Chapter 1
Varieties of Decision-Making

Bert Heinrichs

Abstract Starting with a broad use of the term, this chapter examines different types of decision-making and agents. A typology is proposed that can help to avoid confusion and promote understanding, especially in discussions between different scientific disciplines. The typology includes reporting/non-reporting agents, fully/partially transparent agents, and observer-/self-transparent agents. However, the most profound distinction turns out to be between non-discursive and discursive agents. This is mainly because discursivity and responsibility are closely linked. Discursive agents therefore have a special position: they must decide how they deal with other types of agents and what they use them for and how they use them in their own decision-making. So far, only humans can be considered as discursive agents.

Keywords Decision-making · Artificial Intelligence · Agency · Discursiveness · Responsibility · Morality

1.1 Introduction

We often use the term "decision-making" in everyday language, and it is also common in some scientific disciplines, in particular in decision theory (Steele and Stefánsson 2020) and operations research (Salhi and Boylan 2022), but also in psychology (cf. the chapters in Part II of this volume), and computer sciences (Kochenderfer et al. 2022). In philosophical theory of action, on the other hand, the concept is less prominent and ranks behind other concepts, most notably that of intention (Anscombe 2000; Davidson 2006; Setiya 2022). Like many other terms, "decision-making" is used in a variety of meanings that only partially overlap. While in philosophy the notion is often closely related to the concept of responsibility (Heidbrink et al. 2017), in other

B. Heinrichs (✉)
Institute of Neurosciences and Medicine: Brain and Behaviour (INM-7), Forschungszentrum Jülich, Jülich, Germany
e-mail: b.heinrichs@fz-juelich.de

Institute for Science and Ethics (IWE), University of Bonn, Bonn, Germany

U. Ettinger et al. (eds.), *Decision Making*, Studies in Neuroscience, Psychology and Behavioral Economics, https://doi.org/10.1007/978-3-032-00880-0_1

fields it is used rather broadly to describe the behavior of various types of systems in their respective environments. These systems can be humans, non-human animals, or artificial systems such as robots or computer programs. In what follows, I begin with such a broad use of the term and consider then whether it might be helpful to distinguish different types of decision-making to avoid confusion and promote understanding, especially in discussions among different scientific disciplines.

1.2 Decision-Making Broadly Conceived

Suppose S is a system that exists in an environment E, such that S has a set of observations O about the state of E, and S can perform an action A based on O that causes changes in E or S itself. Then let S be an *agent* and let the process that leads S to the selection of A from a set of possible alternatives be a *decision* D (this definition is inspired by Kochenderfer et al. 2022, pp. 1–2; cf. Russell and Norvig 2020, chap. 1).[1]

Based on this definition, S could be a person who notices the smell of fresh coffee in the house and as a result decides to go to the kitchen to get some. Or S could be a cat strolling through the garden, hearing a rustle in the grass, and decides to crouch down to launch an attack on a mouse it suspects to be there. Or S could be a surveillance system that registers suspicious transactions on a bank account and then decides to issue a warning to the regulatory authority in charge. In each of these three cases, the system S responds to observations O (the smell of coffee, the rustle in the gras, the suspicious transactions) from the environment and selects an action A (going to the kitchen, crouching in the gras, issuing a warning) out of several alternatives. Consequently, at this level of abstraction, all three systems—the human, the cat, and the surveillance system—can be said to make decisions. Note that the concepts of agent and decision-making are inextricably linked in that they describe the same phenomenon from different angles. We are concerned with entities and events that are special in some way. For example, stones are not agents and cannot make decisions.

An interesting question with respect to systems that are agents is how good their decisions are. To answer this question, one needs to introduce some standard V by which to measure the changes C brought about by S through A in E or S itself. Furthermore, one needs to estimate which alternative changes C* would have been brought about by an alternative action A*. One can say a decision D was good if the changes C brought about were positive as measured by V. One can further say a decision D was optimal if there was no alternative action A* so that any other changes C* brought about would have been more positive than C. An agent can be

[1] The term "observations" is to be understood here in a very broad sense and includes many types of inputs based on states of E to A. However, according to this definition, decisions are only those types of selections that are triggered by an external input, i.e. no purely internal status changes of A.

called rational if S acts (often enough) in such a way that the chosen actions $A_{1\text{-}n}$ are guided by the standard V.

