planned things so that this had to happen. Yet, since he is like any other person in our world if the world is determined, it seems that if he is not responsible, no one else is either (Mele, 2006, pp. 188 – 189).

Even if one disagrees over how strong the manipulation argument and the zygote argument is against compatibilism, I think there can be little doubt that, if the future is already determined before we are born, we do not have a strong form of free will. It is not up to us to change the future into anything other than what was already determined before we were born. Libertarians, on the other hand, think that we do have a stronger form of free will than this. They hold that we can be the source of our choices in a more fundamental sense than what compatibilists will allow, but that requires an indeterministic world where different futures are possible and where it is up to us to influence what the future will be like.

There are three main positions among libertarians distinguished by how they understand the causality involved in free choices. *Non-causalists* believe that free actions are not caused at all, but are intelligible in the light of the purpose of the action. *Agent causalists* believe that there is a unique and irreducible kind of causation that only free agents can employ. *Event causalists* deny that actions have special causes, but believe instead that all causes are of the same kind: they think that events cause events, both in the mind and in the world in general.

Those who defend the strongest form of free will are the non-causalists and the agent causalists. Non-causalists argue that human action should be explained by intentions or reasons instead of causes, and that these are not reducible to ordinary event causes.[168] A classic charge against this view was leveled by

istic but reasons-responsive process of deliberation, and he has specified first- and second-order desire. The neural realization of Plum's reasoning process and decision is exactly as it is in Cases 1– 3; he has the general ability to grasp, apply and regulate his actions with moral reasoning, and it is not because of an irresistible desire that he decides to kill.

168 An example of such a theory can be found in Ginet (1990).

Donald Davidson (Davidson, 1963). He pointed out that even if a person has a reason for doing something, that does not mean that his reason is what actually caused the event to happen. People often experience having competing reasons for doing different things when they act. The challenge to non-causalists is to explain what links the personal reason to the action. Agent causalists like Timothy O'Connor hold that agents are enduring, irreducible substances that have a unique ability to perform actions (O'Connor, 2011). Agent causalists are typically criticized for appealing to both a mysterious agent and a mysterious form of causation, which does not fit into the ordinary scientific world view (Pereboom, 2014, pp. 65 – 69). Nor do they explain how reasons make actions happen, for in virtue of what does the agent control her actions? Agent causation can be argued to be an irreducible phenomenon,[169] but I shall argue later that it is superfluous to add anything to normal event causation. This will be my main argument against non-causal and agent-causal libertarian theories: that our behavior can be explained sufficiently in event-causal terms so that there is no good reason to believe in extra agency or causation beyond that.

In my view, the most plausible of the libertarian theories are the event-causal theories. Event causalists hold that mental events can cause free actions. There are two main event-causal theories, distinguished by where in the deliberation process they locate indeterminism.[170] The advocates of *centered* event-causal theories believe that there is indeterminism until and in the moment of choice, whereas advocates of the *deliberative* event-causal theories hold that there is indeterminism early in the deliberation process, creating different ideas in the mind (alternative possibilities), but that the rest of the deliberation process is determined.[171]

Since event-causal libertarian theories are close to my own theory, I shall spend a little time in presenting them here. I start with the centered event-causal libertarian theory of Robert Kane. Kane thinks that free will is fundamentally about the ultimate source of action being in us (Kane, 2007, pp. 13 – 14). More precisely, the requirement is that "*To be ultimately responsible for an action, an agent must be responsible for anything that is a sufficient cause or motive for*

169 For example, E. J. Lowe argues that both causality and agent-causality are irreducible concepts. See Lowe (2002, chapters 8 to 11).

170 Indeterminism simply means that more than one future is possible, so when I and others I refer to speak about the location of indeterminism, the point is to speak about the source of indeterminism: where do the indeterministic effects arise?

171 The distinction between centered and deliberative event-causal theories is from Clarke (2003, pp. 57, 71).

the action's occurring" (Kane, 2007, p. 14).[172] This means that the requirement of alternative possibilities is not necessary for all our actions. But it is necessary in some specific early choices in which we formed our own characters, according to Kane. He calls such actions self-forming actions, or SFAs. Even if one could not have done otherwise in some situations, one is still responsible if the reason that one could not do otherwise was earlier SFAs. For example, if you have formed your character through SFAs so that it is now impossible for you to lie, you are still free, responsible and praiseworthy for not lying in situations where you could.[173] As long as we are free to make some SFAs, we can be free, but if the world is determined, then none of our actions are SFAs and then we are not free.

But this seems to result in an infinite regress, for would not those earlier choices depend on even earlier choices, and so on indefinitely (Kane, 2007, pp. 19–20)? Kane's response to this criticism is that the regress is ended if there is an action in the agent's past that lacked sufficient motive.[174] There could be a situation in which the agent did not know what to do and so did not set her will before the action occurred; then the action would set the will in the very act of choosing. Kane calls such actions "will-setting actions" (which are the same as self-forming actions). He adds that in order for such actions to provide us with free will, they must have been such that the agent could act voluntarily, intentionally and rationally in more than one way when she acted. If the action happened as an accident, it would not have made the agent the ultimate source; rather, the accident would be the source. But if the agent had a motive for both alternatives, then she is the ultimate source of the choice no matter what she chooses, so the regress stops there (Kane, 2007, p. 20).

Even if such a choice is undetermined, Kane still thinks it can be a rationally willed choice. To argue this, he offers the example of a businesswoman on her way to an important meeting who witnesses an assault. In this situation she has reasons to stop and reasons to move on, and she does not know what to do. The conflicting motives stir up a chaos in the brain, which is sensitive to undetermined events at the micro level of quantum mechanics. In this situation the woman must make an effort to choose and, no matter what she chooses, it will be for a reason. When she decides, that decision sets her will (Kane, 2007, pp. 26–28).

---

172 Emphasis in original text.

173 "Could" must here be understood in the sense that it was type physically possible.

174 According to Kane, we have a "sufficient motive" for doing something when our will is set one way on doing so before and when we act (Kane, 2007, p. 19).

The most common critique of Kane's theory is that it runs into a problem of luck. Let us say there is a 70% probability that Jack will decide to have pancakes for breakfast. If history were rolled back a hundred times and played again up to the moment of choice, Jack would decide to have pancakes 70 times and something else 30 times. But if the exact same history up to the moment of choice can give completely different choices – which Robert Kane argues it can (Kane, 2007, p. 23) – it seems to be a matter of luck as to what Jack decides to do. The same point can be made with identical worlds up to the moment of choice, where Jack 1 and Jack 2 make different choices.

Kane's theory is a *centered* event-causal theory since it locates indeterminism in the moment of choice. Deliberative event-causal theories try to reduce the luck component by locating indeterminism at an earlier point in the deliberation process. Such models are also called two-stage models since the deliberation process comprises two stages: First there is an indeterministic stage in which alternatives for actions are generated in the mind, then this is followed by a deterministic stage in which one alternative for action is selected.

One of the best such proposals is Alfred Mele's *daring soft libertarianism*. Mele argues that whereas the other models shun luck and only include indeterminism to avoid determinism, this model embraces luck while still maintaining that the agent can be in control (Mele, 2006, p. 117). The point is that the deliberation process is indeterministic, so it is partly a matter of luck what the agents end up choosing. But the agent learns from experiences over time and the agent's evaluations of these experiences influences how likely it is that the same choice will be made later. In this way, the agent shapes her own character over time. This also explains why we hold children less responsible than adults for what they do (Mele, 2006, pp. 122–123, 131–132).

Neil Levy argues that libertarians can try to reduce the luck component by making their theories almost compatibilist (Levy, 2011, p. 77). But he does not think that Mele's strategy of including luck in the history of an agent works, since luck has been a part of every choice and you cannot solve the problem of luck with adding more luck (Levy, 2011, p. 89). Even compatibilists have a luck problem, according to Levy, since it is also a matter of luck what such agents come to think about or desire (Levy, 2011, p. 90).

In addition to the luck problem, there are two other important arguments against event-causal theories. First there is the problem of the disappearing agent. It seems that in event-causal approaches, choices reduce to desires, be-

liefs and bodily movement, and the agent disappears.[175] If everything is just natural causal processes occurring, where is the free agent? A second and similar charge against event-causal libertarianism is the regress problem, since it seems that we can follow causes backwards further and further to before the agent can make an ultimate choice (Strawson, 1994, pp. 5–7). Then the agent cannot be the ultimate cause of a choice if there must be a cause for why the agent chose as she did.