At this very abstract level, understanding (good or optimal) decisions poses no fundamental problems. To be sure, this does not mean that it is easy for an agent S to make a good or even optimal decision. One reason for this is that the changes C or C* (i.e., the effects of A or A*) are subject to considerable uncertainties. For S it can be very difficult to foresee whether A will have a positive effect, and even more so whether A will have a more positive effect than A*. Additionally, the set of observations O can also be subject to uncertainty, which complicates matters further. Think about the person who smelled the coffee: maybe the smell came from outside. Or think about the cat: maybe there was a mouse there, but the attack came too late. Or think of the surveillance system: maybe it was a malfunction in data transmission. In all these cases, the decision that the agent made would not have been a good one.

Moreover, it can be difficult to define an appropriate standard V and to evaluate A or A* in terms of V. Apparently, the term "utility" is relevant here and hence defining and calculating utility functions is an important endeavor in decision science (Steele and Stefánsson 2020, Sect. 1.2).

Taken together, it should be clear that decision-making can be challenging for an agent, and it can be especially challenging to design an artificial agent in such a way that it makes good or even optimal decisions in a complex environment. Nevertheless, the basic constellation of decision-making as the selection of an action from a set of alternatives based on observations does not seem to be particularly complicated. In the following, I shall deal with some conceptual difficulties that may nevertheless arise.

Saying that the surveillance system decided to issue a warning to the regulatory authority after noticing suspicious transactions on a bank account may seem excessively anthropomorphic and therefore misguided. Some may even consider the notion of an agent for a computer system to be a misnomer. Taking this charge seriously, it might be worthwhile to look for differences in decision-making and consider how these relate to the nature of the agents involved (humans, non-human animals, artificial systems). This is how I shall proceed in this chapter. This will allow for a more precise use of the term decision-making.

1.3 Types of Decision-Making and Types of Agents

Based on the above definition, making a decision is essentially making a choice from a set of alternative actions. As shown, very different systems qualify for decision-making under this definition. Let's start now by looking at a simple artificial system and examine it a little bit more closely. Imagine a (simplified) robotic lawnmower that has various sensors, including a bump sensor to avoid obstacles, a tilt sensor that stops cutting blades if the robot is in a compromised position and a lift sensor that stops cutting blades if the robot is lifted off the ground. Apparently, the robotic lawnmower is a system that has the characteristics mentioned in the definition above.

It is therefore a system that makes decisions, i.e., an agent. If, for example, the lift sensor indicates that the robot has been lifted off the ground and the robot stops as a result, then it has made a decision, i.e., it has selected one of two possible actions—continue mowing or stopping. Note that if it was indeed lifted and if there was no sensor malfunction, then it was a good decision because it was in accordance with the manufacturer's standard aiming to prevent human injury. In fact, the decision was even optimal because there was no alternative action available that would have led to a better outcome.

An external observer might ask why the lawnmower stopped cutting. The lawnmower itself is not able to answer this question. Compare this with the person who smelled the fresh coffee. An observer might ask him why he decided to go to the kitchen. Obviously, and in contrast to the lawnmower, the person could have answered that he wanted to get a coffee. A distinction can be made, then, between two types of agents, those that can provide information about their own decisions and those that cannot. Let's call the first type *reporting* and the second *non-reporting*. Note that reporting systems do not need to be able to use speech. If the lawnmower has a built-in LED that indicates when the lift sensor reports that the robot has been lifted off the ground, this will also count as reporting. Note also that non-human animals are generally non-reporting according to this classification since they usually do not articulate why they made a particular decision. Rather, it is us who interpret their behavior—although often quite accurately. Humans, on the other hand, are by default reporting agents, at least when it comes to healthy adults. In artificial systems both variants—reporting as well as non-reporting—exist.