In addition to this list of philosophical problems come the challenges from neuroscience. Three kinds of findings are particularly relevant. The first finding is that of Libet-style experiments showing that consciousness seems to enter the stage after the brain has already determined what a person will do (Libet, Freeman, and Sutherland, 1999). More advanced experiments let researchers predict (better than chance) how people will act several seconds before they make their choice based on watching brain scans of the test persons (Haynes et al., 2007; Soon, Brass, Heinze, and Haynes, 2008). The second finding is of confabulation-type experiments and related kinds of experiments showing that our own conscious interpretations of our actions are often wrong. Confabulation means that persons are wrong about the real reason for their action. This has been demonstrated clearly in split-brain patients (Gazzaniga, 2005), but also among people in general, typically in examples of choice blindness (Johansson, Hall, Sikström, and Olsson, 2005; Hall et al., 2013). Other kinds of experiments show that non-conscious factors often influence our behavior without us being aware of it (Schnall, Haidt, Clore, and Jordan, 2008). The third finding is that reductionist theories of mind seem to explain all parts of human choices and actions (Damasio, 2010; W. Singer, 2004b). Brain processes are physical processes guided by the laws of nature, and there is no need for concepts like persons with intentions choosing between alternatives and controlling the outcome.

As we have seen, the different kinds of compatibilisms and libertarianisms run into different problems. This has led various philosophers to conclude that we do not have free will, since free will is incompatible with both determinism and indeterminism (Pereboom, 2014). But most philosophers still hold on to the idea that we have free will, since this seems best to fit our experience of having free will and being responsible for our actions. Now that we have an overview of the main positions and main problems, I am ready to locate my own theory as a newcomer to this map and indicate how I will relate to the different problems.

175 This is mentioned as a main objection against event-causal theories in, for example Pereboom (2014, pp. 31–33) and Steward (2012, p. 62).

As a brief preview, I will suggest that free will and responsibility come in degrees. A person can be involved in her choices to varying degrees from when a desire immediately causes an action to where an independent self causes an action. I use the theory of the self presented in Chapter 5 to argue that the self can gradually cause its own content over time and thus be the cause of itself and the cause of a person's actions; this way the person is free in the sense of being the ultimate source of her choices. This presupposes a specific understanding of causation as contrastive (presented in Chapter 4) and that the world is indetermined at the macro level of human interaction (to be defended below).

In more detail, the chapter is structured as follows: I start with the topic of determinism and indeterminism in Section 8.2 to show how I deal with the manipulation argument and the zygote argument. I argue that the world is indetermined at the macro level of human interaction and that this is required for free will in a strong sense of the term. I then move on to the topic of causation in Section 8.3 to show how I deal with the regress problem. After this, I move on to the topic of person, self and mind in Section 8.4 to show how I deal with the problem of the disappearing agent. I argue that people are involved in their choices to varying degrees, and that agents do not disappear even if they get explained in a more fine-grained way. I continue to show how free will can be built gradually via how, over time, the self can be the cause of itself and of actions. This is a more detailed response to the regress problem.

After a short discussion of responsibility in Section 8.5, I show how this theory responds to the problem of luck (Section 8.6), the question of weakness of the will (Section 8.7), and some other objections (Section 8.8). The problem of free will is related to many big questions and so cannot be fully defended in one chapter. At several points I must refer the reader either to other chapters or to further details in a book I have written about free will (Søvik, 2016).

8.2 Determinism and indeterminism

We saw in the introduction that free will can be understood in many ways and that some defend a strong version of free will while others, like the compatibilists, defend a very limited version of free will. According to compatibilists you have free will even if every action you do was determined before you were born. If we are to have free will in a stronger sense than compatibilist free will, it requires that there is indeterminism in the world at the macro level where humans act. Indeterminism here means that there are several possibilities open when it comes to what the content of the future will be, whereas determinism means that the content of the future is already set. However, as I will argue

below, external indeterminism suffices for free will. That means that there do not have to be indeterministic processes at specific places in the brain, such as envisioned by Robert Kane in his theory of free will (Kane, 1996, 2007), but only some indeterministic processes occurring somewhere in the world with effect at the macro level of human interaction. Everything in the brain may happen as if the whole world was determined, yet it is important that the world itself is not determined.

The reason such indeterminism is important is that the content of the future will be open so persons can be the causes of which future content becomes actualized. It may seem strange to make an ontological divide between the brain and the rest of the world, but that is not what I do. Maybe there are some indeterministic processes in the brain as well. The point is merely that some indeterminism somewhere is required in order for people to have free will, since that opens up the possibility of different futures and an opportunity for us to influence which future will be actualized. Alfred Mele has claimed that it would be preposterous to base a theory of free will on external indeterminism (Mele, 1995, pp. 195–196), but I shall try anyway, and will respond to his detailed criticism of such attempts later in this chapter.

Do we have reason to believe that there is indeterminism at the macro level of human interaction? The most common place to go for support is quantum mechanics. Quantum mechanics can be given indeterministic interpretations (like Copenhagen and GRW) and deterministic interpretations (like de Broglie-Bohm and Everett). Nevertheless, all interpretations will agree that the guiding laws are merely probabilistic, saying only that something will occur with a certain probability (Ney and Albert, 2013, p. x). This still leaves open whether there is a determinism at a deeper level and whether indeterminism at the micro level of elementary particles can be scaled up to the macro level of human interaction.

If there is indeterminism at the micro level of quantum mechanics, there may nevertheless be determinism at the macro level, because the events at the micro level cancel out at the macro level. It may be undetermined whether a single particle goes here or there, but determined that 50% will go here and 50% will go there, so that the macro result is the same in any case. If it is undetermined where the particle will go, we could set up contrasts and ask, "Why did the particle go here as opposed to there?", and the answer will be that there is no cause, but rather it is a causeless, indetermined event. However, it could also be that indetermined micro events can scale up to the macro level.

James Ladyman offers the example of a scientist who decides to take lunch after so many clicks on his geigerteller (Ladyman et al., 2007, p. 264). Geigertellers measure events that according to some interpretations are indeterministic. We could expand the example and say that a scientist may decide to invite

her male colleague to lunch if she gets a click on her geigerteller before 12. This decision may make them have lunch, fall in love and get married – or not. The world may then be very different in the future depending on undetermined events. We do not know whether quantum mechanics should be interpreted deterministically or non-deterministically. But note also that also in Newtonian physics indeterminism at a macro level can occur, for example if several identical particles with the same speed collide (Earman, 1986, pp. 30 – 32).[176]

Here is an additional argument I suggest in favor of indeterminism: evolution makes more sense if what the content of the future will be is genuinely open than if it is determined at the micro level. If many possible scenarios could have happened, we easily understand why the one that actually happened was where the ones best fit for survival had many children. If the one scenario that actually happened was determined solely by laws interacting with particles at the micro level, it seems that this scenario could just as well have been one where what happens at the macro level is very chaotic and unsystematic. The selection effect makes more sense as a selection between genuinely possible futures than if only one future was possible anyhow. And when there is physical indeterminism, qualia can also play a causal role as seen in the chapter on consciousness.

Taking these arguments together, I find more support for the view that the world is indetermined, so I will presuppose here both that quantum mechanics is indeterministic and that the world is undetermined at the macro level of human interaction. If the world is nevertheless determined and I am wrong in making the presupposition that it is not, I blame the Big Bang for making me screw this up.

While I think that the manipulation argument and the zygote argument are good arguments, they are not arguments against the theory I present here, since that theory presupposes indeterminism at the macro level of human interaction. The zygote argument cannot be used against my theory since a divine goddess cannot plan the life of a zygote in an indeterministic world. The argument thus rather supports the claim that indeterminism in required for free will. As for the manipulation argument, compatibilists are challenged to show the relevant difference between a determinism case and a manipulation case. I will argue below that a person is the ultimate cause of the action in cases of free action while, in manipulation cases, the manipulator is the ultimate cause of the action.

176 Earman also gives other examples from Newtonian and relativity physics. Important examples are briefly summarized in Sklar (1992, p. 203).

8.3 Causation and choices

I now move to the criticism of non-causal and agent-causal theories. The charges against them were that they invoked mysterious agents and forms of causation and were not able to explicate how choices, actions and control occur. Such charges are not valid against the theory I propose here, since I defend an event-causal understanding of the mind and the self, where causation is understood as being of the same kind everywhere in the world. Given such an understanding of the self, how can we understand how choices are made and lead to action?

A person can be involved in her choices to varying degrees, which makes persons have different degrees of free will and responsibility. This gradual understanding of free will solves many problems, as I will show below. I start with an overview of how persons can be involved in their choices to varying degrees, from an action caused by a desire, to an action caused by the autobiographical self, to an action caused by an independent autobiographical self. I presuppose a desire-action model where the strongest desires lead to action, as argued in the chapter on the mind.

At the first level, we find actions caused by desires, and the desires can be conscious or non-conscious, innate or acquired. Sometimes desires lead directly to action without the occurrence of any additional thoughts or feelings between the desire and the action happening. For example, a woman might see someone being mean to her boyfriend, desire to hit that person, and then immediately do so. In our brain, we carry several evolutionary old and simple systems that can sometimes execute an action immediately, bypassing the influence from reason (Schroeder et al., 2010, p. 103).