Let's go back to the lawnmower and assume that it does not have a built-in LED that indicates whether one of the sensors has passed on a signal. Since the system is non-reporting, the observer may still want to know why the robot stopped. For this purpose, the observer could disassemble the lawnmower to try to understand its functioning. Considering the rather simple setup, the observer could certainly find out why the decision to stop was made. Even more, the observer could, with some technical skill, figure out the entire functioning of the lawnmower. Let us call agents where this is possible *fully transparent*.[2] Humans and non-human animals apparently do not fall into this category. Even if we dissected a non-human animal (which would, of course, be problematic for ethical reasons) we would not find out why it makes the decisions it does. Animals are simply far too complex to understand their behavior in this way. We must therefore classify animals as only *partially transparent*. Although we can often capture their decisions-making quite accurately through observation, there is always some uncertainty and room for interpretation, so predictions are never entirely certain when it comes to animal behavior. Take the cat from above: Perhaps it did not lie in wait because it suspected a mouse in the grass, but it wanted to hide from a supposed predator. The original interpretation of the decision situation might be more plausible but is by no means beyond doubt.

[2] Transparency here refers exclusively to the decision-making process, i.e., the action selection process, but not to the observations or evaluative standard.

Although humans can provide information about why they made a decision, it would be naïve to believe that they are fully transparent. Our decisions are often influenced to a considerable extent by unconscious factors. Humans are therefore also only partially transparent, both for themselves and others (more on this below).

What about artificial systems? Apparently, there are some such systems which are fully transparent as is the case with the lawnmower. Others, however, are only partially transparent or perhaps even non-transparent. The latter is true, for example, for deep learning systems. To be sure, such systems have been designed by programmers. In the course of training with large amounts of data, however, these systems change the weights of their hidden layers, which ultimately leads to their decision-making being incomprehensible to humans including their programmers. This fact raises a number of problems, including ethical ones (Müller 2023, Sect. 2.3). Therefore, there is a demand that such systems should become explainable or, in the terminology suggested here, partially transparent, which has been the subject of intensive work for some time now (Holzinger et al. 2022).

Let's briefly summarize: One can distinguish between *reporting* and *non-reporting* agents on the one hand and between *fully transparent*, *partially transparent* and *non-transparent* agents on the other hand. Obviously, all agents that are reporting are at least partially transparent.[3] Humans are usually reporting and partially transparent, while non-human animals are non-reporting and partially transparent. For artificial systems, all variants are possible.

Another distinction that can be made has to do with how an agent is informed about its own decisions. The person who decided to go to the kitchen to get coffee usually knows why he went to the kitchen. This is why he can answer the observer's question. This is different from the lawnmower, which has an LED that indicates when a sensor has passed a signal. The lawnmower itself does not have the information that the sensor was active, i.e., this meta-information is not represented in the robot. It is simply an information that indicates to an external observer why the lawnmower has decided to stop. One could say the lawnmower is merely *observer-transparent*. In contrast, humans know about the factors that led to a decision—at least partially. One can therefore say that humans are (partially) *self-transparent*.[4]

It is important to distinguish between reporting/non-reporting, degrees of transparency and observer-transparent/self-transparent as each of these categories indicate different capacities of an agent. Above all, this seems important for avoiding a particular misunderstanding: One might think that artificial systems cannot be self-transparent and that this is something that distinguishes them from human agents, but this is not the case. It is true that the lawnmower with the built-in LED does not have this capacity. But more complex artificial systems are conceivable that have

[3] To be precise, another qualification may be necessary here: The system could be non-transparent and still be reporting. In this case, the reporting would of course very often be inaccurate. However, it is questionable whether the system's statements should still be regarded as reporting in this case.

[4] The notion "self" in "self-transparency" is not used here in a reifying way. Rather, it only indicates a self-referential structure. John Flavell coined the term "metacognition" for "knowledge and cognition about cognitive phenomena" (Flavell 1979, p. 906). The notion of self-transparency is meant here in the sense of such metacognition.

representations of their own decision-making. One can imagine, for example, that the program for monitoring bank accounts has internal representations of this kind. If this were the case, then in the event of an alarm, the program could be questioned as to why it made its decision. For example, the program could specify factors such as the frequency of transactions, the amount of money involved, and the type of source and destination accounts that led to the alert. Note that self-transparency is obviously only possible if a system is at least partially transparent, while the reverse is not true.