A new level of personal involvement is reached when the autobiographical self is activated between desire occurring and the action happening. Sometimes the autobiographical self is activated and changes the initial desire. For example, a person may see a chair and desire to sit, but also see a man approaching the same chair. Autobiographical memories are activated about not having offered a seat to someone previously and receiving negative feedback and of another time having offered a seat with positive consequences; then the initial desire to sit weakens and the desire to offer the seat to the other person strengthens. Initial desires may also change because of new thoughts and feelings as the brain is capable of making new connections between neurons.

The autobiographical self can be more or less involved between an initial desire and action, depending on the number and emotional strength of memories considered before action. The autobiographical self may also be more or less independent. An autobiographical self becomes increasingly independent

throughout life as it changes initial desires by a process of thinking and feeling about alternatives. Even if the autobiographical self does not decide what to think or feel, such a change in initial desires is nevertheless caused from within the mind.

There were two charges against event-causal theories mentioned in the introduction: the disappearing agent and the regress problem. I will say more about the development of the autobiographical self and the regress problem below, but first a brief comment on the problem of the disappearing agent.

In Helen Steward's book, *A Metaphysics for Freedom*, her main objection against event-causal libertarianism is that choices reduce to desires, beliefs and bodily movement, and the agent disappears (Steward, 2012, p. 62). But the agent does not disappear. Rather, what an agent is and what it is for an agent to make a choice is explained in a more finely grained way. It is like explaining hardness by describing tension between atoms. The hardness does not disappear, it is just that what hardness is – and, as a result, what is hard – is explained in a more finely grained way.

Concerning the regress problem, this has been formulated by several critics. For example, Hilary Bok argues that since every event has a cause, it seems that we can always follow the causal chain back to events that took place before the agent made a choice (Bok, 1998, pp. 201–205). I will respond to this critique now by first saying more about how the autobiographical self develops.

8.4 Developing an independent autobiographical self

Even if nothing can be the cause of itself, an autobiographical self can over time be the cause of its own content. When we start life, we are not free. At the beginning of life children follow their initial desires, and they are told by their caregivers who they are and what is right or wrong or good or bad. But most children are born with a capacity for reasoning and feeling and thinking new thoughts and making new connections in their minds, which gives them a general ability to find out what is true and good and right. When they make choices and get experiences, these are added to their autobiographical self.

Later, they experience new processes in which new thoughts and feelings change the initial desires and cause action. Again, the autobiographical self is changed from within the mind of the person. In later choices, the old choices and experiences from the autobiographical self can change the initial desires, and again the experience is stored in the autobiographical self. So sometimes a mind-internal process happening independent of the autobiographical self can change that self, for example, if a new experience gives rise to a new thought

or a new feeling. But most often, when the autobiographical self changes the process will involve the autobiographical self, so that the autobiographical self causes choices that cause changes to itself.

This means that as long as the world is not determined (including not determined at the macro level), it will often be right to select the autobiographical self as the cause of a change in initial desires *and as the cause of new changes in the autobiographical self.* This is illustrated in Figure 3.

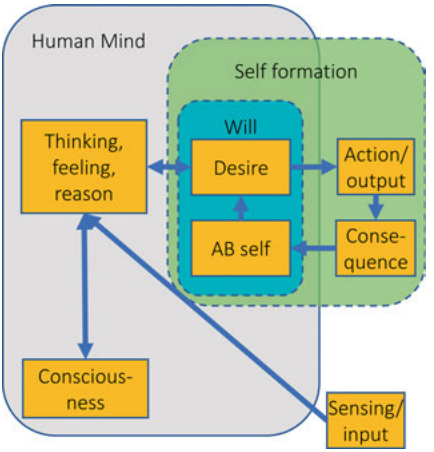

Fig. 3: Self-formation

It may seem of no help to argue that a self can cause itself, since that again must have previous causes and you get a regress going back to before the agent was born. But consider this example: Let us say that a person faces a storm that was undetermined, and sees a young child in the sea about to drown. The person imagines two alternatives: to help the child or not. The process goes on in the person's brain as if determined and leads to action.

The reason it is important that the world in general is not determined is that people will then continually face situations that are not determined by the past. When they face these situations, processes go on in their mind that lead to one desire becoming the strongest, which then leads to action. But even if the process in the mind happens *as if* the whole world were determined, that is not important. What is important is that if the world *is not* in fact determined, then it is sometimes right to select the autobiographical self as the cause of the strongest desire, which again is the cause of which alternative for action is chosen. Indeterminism in the brain is not required, but indeterminism in the world is required for it to be right to select the autobiographical self as the ultimate cause.

Why is it right to select the autobiographical self as the ultimate cause of the choice that is made if the world is not determined? In the example just given, the storm was undetermined, meaning it was not determined before the person was born whether she should save the drowning child or not. Since the storm was not determined to happen, it is right to select the autobiographical self as the cause if the following happened: The person saw the child and imagined two alternatives: to jump into the sea and save the child or not to. The immediate desire is not to jump into the cold sea, but then memories from the autobiographical self are activated about being a person who does the right thing, thoughts about what people think of those who do not help children in need, and so on – and maybe also fear of death – so that the person has problems choosing. But when the person finally decides to jump into the sea, it will (sometimes) be the autobiographical self that causes saving the child to be the strongest desire, causing the person to jump into the water.

When we select saving or non-saving as the contrastive effects and ask for the cause of why the person saves the child as opposed to not saving it, the answer is that the cause is the autobiographical self of the person. This means that even if everything in the person's brain happens as if the whole world were determined, if the world is in fact not determined, then sometimes it will be right to select the autobiographical self as the cause of a choice, and sometimes it will be right to select a person as the cause of a choice.

But is not the undetermined event that led to the storm now the cause of the person jumping in, not the person and not her autobiographical self? No, as one can see by considering contrasts. The undetermined event leading to a storm is not the cause of why the person jumps as opposed to not jumping. The undetermined event is the cause of why there is a storm as opposed to no storm. It creates a new setting in which a choice must be made, but it does not cause the choice. In the undetermined setting, the person and her autobiographical self is the cause of why she jumps, and that is why contrastive causation and macro level indeterminism are important presuppositions to show that this theory of free will is coherent.

But is not the undetermined storm a sufficient condition for the person jumping into the water, since I have said that everything may happen as determined in the mind of the person? No, since the distinction between sufficient and necessary conditions does not work to explain what a cause is. For example, both oxygen and ignition are necessary for a house to burn, but neither of them alone is sufficient, and yet both can be selected as causes. The reason is that we find causes by setting contrasts depending on what we already expect and what interests we have. In the case of the house burning, we are interested in why the house burnt as opposed to not burning, given that there was oxygen present. In

the case of the storm, we are interested in why the person jumped as opposed to not jumping, given that there was a storm. If you say that the house burnt because there was oxygen present, I will keep asking why it burnt because oxygen is not the condition I am interested in. And if you say that the person jumped into the water because there was a storm, I will keep asking why the person jumped in because the storm is not the condition I am interested in. But when you cite a certain autobiographical self as the cause (a person with a heroic character, for example), I am satisfied with the answer. In questions concerning free will, our interest is in why people do A as opposed to B given certain conditions. If the cause is their autobiographical self, they act freely.

There is thus no vicious regress here. A person starts life without an independent self, but by use of the general capacity for thinking and feeling in meeting with new and undetermined experiences, that person builds up from within an autobiographical self from those experiences, good or bad, and the autobiographical self gets an increasingly larger role, by which the person gets more freedom and more responsibility. It is the indeterminism that breaks off the regress, since, like in the above storm, when the previous causes like the storm were not determined to happen, it was right to select the autobiographical self as the cause. In the beginning it is not an independent autobiographical self we select as cause, but over time a more and more self-caused and independent autobiographical self can be selected as cause.

Of course, the choices made by this autobiographical self do depend on the general capacities for thinking and feeling that it began with, and which experiences it has will often be a matter of luck, but that is the start package from which an autobiographical self is built, and we cannot demand from free will that there should be an uncaused beginning – that would not give any more freedom, either. It is logically impossible for anything to cause itself before it starts to exist, so we must start with something given, and from there more freedom and responsibility can be achieved. If this general starting point does not work properly, or gets destroyed by outside causes, we find that this reduces freedom and responsibility.

Why all this focus on the autobiographical self? How is it related to me having free will? When I am concerned that "I" have free will, what does "I" refer to then? What is it that makes it important for us to have free will? What should be the cause of our choices in order for us to have a free will worth wanting? Free will is here understood as inner-directedness. When a choice is made, a desire in a person causes that choice. In that sense it is inner-directed since it was caused by something inside a person. But for "me" or "I" to have control, we also want our desires to be something we have formed from within and have control over. The way in which that happens is that desires cause actions, the memory of the

experience of those actions is stored in the autobiographical self, and the experiences stored in the autobiographical self can then later change desires and thus be the cause of future choices. When the autobiographical self causes the choice of a person, that person is not just inner-directed in the sense that actions are caused by desires in his or her body, but it is also an inner-directed inner-directedness since choices over time have shaped the autobiographical self that caused the choice. So when the autobiographical self causes choices, we are inner-directed to a larger degree since the choice is then inner-directed by something inner-directed; in other words, it is an accumulation of inner-directedness.

What I am saying is that free will lies on a continuum, where the degree of freedom has to do with the involvement of the autobiographical self – how strongly it is involved in the deliberation process and how independent it is by having been involved in earlier deliberation processes. The autobiographical self being the cause of a person's actions is, in my opinion, the content of the term "self-control". What it means to control one's actions is to cause one's actions, as argued by Alfred Mele, for how can you have control over something to which you are not causally linked? In order to have an effect on something, one must be causally linked to it (Mele, 1995, p. 10). Many who write about free will focus on control, but without answering by what means it is that the agent controls her actions. Here I answer that to control one's action is simply to cause it. The degree of control that we have over our actions is the different kinds of action I have described, where we have most control when it is an independent autobiographical self that causes the choice. This theory of free will is therefore also a theory of self-control.[177]

This theory of free will is also a theory of autonomy. Autonomy is self-governing in the way it is described here, as something coming in degrees, since choices can be caused by an autobiographical self which can be independent to different degrees. Autonomy can be undermined in the different ways I describe that persons can lack capacity for responsible behavior.

I find that the regress problem is the main objection people have against this theory, so I will now start at the other end and describe the journey from a newborn, unfree human being on his way to getting free will. Let us call this person Willy. I assume here that Willy is born with a normal capacity for thinking and feeling, since this is the kind of person I argue has free will. Willy has neither decided his genetic make-up nor the context he was born into, but when he is

177 Nicholas Rescher distinguishes between having control and exerting control to make the point that you can have control over something if you can cause it or prevent without necessarily doing it (Rescher, 2018, p. 41). I accept the point, but focus on exerting control here, in the sense of being the cause of an action or a choice not to act.

born, a new individual finds his place in the causal nexus of the world. Also, he is born into an undetermined world where different futures may take place.

There will be scenarios, not determined to happen, where we will look for the cause. The answer will be that it happened because that was what Willy's autobiographical self felt was best to do. In the beginning, if we ask why the autobiographical self of Willy felt that this or that was best to do, it will be caused by states of affairs other than Willy's autobiographical self, but quite soon, when Willy acts in a scenario not determined to happen, he will make experiences that will feel good or bad, and these will be stored as memories.

Important memories are stored as part of the autobiographical self, and thus the memory stored causes a change in the autobiographical self. In a similar scenario next time, the changed autobiographical self may find another alternative to be the best – which again gives new experiences which are stored in memory and further change the autobiographical self. The process here described is what I have called the development of a more and more independent autobiographical self; I have said that a person is gradually more and more free the more his or her autobiographical self causes itself and the actions the person does. But this is a process where the autobiographical self does not *determine* what it finds good; it just *acts* on the basis of what it finds good. And the autobiographical self does not control the events that shape what it experiences or how it is shaped by them. So why is this free will?

It is free will because there is nothing more to free will than being the most important cause of actions. There is no possible control beyond being the cause, for what would such control amount to? If it should be more, it would require a detached soul or a meta-self, but why should the actions of such an entity be freer than what I have described here? From where did this detached soul get its criteria for choosing? The only way to build a person who can will what he wills is by a gradual process as described here, where experiences of what is good are made in an undetermined setting.

Many envision free will requiring some entity completely free of external influence, but this does not give freedom. Instead, it gives randomness (because the entity did not choose its own criteria for choosing) and it gives the homunculus problem. The best we can achieve in establishing something that fulfills the definition of free will by being the cause of its own will is an entity that over time shapes its own criteria of choice. I suggest it must be a self that makes experiences of what it finds good and changes itself in light of that. Even if it did not choose its own constitution from the start, it would not have been freer if it was to choose its own constitution from nothing, for what would then have been the criteria of choice?

"What an autobiographical self found good" is the basis of free will and the description of how free will is built step by step. In an undetermined world there are many steps of what Willy's autobiographical self found good, which built Willy's autobiographical self and his free will. He did not cause the building blocks of his own body and brain or how they respond to experience or what he finds good. But he is the cause of every new reaction to scenarios in the world, and that is what is relevant.

Why is that what is relevant? Is this not too small a basis for holding people responsible? Surely where they were born, which experiences they had, etc., has influenced them greatly? It certainly influences them much where they were born, etc., but free will nevertheless deserves the name because it suffices to make responsibility meaningful. If the world is as I have described here, then holding others responsible and blaming them will influence what autobiographical selves find as good, which may cause better choices in the future. I hold a consequentialist understanding of responsibility, which is typical for compatibilist theories of free will, but in compatibilist theories every event of holding someone responsible was determined to happen before the person was born. Holding others responsible becomes more important and meaningful if it is not determined to happen. I will now present my views on responsibility further to show how it coheres with the rest of what has been said.

8.5 Responsibility

The following discussion is a brief presentation of responsibility, based on a larger presentation I have published elsewhere (Søvik, 2019). According to philosopher Derk Pereboom, the idea of responsibility as basic desert is the most common and traditional view that is operative in the larger literature but rarely explicitly formulated (Pereboom, 2007, p. 86). Pereboom argues that responsibility should be defined in terms of basic desert, and offers the following definition of responsibility and basic desert:

> For an agent to be morally responsible for an action in this sense is for it to be hers in such a way that she would deserve to be blamed for it if she understood that it was morally wrong, and she would deserve to be praised if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations. (Pereboom, 2014, p. 2)

The quote explains negatively what it means that the desert is basic by saying what it is not, but it does not positively explain what basic desert is or why the desert is basic. Michael McKenna refers to conversations with Pereboom and explains that according to Pereboom there is no justification for this view since it just is a basic relation which cannot be defined in terms of anything more basic (McKenna, 2011, p. 121). I will suggest here that we should not believe that there is such an irreducible kind of responsibility, but rather that responsibility should be understood differently. Instead of understanding responsibility as basic desert, we should have a consequentialist understanding.

In the following, I will present the best consequentialist theory of responsibility that I know of, which is developed by Manuel Vargas. He argues that holding others responsible is a general strategy for cultivating morally good agency in a society (Vargas, 2013). We will have to dig more deeply into the concept of responsibility, but a good introduction to this theory can be given by looking at how Vargas responds to common criticisms of consequentialist theories of responsibility (Vargas, 2013, pp. 187–195). We start by looking at how Vargas answers four typical objections.

Firstly, we can influence people and animals in many ways, but surely responsibility is more than mere influence. Vargas answers that holding people responsible is a special form of influence since it is about giving people reasons to act differently, so it applies only to people who can respond to reasons.

Secondly, it seems that in many individual cases, it will not have the desired effect to hold a person responsible. He or she need not become a better person by being held responsible. Vargas agrees but points out that holding others responsible is a *general* strategy for cultivating morally good agency in a society and it obviously works in general at that level, even if it does not work in many individual cases.

Thirdly, it seems not to distinguish between holding people responsible and holding people *appropriately* responsible. If holding people responsible is just about influencing their behavior, we could put innocent people in jail as scapegoats in order to achieve the desired result (that people behave better). But that would not be appropriate, so a theory of responsibility must include more than just influence to explain when it is appropriate to hold others responsible. Vargas answers again that a consequentialist theory of responsibility must be seen as a general way of cultivating good agency in a society, and as a general strategy, it will not work to put innocent people in jail. Also, there are other ethical norms which are reasons not to put innocent people to jail; for example, it is not just or not the best way to actualize the best world (more on this in the chapter on ethics).

Fourthly, it seems that we often blame people without being interested in influencing them. We may even blame dead people, where influencing them is impossible. Vargas responds that others can learn when we blame dead people even if the dead cannot, but points out that, in many cases, people will not have the intention of cultivating agency when they blame people. Still, the whole practice of holding each other responsible has this effect.

To sum up so far, we hold others responsible as a general strategy for cultivating morally responsible behavior. But does not *holding* someone responsible presuppose them *being* responsible? We have seen that in the philosophy of responsibility there is a distinction between basic desert views and consequentialist views. These two views have a very different understanding of the relation between a person having capacity for responsible behavior and a person being held responsible. The basic desert view sees the capacity for responsible behavior as primary, and if this capacity is present and a person does something wrong, then it follows that the person deserves blame or punishment, regardless of what consequences blaming the person has. The consequentialist view sees holding others responsible as primary. Holding others responsible by praising, blaming or punishing them is what creates in them the capacity for responsible behavior. It gives them an understanding of the world and adequate feelings about what is good and bad, in light of which they can guide their behavior.

Briefly put, the first view says that only if you have capacity for responsible behavior, you can be held responsible, while the other view says that you can achieve capacity for responsible behavior by being held responsible. But even shaping responsible persons by holding them responsible does require a normal capability of thinking and feeling. Even if one thinks that holding persons responsible is primary in understanding what responsibility is about, one could still say that it presupposes a basic capacity for responsible behavior. This capacity then means that is possible for the person to be influenced by being held responsible through a normal deliberation process. In the following, I will try to describe in more detail how this happens.