Let us briefly summarize again: I have argued that there are different types of agents capable of different types of decision-making. Humans have been classified as systems that are (typically) reporting and both partially self-transparent as well as observer-transparent. Non-human animals, on the other hand, are systems that are non-reporting and partially observer-transparent. Artificial systems cover a wide spectrum in this classification schema. They can be reporting as well as non-reporting; they can also be fully transparent, partially transparent or non-transparent; finally, they can be self-transparent or merely observer-transparent.

Generally speaking, complex systems perform worse than simple systems when it comes to transparency. However, this does not only apply to artificial systems, but also to humans and non-human animals. We, too, are only partially transparent to ourselves as well as to our fellows when it comes to our decisions. By the way, one can speculate whether there are also completely self-transparent agents. A divine being—if there is one—might fall into this category. But if such a divine being exists, it seems to be non-reporting.

Does the typology suggested here imply that the charge of anthropomorphism, mentioned at the outset, is completely mistaken? Apparently, humans as well as artificial systems can be reporting as well as partially self-transparent and turn out to be quite similar in this respect. This might suggest that the charge of anthropomorphism should be dropped. Or is there another difference, between artificial systems (at least current ones) and humans, in terms of decision-making?

1.4 Discursive Decision-Making

Let's recall the person who decided to go to the kitchen to get a coffee and compare it with the surveillance system that decided to send an alarm because there was a suspicious transaction on the account. Let us assume that the surveillance system is partially self-transparent, i.e., that it has internal representations of the factors that contributed to the decision it made. We could ask both the coffee drinker and the surveillance system about their respective decisions. To that extent, the two systems are similar. If we were to ask about the rationale behind the decisions, in one case we might get the following answer: "The house smelled so good of coffee and since I didn't have any before, but I really like to drink coffee I decided to go to the kitchen because I suspected that's where the smell came from." In the other case, the answer might perhaps be as follows: "I registered in quick succession a large number of transfers to and from the account of large sums going to or coming from various

countries. Having ruled out a malfunction, as the bank's server did not document any such failure, it could be that these were illegal transactions intended to conceal criminal activity." Both answers would be appropriate in the sense that they make the choice of the particular action from the set of possible actions sufficiently plausible. (Of course, the information could still be wrong because humans are never completely transparent to themselves. A person could therefore be mistaken about their own reasons.) Both responses also indicate that the selected courses of action are rational in that they meet predefined standards (i.e., maximizing well-being, detecting crime). Still, in the case of the coffee drinker, one can imagine that an observer might have doubts about the soundness of the decision. The observer might say something like: "I understand why you decided to go to the kitchen, but it would have been better to stay at your desk and to continue to work to finish that book chapter you agreed to write." It is hard to imagine such a reply in the case of the surveillance system. What does this tell us about the decision-making of the two different agents?

In a famous passage of his *Empiricism and the Philosophy of Mind,* Wilfrid Sellars considers what it means to conceive of a mental episode as knowing:

> The essential point is that in characterizing an episode or a state as that of knowing, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says. (Sellars 1997, § 36)

The point of comparing the coffee drinker and the monitoring system is that we can replace the term "knowing" with "decision-making" in the quote from Sellars. If a reporting self-transparent agent (human or artificial) provides information about which factors influenced a decision, then this agent only provides an empirical description in the sense of Sellars. The agent discloses what factors were computed and how this ultimately led to the selection of a particular action rather than another. However, humans can also provide *reasons* for a decision. If they do, they are no longer just giving an empirical description. Rather, they place the decision in the space of reasons, i.e., they enter into a *discourse* about it. They explore not only whether a decision was good or bad according to a predefined standard, but they examine whether an agent may have made a *mistake.* In other words, decisions then appear under the normative distinction of right and wrong. In this, such agents differ from those considered so far which suggests a final distinction, namely that between *discursive* and *non-discursive* agent.[5]

The space of reasons is a normative space, and it is a socially structured space (Brandom 1994, 1995). Making a move in this space, such as making a discursive decision, means taking on commitments and assigning entitlements to others. This includes, above all, accepting the obligation to give reasons when someone else doubts the legitimacy of the move. The entitlements that others receive are essentially