What usually happens when we hold others responsible is that we compare a person's action with a moral standard concerning what a person in general should have done in such a situation. This means that when you are in a situation, people will hold you responsible according to a standard related to that situation, and this means that you can also be held responsible for something that you have not done but should have.

For example, if a child is drowning while you are nearby, people think that the morally right thing to do in such a situation is to jump into the water and try to save the child. If you do not, people will hold you responsible and blame you according to this standard, even if you did not cause the child to drown. If you

do not do what people think you should have done – or if you do what they think you should not have done – it is considered blameworthy, whereas a high score on the moral standard of situations is considered praiseworthy (Bok, 1998). I think that this makes it clear that one can be morally responsible for something one is not causally responsible for, contrary to what, for example, Michael McKenna holds (McKenna, 2011, p. 7).[178]

People are usually understood to have a capability of formal reasoning and a normal emotional life. Because they have this, other people expect that they will understand what is true and what is right to do in various situations. We expect that they learn certain norms and that these norms are alternatives they consider when they make choices. As people grow older and become adults, we expect that they have had enough time to think and to make their own experiences in order to understand essentially what is true and right. Then we also expect them to act in accordance with that understanding.

What does it mean to *be* responsible (in the sense of having capacity for responsible behavior) as opposed to being *held* responsible? In the theory offered here in this book, a person is responsible for what he or she does or does not do in a situation if it is type – not token – physically possible for that person to act in a morally different way, in a way which can be influenced by being held responsible.

This type/token distinction is important for discussions about the possibility to act otherwise in a situation, since in a specific situation there may just be one action that is token physically possible for a person to do given the causes that led up to the situation. Nevertheless, it makes sense to hold persons responsible in such cases because the action of holding others responsible influences what is token physically possible to do in the specific situation. Why is that important?

The relevance of this point is that many (typically physicists or neuroscientists) will argue that no humans are responsible for their actions since all our actions are results of causal chains in the brain where it never makes sense to speak of an agent considering alternatives and controlling the outcome. There is just one alternative action that is token physically possible for a person to do in a specific situation because of the causal chain that led to this action. But one may accept that and reply that holding others responsible is a part of

178 This understanding of responsibility as based on a standard for what people should have done is helpful in matters of pulverization of responsibility, where the causal links are vague between people and consequences and it is hard to find anyone responsible. To find the responsible, one should ask who should have acted differently to avoid this consequence. Often it will be the ones with the most power, explaining why it is true that with great power comes great responsibility. But sometimes the answer may also be that nobody is to blame.

the causal chain that led up to the action and influences exactly which alternative is the token physically possible one. Jack seeing the telling look from his father may cause another alternative action becoming the one that Jack's brain executes.

If the world is determined, there is only one token physically possible chain of events that can happen from the Big Bang until the end of time, and then it does not make sense to say that holding others responsible is a meaningful way to change what is token physically possible in a specific situation. But as seen above, there are in fact good reasons to think that the world is not determined.

What I am saying is that there exist no entities in the world to which the concepts of basic responsibility or basic desert refer. All that exists are people who influence each other by comparing actions with a moral standard for what they think should have been done in that situation, and then praising, blaming or punishing people accordingly, which makes persons consider such reactions when deliberating. This practice only has the effect of cultivating moral agency in cases where people have the capacity for considering such reactions when deliberating. Thus we may define capacity for responsibility as capacity for considering such reactions when deliberating.

The standard that we compare the actions of people with is a general standard which presupposes normal development of people with normally functioning minds. Such a normal development was how I described the development of an independent autobiographical self above. To summarize: So far I have said that *holding* people responsible is a general practice of cultivating moral behavior through blame and praise, even if this may not be the motive in many specific cases. This practice presupposes people *being* responsible (in the sense of having capacity for responsible behavior), which means that they can take praise and blame into consideration in a normal process of deliberation.

When it comes to *being* responsible, we should distinguish between being a responsible person in general and being responsible for *x* in a particular situation. Being a responsible person *in general* means that you have capacity for responsible behavior and can take praise and blame into consideration in a normal process of deliberation in any situation where it is possible to deliberate. Being responsible for *x* in a particular situation means that there is something (*x*) that a person has done or not done in this particular situation which could have been influenced by praise or blame through a normal process of deliberation.

Being blameworthy or praiseworthy for *x in a particular situation* means that there is something (*x*) that a person has done or not done in this situation which, according to a moral standard, is blameworthy or praiseworthy. Responsibility in itself is nothing more than a concept referring to people being responsible for

something in a particular situation, which again is reducible to a certain situation involving a person with capacity for responsible behavior, and "capacity for responsible behavior" means the capacity for being influenced by praise or blame through a normal process of deliberation. All these concepts have content which refers to normal processes between humans. There is nothing mysterious, unknown or metaphysical, in the sense of non-empirical or irreducible or undefinable, about them.

8.6 The problem of luck

In the introduction I mentioned three problems for event-causal libertarian theories of free will such as my own. They were the problem of the disappearing agent, the regress problem and the problem of luck. I have discussed the two first problems already, but now it is time for the problem of luck.

The problem of luck is considered by many the greatest, as with Neil Levy in his book *Hard Luck*. He defines luck as an instance of chance which is significant. Chance is not enough for something to be luck, since, for example, the number of trees in the forest is due to chance, but it is not a matter of luck. Levy distinguishes between chancy luck and constitutive luck, where chancy luck are significant chance events in your life and constitutive luck is the luck involved in what kind of person you were born as (Levy, 2011, pp. 13, 29).

Levy argues that we do not have free will, not because of determinism or indeterminism, but because of luck. Libertarians have a problem of luck because of indeterminism, but compatibilists also have a problem of luck, according to Levy, because so much of what happens in our mind and in our life is due to luck. For example, many external factors determine which ideas pop into our mind (Levy, 2011, p. 90). Trying to solve the problem of luck by arguing that a person has a life history where the effects of luck can be cancelled out over time does not help, according to Levy, since every choice in history has been influenced by luck and one cannot solve a problem of luck by adding more luck. Even when we change our character from within, such changes are based on our character, which was a result of luck, whether now or earlier (Levy, 2011, pp. 89 – 92).

I have three main responses to the problem of luck. The first is that we are born with a quite common capacity for thinking and feeling, and those who do not have these capacities are considered less free and less responsible. This basis is used as a standard when we ascribe people free will and responsibility, and luck is then taken into consideration in many ways. We find it mitigating if peo-

ple are born in uncommon circumstances or with uncommon problems, and we find it mitigating if luck plays a strong role in specific situations.

We also give people time through which the effects of luck should be cancelled out. So maybe you are more aggressive than normal, and maybe your parents were not that nice, but after a period of several years we think that, given normal capacities for reasoning and feeling, you should understand that it is wrong to steal, murder and lie. Especially in clear cases we expect that people understand what is right and wrong, whereas we can mitigate these expectations in more difficult ethical cases. So my first response is to accept that there is much luck involved in human action, but that this fact is included in how we evaluate people's free will and responsibility.

Is that a matter of solving the problem of luck with more luck, as Levy argues? No, it is a matter of letting reason and feelings find out over time what is right and wrong, even when disturbed by luck, and adjusting our view of a person's free will by the amount of luck involved. For example, a person growing up and traumatized in a war context may receive treatment and help instead of punishment to deal with his aggressive behavior.

Against this, one may argue that we cannot know to what degree people's choices have been influenced by luck, so we are not capable of taking the role of luck into consideration. I agree that we usually cannot know this to any exact degree, but that is not so important either when I consider responsibility to be something that has the goal of improving behavior. Holding others responsible, even those who have suffered much bad luck, can be a way of helping them to act better morally. It can also be a way of giving them more free will by making them aware of alternatives for actions that they can reflect upon and involve their autobiographical self in the process.

For example, by luck a person can have grown up in a family where they speak all at once, and he does not reflect on this behavior. He finds it normal and thus interrupts other people all the time. When someone then holds him responsible and in some way communicates to him that this is not good behavior, he can realize that there is a choice to be made between talking all the time and letting other people talk, and realize that letting other people talk is a good choice. This makes the person more inner-directed than before and thus freer than before.

One could argue that it was a matter of luck that he was held responsible, but it was nevertheless his own thoughts, feelings and experiences which caused the response. Of course one can say that everything is luck, including that we have a capacity for reasoning and feeling at all. But then the concept of luck has become too broad to be a problem to worry about. As mentioned, we start off with a given basis as standard, and consider luck, freedom and responsibility

in light of that. You may call everything luck, but that does not make it incoherent to continue thinking of free will, responsibility and luck in a narrower sense in the way I have done here.

Still, even if luck is taken into consideration, and holding others responsible can have good effects even when we do not know the amount of luck involved in a person's choice, we do not actually know whether that person has been so struck by bad luck that in effect he has no free will. That is true, but it does not threaten this theory of free will – this is my third main response – because of the fact that people have free will and responsibility to varying degrees in different contexts. People experience different degrees of good luck or bad luck in their lives, but people also have different degrees of free will and responsibility, and we may never know exactly how free and responsible a person really is.

But as I have argued, there are many cases where the autobiographical self is the cause of an action. And some people have had many chances to feel and think and make choices and change their autobiographical self from within to become more and more independent and free. Let us say my grandmother is about to fall off a cliff, but I rescue her. Why did I rescue her as opposed to push her off the cliff? If you answer that it was just a matter of luck, I answer that the concept of luck is used too broadly to be of practical import. Then one has constructed a theoretical framework with a widely defined concept of luck within which everything can be defined as luck and every event explained by luck. Such a theoretical framework is too coarsely grained. We need a narrower concept of luck in order to differentiate between free and responsible actions on the one hand and cases of luck on the other hand – with a lot of mixed cases in between.

8.7 Weakness of will

A good theory of free will should be able to explain weakness of will. I will deal with this problem here, then also return to the topic of strength of desire, as I promised when writing about desires above in Chapter 5.

"Weakness of the will" is a term used for those who do something they do not want to do. For example, I may desire not to eat chocolate and yet I do, and that is called weakness of the will. That the will is weak seems to imply that there is something called will which can be strong or weak. I do not think that there is one entity deserving the name of "the will". Rather, I think that all there is are different desires of different strengths competing to be the one that finally triggers the motor neurons to make the person act. These different desires are what the person wishes to do, and the one that is actually acted upon is the one we

say the person "wanted the most" or "willed". But sometimes we have contradicting desires; that is when it is relevant to speak of weakness of the will.

I find it useful to distinguish between different ways in which the term "weakness of the will" can be used. I concentrate on cases where people want to act in one way but do not do so. This can happen in situations in which a desire caused by the autobiographical self contradicts an innate desire, like an acquired desire to control aggression against enemies contradicting an innate desire to be aggressive against enemies. It can also happen in a conflict between desires that are both caused by the autobiographical self, such as a recent desire versus an older desire, both caused by the autobiographical self. Perhaps you enjoyed drinking good wine before but after having children you have decided not to touch alcohol and now you feel contradicting desires regarding drinking wine. The term "weakness of the will" can also be used for competing desires where people think that one desire is morally good and the other desire is morally bad, so when people act on their morally bad desires, they are said to have a weak will – not able to control their bad desires, for example, for too much alcohol. These may be desires caused by the autobiographical self or they may be innate desires.[179]

I shall disregard the moral examples and only consider the first two examples: where there is a conflict between desires caused by the autobiographical self and desires that are innate, and conflicting desires within the autobiographical self. How can such scenarios lead to acts of weakness of the will? I will start with the first case: a conflict between a desire caused by the autobiographical self and an innate desire. I have said that a person is freer (more inner-directed) when the autobiographical self is the cause of the desire that leads to action than when it is not. This means that if the desire caused by the autobiographical self wins a conflict with an innate desire, then the person exerts more free will (and more self-control, as I argue below) in that choice than if the innate desire wins. If the desire caused by the autobiographical self is for a person not to eat sweets on a weekday and yet that person does because of an innate desire for sweet food, then that is a case of weakness of the will, which more precisely means

179 Richard Hare uses akrasia to denote lack of moral strength (Hare, quoted in Mele (1987, p. 5)). There are also other variables and distinctions concerning akrasia that I have chosen to omit since they are not important to my overall project. A person can show weakness of will in different areas at different times to different extents; it is a difference between weakness of the will in the moment of action and that concerning decisions about what to do later. Weakness of will can relate to control over thoughts, feelings, desires, and actions, but I focus on actions here. The distinctions in this footnote are from Mele (1995, pp. 8, 32, 121).

that the desire caused by her autobiographical self was weaker than an innate desire.

The second case is when there is a conflict between desires that are all caused by the autobiographical self. If these desires are not sorted out, and the person does not know what she prefers, it is not right to speak about weakness of the will but rather confusion of the will. But if there is a new desire which the person now has as a voluntary desire (a desire she is happy with having), and an older desire that the person wants to get rid of, we can speak of weakness of the will when the person acts on the old (and now involuntary) desire she wants to be rid of. For example, this would be the case if a person joined a sect and there acquired some strange moral or religious views that influenced her desires, then left the sect and considered her views wrong, but was still unable to shake off some of the old desires. In this case, there is also weakness of the will if the old desire wins over the new desire since the new desire is caused by a more independent autobiographical self than the old desire – this is a weakness of the will in this context. The reason the new autobiographical self is more independent than the old one is that it has changed at least one more desire than before.[180]

Is it not strange that a person can say she wants one thing the most, yet she does something else? Have I not said that the strongest desire leads to action? How is weakness of the will then even possible?[181] It is time to consider, in more detail, what it means that one desire is stronger than another. I have already mentioned the consciously experienced side of strength of desires – that it has to do with the intensity and amount of pleasure or displeasure – but I have also said that, by definition, the strongest desire is the one that leads

180 Alfred Mele discusses the unorthodox case of a thief who thinks it would be best not to steal, but decides to do so anyway. Even if he is afraid, he steels himself and commits the crime. In this case, it seems normal to call it weakness of will if the thief did *not* steal, whereas it is self-control when he steals, even if he judges it best not to steal. Mele's own solution is to distinguish between an evaluative judgment and an executive judgment, so whether the thief steals or not will be in accordance or not with his executive judgment (Mele, 1995, p. 74). I interpret the story differently. If he does not steal because he is afraid, I would not think of it as weakness of will, but weakness of courage. If he steals, it may be a case of weakness of will, but this depends on the details of the story. Is his desire to steal in conflict with a newer desire not to steal, or is it not? This is what determines whether it is weakness of will, and not just that he thinks that stealing is not good, since if he desires to steal he must also think that stealing is good in one sense or another.

181 Some have argued that weakness of the will is impossible. As examples Mele mentions Socrates, Richard Hare, Gary Watson and David Pugmire (Mele, 1987, p. 8).

to action.[182] However, it seems that sometimes one desire can feel stronger than another yet not be the one that leads to action. I can feel a strong desire for pizza yet eat fish, which I do not feel a strong desire for. Is there then a contradiction in what I am saying here?

One part of the solution to this problem is that a consciously felt desire may activate many strong, contradicting desires in the mind, even if they are not consciously felt or consciously thought. Let us say a doctor suddenly feels a strong desire to have sex with a patient, yet does not act on this desire at all.[183] The desire probably activated many thoughts in the doctor's mind about scandal, prison, broken marriage, shame and so on, all with strong negative feelings connected to them (even if they were not consciously felt or thought), so there was a stronger desire in the doctor's mind not to take initiative to have sex with the patient even if not consciously felt. Mele mentions a problem raised by G. F. Schueler for those who think that the strongest desire leads to action; namely, how a person can intend to go to a school meeting even if he does not desire to do so (Mele, 2003, p. 29). My answer is that in the person's mind there are thoughts and desires concerning what people expect and what will happen if the person misses the meeting for no good reason, so the strongest desire is to go to the meeting even if it does not consciously feel that way.

Strength of desire then has to do with how good or bad the connected emotions are and how much they occupy the mind, but it may be both conscious and non-conscious. This determines the strength of the desire itself, but some desires are more easily executed than others because of physical factors in the brain. In our brain, certain neural patterns have stronger synaptic connections than others. Every time a pattern is activated, its parts become more strongly connected to each other.[184] This is why negative thoughts you have thought many times easily come back and it is also why many psychologists tell you to repeat good thoughts many times. Synaptic strength cannot be experienced directly, but it

182 Supported by Singer in Geyer (2004, pp. 56–57). Mele discusses the critique that referring to the strongest desire does not explain an action. For example, if I ask why Bob threw the rock, the action is not explained by saying that he desired that the most strongly. Mele suggests that we do not think of it as an explanation because we already take it for granted that he desired it the most, but want to know why he desired to throw the rock (Mele, 2003, p. 165). So, even if referring to the strongest desire does not explain a particular action, it does explain in general which desires lead to action when it is further explained what strength of desire means.

183 The example is from Mele (2003, p. 162).

184 As quoted above on Hebb's rule ("Neurons that fire together, wire together"). Synaptic strength does not make a desire stronger in the sense that it is connected with stronger feelings, but stronger in the sense that it more easily leads to action.

can be experienced indirectly or inferred by how easily thoughts come to you, or how hard they are to get rid of. Synaptic strength is also part of the reason it is easier to jog if you engage in it as a regular habit than if you just have to decide at one moment between jogging and the sofa. If you are a regular jogger, you may feel a strong desire for the sofa and yet the desire to jog at the regular time makes you jog because of physical aspects of neural patterns in the brain. In that case, the desire for the sofa may in itself have been emotionally stronger than the desire to jog (at least that was how it felt consciously), but since the desire to jog had an easy path to be executed, it led to action, and was therefore strongest in the total sense where strongest means the one that led to action.[185]

What I tried to say in the previous paragraph was that not only the strength of desires (which have already been influenced by thoughts) matters, but also habits. Schroeder et al. describe this by saying that the motor basal ganglia in the brain select which action to execute based on desires, but also based on the internal structure of the motor basal ganglia themselves, which are again a result of habits (Schroeder et al., 2010, p. 82). Desire strength is thus both the strength in the moment, but also the strength it has built up over time by developing habits.