[5] It should be noted that the analogy to Sellars's notion of knowing is imperfect in at least one respect, insofar as it might suggest that the process of choosing an action, if it is not discursive, is not a form of decision-making at all, just as a particular description of a mental state does not describe this state as knowing. Here, however, I assume that such a process can be taken as a form of decision-making, if only as non-discursive decision-making. Unlike Sellars, I allow for qualifications within the concept. I am grateful to Tommaso Bruni for pointing this out to me.

that they may use the prior move as a starting point for their own future moves. These normative interrelations indicate that beings moving in the space of reasons must be able to take responsibility. Discursive agents are therefore necessarily *responsible* agents. This brings back a notion that had already briefly appeared at the beginning, and which has traditionally been closely associated with the concept of decision-making, especially in philosophy. Now it becomes clear that there is, indeed, an important dimension of decision-making associated with the notion of responsibility.

The exercise of responsibility presupposes a community of agents linked by a normative practice. There is an argument to be made that such a practice is tied to a shared life form. It is true that the normativity in the space of reasons, as conceived by Sellars and Brandom, is initially of a logical or inferential nature. However, it is reasonable to think that it must be anchored in elementary structures of a life form. Wittgenstein drew attention to this with his example of the talking lion, which we do not understand (Wittgenstein 1984, § 568). Floridi has a point when he argues against Wittgenstein that we share a lot with the lion (hunger, fear, pain, and pleasure), so that understanding might be possible if the lion started to talk to us (Floridi 2013, p. 298). However, the basic idea remains valid: there must be substantial common ground for different agents to form a discursive community. It is difficult—though not impossible, as some science-fiction shows—to imagine that artificial systems become members of our human community. If this is true, then artificial systems cannot be discursive agents at least as long as they are not members of our human community. But for this to happen, they would need to have some properties that are essential to us, such as embodiment and feelings—by the way, properties that some critics believe are necessary for artificial general intelligence (the *locus classicus* is Dreyfus 1992).[6]

For discursive agents, the evaluative standard for choosing an action is always up for discussion as well (and not only the particular decision taken). While some artificial systems can provide information about how they have reached a decision—which is quite impressive, especially if one considers that such systems can make extremely complex decisions for which they process a great number of factors—the standard of evaluation itself is fixed. For this reason, it is important to consider exactly which standard ensures that a particular goal is achieved by a complex system. After all, it could turn out that an artificial system pursues a predefined goal extremely efficiently, but then it turns out that it was not the goal we had initially in mind. Ironically, this decision would still be good or even optimal given the definition from above. This problem has been discussed for some time under the title "alignment problem" (Christian 2020). It is a problem not least because artificial systems do not change their goals or their evaluative standards. Therefore, we need to make sure that the goals we implement in them are appropriate. Or else we would have to build AI in such a way that it can change its goals on its own. However, this approach might be even more problematic. If artificial systems were discursive in

[6] Note that it could be that artificial agents form an entirely separate kind of discursive community that would exist beyond and entirely independent of our human discursive community. This is, however, an extremely speculative idea which I do not want to pursue any further here.

the sense just outlined, then they could rethink their decision-making practice under the impression of good reasons and, if necessary, realign it themselves. Just as the coffee drinker might concede to the observer's criticism that in the future, he will curb his desire for coffee until the book chapter is finished, so too the surveillance system might change its goals by itself under the impression of criticism. However, this would have to be accompanied by the ability to take responsibility for one's own decisions. Even if it should be possible to build artificial systems that can change their goals and standards themselves, we must think carefully about whether we want to do so. It could be that that would lead to profound conflicts with our social practice and, ultimately, even with the future of our life form (Bostrom 2014). In any case, the construction of such artificial agents would be a decision for which the persons in charge would have to take responsibility within the human community.