However, it is even more complicated than that. There are several ways in which thoughts, feelings and desires influence each other while physical factors also play a role, and this makes it difficult to set it all out in an easy formula. My aim here is only to point out the main contours of how it works. Many exceptions will not be treated, but I shall mention some here to indicate certain complicating factors. Alfred Mele presents the case of Ian, who most desires to sit and watch television although he should go out and paint the shed. After a while he shouts to himself: "Come on, get out!" And out he goes. How is that to be explained? Mele mentions several possibilities. One is that perhaps Ian has a habit of following orders and shouting to himself provides the necessary adrenaline boost to change what is the strongest desire. Furthermore, we can have several desires simultaneously, so even if painting the shed is not the strongest desire, Ian may try a technique for making it the strongest desire, like shouting to himself. Other techniques for changing the strength of desire can be things such as imagining a piece of chocolate cake to be a piece of chewing tobacco (Mele, 1995, pp. 45 – 46, 179).

185 Synaptic strengthening also explains why people over time develop more stable characters, habits, and preferences. At least I think so, and it fits very well with psychological research on personality, as shown in Mischel and Shoda (1995).

Thoughts influence desires. In my own experience, when a thought gives me a specific feeling or desire, I can change the feeling or desire by taking a meta-perspective: thinking about the fact that I am thinking about something which feels like that. It will often then feel different, since it is a new thought with a new feeling connected to it. This experience of mine supports Damasio's somatic marker hypothesis. Our thoughts about how likely something is to occur or when it will happen are thoughts that influence our desire for something. For example, I desire to have a million dollars, but that does not make me buy lottery tickets, because I think it so unlikely I would win.

What I am saying is that there are several things determining strength of desire, so it is not strange that one desire leads to action even if a person has another contradicting desire which may consciously feel stronger. This is a kind of conflict we have all the time and it is a life project to create an autobiographical self which integrates our desires in a coherent way that we like. I realize that my theory about the strength of desire is unfalsifiable, since I can always appeal to non-conscious desires or physical neural connections in order to defend my claim that the strongest desire leads to action. However, there is independent support for the claims and they do explain features that otherwise seem strange. I have already mentioned examples which clearly indicate a physical side of desire strength: The lateral hypothalamus can be stimulated electrically and make a person or animal feel compelled to eat and drink, whereas when it is destroyed, animals must be force-fed not to die. Likewise, other areas of the brain could be stimulated to make people suddenly very sexually active (Joseph, 1996, pp. 171–172, 187). These data fit very well with an understanding of desires having different strengths and causing action. There is much neuroscientific evidence supporting that a choice consists in brain activity where desires reach a certain threshold, which then activates motor neurons leading to action (Roskies, 2014).

Helen Steward claims that there is no empirical evidence of non-conscious decisions and that desire strength is an empty concept (Steward, 2012, pp. 159–160, 170–171), whereas I think that the examples in the previous paragraph show the opposite. She nevertheless accepts that conscious beliefs and desires guide our action, while others reject the whole idea of beliefs and desires explaining action. Alex Rosenberg is an example of someone who argues that desires and beliefs are illusions irrelevant to explain what really causes action in the brain (A. Rosenberg, 2011; A. Rosenberg, 2018).

According to Schroeder et al., it is standard neuroscience to think that desires are something physical in the brain causing action (Schroeder et al., 2010, pp. 84–87). I am open to the idea that the physical realizers in the brain of what we call desires and beliefs may turn out to be and work quite differently from how we describe the causal chain of strongest desire leading to ac-

tion. Nevertheless, I have offered a theory above of how conscious experiences have influenced which brain processes have evolved, which explains why we have the systematic relations we have between conscious beliefs, desires and actions (and it is strange that it should be so systematic if they are just illusions). This means that even if brain processes realizing beliefs, desires and actions are very different from how we describe conscious beliefs, desires and actions, it does not imply that our understanding of the relation between beliefs, desires and actions are wrong. Instead we would just have learned more about the physical side of the process. If a new theory shows that it is more coherent to abandon any role for conscious beliefs and desires, we should of course give up that belief, but so far it just seems to be a self-contradiction for Rosenberg to desire us to believe that there are no beliefs and desires.[186]

8.8 Some final objections

In the introduction, I presented some objections against free will from neuroscience. However, I will not go deep into the neuroscience. The reason is that the findings from neuroscience often considered to contradict free will merely contradict specific theories of free will like agent and non-causal libertarian theories. They lose their force if both conscious and non-conscious mind are understood as causal processes in any case, and they are even less relevant for theories of free will that emphasize free will as a result of decisions made over a long period of time, such as this one, since neuroscientific research on free will is usually made on spontaneous decisions.

I now turn to the objections from Alfred Mele that a theory of free will cannot be based on external indeterminism. Mele has four counterarguments (Mele, 1995, pp. 195 – 196). The first argument is that if an indeterministic bomb exploded, for example, on 15 September 1969 it gives another future than if it had not exploded, but it does not give us free will. I argue that it does, because in this scenario it is not determined before our birth what will happen. In this scenario, one can select the self as the cause of an action without determinism also being the cause of the action, and thus the self becomes the ultimate source of the choice. In this scenario, since undetermined events like the bomb are possible,

186 At the University of Oslo, 27 Nov. 2019, Rosenberg gave a lecture ending with him acknowledging on the last PowerPoint slide that his suggestion was self-contradictory and that this was an objection he had to continue working with.

other events will also be undetermined, and so our selves will be the cause of many actions.

Mele tries to push the point further by suggesting that if the bomb did not explode on 15 September 1969, and that this bomb was the only indeterministic device in the world, then we would no longer be free after that date. But Mele finds it preposterous to suggest something like that. I reply that such a world would be extremely close to a deterministic world. If the bomb were set such that it would either explode or not on 15 September 1969, then only one future would have been possible up to that date and after that again only one future would be possible. In effect, the world would have been determined up to that date and after that date, and so I agree that we would not have free will in such a scenario. But it is still important that the world is undetermined in general, with several undetermined events, and this argument does not refute that. The more undetermined the world is, the larger the role our selves can play.

Mele asks us to consider two worlds, one of which has undetermined bombs and the other determined. Let us then say that none of the bombs go off, so that everything that happens in the two worlds is identical. Could it then be right to say that those living in the determined world are not free, but those in the undetermined world are free? Here it is open whether the two worlds by accident happen to be identical, or whether everything in the world with the undetermined bombs (except for the bombs) is determined. If everything is determined but the bombs, my reply is like above, that it almost becomes a determined world. But if it accidentally is the same, but many events were undetermined, our selves will play a larger role. So an example with only a few undetermined bombs not going off does not show that a world with indeterminism as a general feature cannot support free will.

Mele's final point is that external indeterminism does not secure free will, as libertarians want the future to be up to them. I agree that free will implies that what happens in the future is up to the agent, but what that means is that the self is the ultimate cause of the action, and the self can be that as long as the world is not determined. As far as I can see, none of Mele's counterarguments refutes the solution I here offer to the problem of free will.

I move on to considering another kind of objection, namely that I have overlooked the importance of social conditions for freedom. I agree that this is a perspective which should complement my presentation, while being fully compatible with it. To introduce the point, I will introduce a famous discussion between Isaiah Berlin and Charles Taylor. Two main points will be considered: First, a point from Berlin on how social recognition determines both who we are and what we want and what is socially possible and impossible actions. Then a point from Taylor on how internal constraints can diminish our freedom.

Berlin wrote a famous essay where he distinguishes between positive and negative freedom. Negative freedom means that others should not deliberately interfere to prevent you from doing things you could otherwise do, while positive freedom means that you are the person determining who you are and what you do (Berlin, 1969).[187] Berlin also added a freedom he called social freedom, where the point is that you need to be recognized by another as a person with a will in order to become a free person who realizes that you are an individual with our own will. It is through the recognition from others of us (as English, church members, football fans, etc.) that we get our identity and desires (Berlin, 1969).

The concepts of positive and negative freedom seem very similar to what I have called freedom of will and freedom of action. Freedom of will is being the source of your choices, while freedom of action is having alternatives for action. While my focus was on the conditions for free will in relation to determinism, indeterminism and luck, Berlin adds the importance of the social conditions for free will. While I focused on alternatives being type physically possible, Berlin describes how social interaction makes it possible for us to understand ourselves as persons choosing among alternatives.

Doing this, Berlin underscores the role of recognition in this social process. "Recognition" is a term which is used in many ways. Arto Laitinen and Heikki Ikäheimo distinguishes between three meanings of recognition: There is first a basic sense of recognition as *identification* of something as something (e.g. "this is a person"). Then there is a second sense of *acknowledging* norms, facts, and values as valid. Third there is a *mutual recognition* process between humans or groups recognizing each other as recognizers (Laitinen, 2002). All of these forms are relevant for free will, since it shapes what we understand as relevant alternatives for actions to choose among.

Recognition both confirms reality and creates reality. For example, you cannot be popular if nobody recognize that you are popular or be somebody's friend if the other does not recognize that you are their friend. Recognition (and with it language) confirms and creates the understanding of ourselves and our surroundings in which we develop an independent autobiographical self. I have focused in this chapter on physical possibilities and impossibilities, but I think that this perspective from Berlin is a good description on how social conditions determine what we think of as possible and impossible alternatives at the macro level where we make our choices. For example, in order for me to think of myself

187 In this article, Berlin also used the terms "freedom from" for negative freedom and "freedom to" for positive freedom, but as others have pointed out, this description is quite misleading, since many scenarios are equally well described as freedom from or freedom to (Nys, 2004, pp. 216 – 217).

as someone who can get married and in order for me to have a desire to get married, there must be something called marriage and others who think of themselves as possible candidates for marriage and who recognize me as a possible candidate for marriage. Language, society and recognition partly create and determine how our autobiographical selves are experienced and understood, what we desire, and what alternatives for actions we perceive.

It is not only that recognition is important and constitutive for our understanding of our alternatives for actions – recognition is also important and constitutive for what are our alternatives for action. Whether people recognize us and our intentions or not determines what is possible for us to do and not do at the social level, since so much of what we do happens in cooperation with and by means of others. All of this is much more complex than what I can describe here, but I wanted to recognize the point. Elsewhere I have written about social conditions for freedom and what positive and negative effects different form of influence have on different aspects of freedom (Søvik, forthcoming-b).

This is fully compatible with what I have said in the rest of this chapter, and what this book says of the world, mind and language should be enough to see how to translate between these theoretical frameworks on free will. But the social level is an extra layer of complexity shaping the interacting persons, desires and states of affairs considered as alternatives.

The second point I shall consider, is a point Charles Taylor makes when offering some interesting comments to Berlin's article. Taylor defines *negative freedom* as having opportunities to act and *positive freedom* as exercising control over your actions, but argues that the negative freedom does not make much sense without the positive freedom: having opportunities does not make you free unless you actually are a person exercising your will (C. Taylor, 1979, pp. 177–178).

But even being a person exercising your will and having opportunities free from external constraints may not be enough to make you free if there are internal constraints on your actions. Maybe you are paralyzed by fear, or people have forced you to internalize their standards, or you have false conscious perceptions of matters. Taylor argues that you may be wrong about what you really desire. This can be both because you misunderstand what is a good way to your goal, but also because you have false beliefs about the goal (C. Taylor, 1979, pp. 176, 187–193).

Taylor discusses this problem as a problem of how much freedom society should give people and in what way, and I shall return to that problem in Chapter 15.3. Here I shall consider Taylor's point as a possible objection to the theory of free will in this chapter. One could object that in order to secure free will it is not enough that an autobiographical self causes itself, if conditions make it the

case that the autobiographical self causes desires we do not think that would have arisen in more normal conditions.

To answer this objection, we should distinguish between three things: freedom of will, freedom of action and responsibility. I have argued that freedom of will is for the autobiographical self to cause its own content and actions, but that this comes in degrees. In a setting with strong social pressure or where things do not work properly in the brain, the autobiographical self may have less opportunity to cause itself than in a setting with more possibilities to explore without external pressure. The contexts within which people develop their autobiographical selves will be different for all in any case, and so that does not contradict my theory which says that people are independent to different degrees.

When it comes to freedom of action, different contexts will give people different opportunities, which will influence how much freedom of action they have, and also how happy they become. They would have made different choices in different contexts, and there is no agreement on what is the normal or the perfect conditions. I shall return to what are the best conditions in Chapter 15.3, and suggest that since we do not know what they are, that favors that people should have freedom of action to explore alternatives.

When it comes to responsibility, we do compare people's actions with a standard for what we think they should have done given normal conditions. If people have grown up in a non-normal condition (like a very mind-controlling sect) or there is something non-normal with their brains, we may hold them less responsible since their actions have been strongly shaped by mind-external factors in non-normal conditions. Again, this raises the question of what should be considered normal conditions, a question I will return to in Chapter 15.3.

When discussing positive and negative freedom, Belin and Taylor are most interested in how society should be formed in order to secure as much freedom as possible for its citizens. This is the discussion I shall return to in Chapter 15.3, and then I will pick up again the debate from Berlin and Taylor. Here I will just mention one final objection, which is more of a speculation, from Yuval Noah Harari.

Harari has speculated that in the future, it will make sense for people to let machines choose everything for them, since the algorithms know people better than they know themselves (Harari, 2017, p. 384). The idea of "free will" will cease to make sense, but Harari argues that belief in free will is a result of faulty logic and the idea of the self is just an imaginary story (Harari, 2017, pp. 331, 353). I have suggested a coherent theory of the self and free will in this book, which lets us see why it does not make sense to leave our choices up to a machine. The reason is that free will means a continuously increased independence by being the cause of your own choices in an undetermined world. This happens

gradually by making the choices in undetermined settings. If the machine chooses, it and not you, is the cause of your choices, and even if you gave the initial instructions, you could not know what would happen in the future. Since a person is a whole life process, it means that she loses her free will if a machine instead of her is the cause of her choices. To sum up this last line of reasoning, finding your meaning of life and the best way to the best world, presupposes that both governments and machines allow us freedom of action to form our own independence.

In Part Two I have presented an understanding of how dispositions, desires, choices and thoughts could evolve gradually and function as a causal process. This description fits very well with how artificial intelligent agents are developed. An artificial intelligent agent is something that can perceive its environment through sensors and act upon that environment through actuators (S. J. Russell et al., 2010, p. 34). In almost all artificial intelligence you will find four types of agent, with an increasing degree of complexity: Simple reflex agents, model-based reflex agents, goal-based agents and utility-based agents.[188] *The simple reflex agent* acts directly on its current percept with a condition-action-rule "If A, then B". For example, a self-driving car could have the condition-action-rule: "If the car in front of you brakes, then brake", and humans have something similar, like "if something approaches your eye, then blink".

The model-based reflex agent does not only have an automatic response to what happens, but in addition it has a model of the world which says what happens if the agent does A or B, based on previous percepts. This allows it to make more advanced choices, not only based on what is currently perceived, but also previously perceived. *Goal-based agents* are even more advanced, since they have models for different possible worlds describing what happens if the agent does A, B or C, and in addition they have goals that the alternatives can be matched against. This can be made very complex with several goals for finding the best way to a goal. *Utility-based agents* add a utility function to measure to measure goals against each other and choose goal-based on maximal utility. Russell and Norvig write that this is the same as if the agent was to ask itself how happy the different goals would make it, but that since "happy" does not sound very scientific, computer scientists use the term "utility" instead.

In addition, you can have learning agents, and there are four components in a learning artificial intelligent agent: a learning element responsible for making improvements, a performance-element selecting actions, a critic which evaluates

188 The descriptions of the different agents are from S. J. Russell et al. (2010, pp. 46 – 58).

and determines modifications for the future, and a problem generator, suggesting new actions to be tested. The agent can learn either based on an internal measurement of how well it predicted outcomes, but it can also learn from an external standard of what is useful. The way that happens is that feedback is interpreted either as reward or penalty. Russell and Norvig compare it with pain or hunger in animals. All learning is about making the parts fit better together to increase the utility for the agent.

There are obvious parallels between artificial intelligent agents and the description of animal and human minds given in this part of the book. The simple reflex agent is like simple animals with brains driven by "if A, then B"-dispositions. The model-based agent has more advanced representations of the world. Goal-based agents can represent alternatives to choose among, while utility-based have desires of different strength making the select the presumed best alternative. In addition, the agents can learn, based on feedback, in the same way as we can develop more independent autobiographical selves, and in the same way as we learn moral behavior by being held responsible. I write more about the comparison of humans and artificial agents in Chapter 15.2.

To sum up the whole chapter: Free will is a matter of inner-directedness and exists on a continuum. People have a small degree of free will when their desires cause their choices without influence from their autobiographical self. They have more free will when their autobiographical selves are the cause of their choices and even more free will if they have independent autobiographical selves. This means that different people have different degrees of free will in different situations. It also means that you can attain more free will than you have now by exploring alternatives for action.

This chapter concludes Part Two on the mind. We now have a much deeper understanding of what happens when we understand and discuss something as true, which we can use when trying to understand other things in the world. On the other hand, we have also laid out a theory of mind which fits into the natural world that we shall consider further in the next part.

Part Three of the book deals with different elements in the world different from mind, and again an important goal is to show how very many things that could seem to be irreducible entities can be reduced to values actualized in a field. The topics to be discussed are time, fundamental concepts in physics, mathematical truths, and probability. First, it is time for time.