1.5 An Alternative Approach

In his *The Ethics of Information*, Luciano Floridi proposed a typology of agents too. A brief comparison shows where his approach and the one suggested here differ. First, Floridi introduces the notion of "level of abstraction (LoA)" to allow for different levels of analysis (Floridi 2013, pp. 31–34). On one level of abstraction, according to Floridi an agent is "a system, situated within and a part of an environment, which initiates a transformation, produces an effect, or exerts power on it over time" (Floridi 2013, p. 140). According to this definition, earthquakes are agents, as Floridi points out. However, since this is implausible, he suggests changing the LoA to include three criteria, namely "interactivity", "autonomy", and "adaptability". While interactivity is self-explanatory and, moreover, in line with the definition proposed above, autonomy and adaptivity require explanation. According to Floridi, autonomy simply means that a system can carry out internal transactions independently of its environment. This is also in line with the definition above. The situation is different with adaptivity, by which Floridi means that "the agent's interaction changes (or can change) the transitions rules by which it changes states." (Floridi 2013, p. 141). This was not required in the definition above. Otherwise, the robotic lawnmower, for example, could not have been classified as an agent. In this respect, Floridi places higher demands on a system. According to him, agents are always adaptive systems. Floridi goes on to introduce another subcategory, and that is that of a "moral agent" (Floridi 2013, pp. 146–148). By this he means agents capable of "morally qualifiable action". Note that his does not entail responsibility. On the contrary, Floridi wants to escape the following dichotomy:

- moral agency, therefore responsibility, therefore prescriptive action, versus
- if there is no responsibility then there is no moral agency, but without the latter there is no need for any prescriptive action. (Floridi 2013, p. 159).

In other words, he worries that agents can perform actions that are morally relevant (i.e., it can cause moral good or evil), but that no responsibility can be attributed. One

can share this concern without having to accept Floridi's typology. On the contrary, one can adopt a more liberal concept of agents that includes non-learning systems on the one hand and assume a more nuanced typology that comprises discursive agents which are responsible for the consequences of action on the other hand. If one takes this approach, then the question of attribution of responsibility can also be answered: it is always discursive agents who are responsible for consequences of action (if it is not a matter of accidents or natural events). Moreover, this approach has the advantage of linking types of agents directly to types of decision-making, i.e., to specific capabilities, rather than to consequences in the environment. This is less revisionist and preserves common language and established normative practices. Floridi's approach, however, makes clear that agent interaction raises important questions.

1.6 Decision-Making Involving Agents

It is important to realize that agents can be present in the environment of other agents, so that different forms of interaction may occur. In particular, other agents may play a role in the decision-making of agents. They may, in the terminology suggested above, influence the set of observations O and the set of possible actions A in a variety of ways. One can therefore consider whether general rules can be derived from the typology developed earlier. For example, should we deal with reporting agents in a fundamentally different way than with non-reporting agents? Or should we always use non-transparent agents differently than transparent agents?

To begin with a fairly simple observation, the average quality of an agent's decisions does not seem to depend on the type of agent. The evidence suggests that robotic lawnmowers perform very well, i.e., make mostly good or even optimal decisions when measured against the specified standards (getting the mowing done, doing it quickly and accurately, not hurting anyone, etc.). This, of course, is mainly because they must base their decisions on few observations (three sensors), that processing these observations is rather easy, and that the set of possible actions is very limited (mowing in a certain direction, stopping). There is, therefore, little room for error. With complex systems such as the surveillance system, things are different. It is the complexity of the environment, or the set of observations that goes into a decision, as well as the possible courses of action that affect the quality of an agent's decisions. The uncertainties associated with the given parameters is equally important. In this regard, the typology of agents and decision-making is not decisive.

The picture is different when it comes to agents who can make decisions with serious consequences. It can be helpful to have a reporting agent for this may allow to intervene directly in the decision-making process to avoid unfavorable effects. Likewise, it can be helpful to have partially transparent or fully transparent agents for understanding unwanted consequences in retrospect. Efforts to create explainable AI take this fact into account. Especially in sensitive areas, we should not rely on non-transparent agents.

There is something else to consider: According to an established distinction, non-discursive agents are means or tools.[7] They are used by discursive agents as a means to an end. They can never take the position of a discursive agent within ends-means relationships because they cannot bear responsibility. For discursive agents, this means that they need to look very carefully at the types of agents they use or rely on in making their own decisions. Intransparency can become a problem here because it may make it more difficult to assign responsibility. Some have even argued that a "responsibility gap" is opening up (Matthias 2004), while others have disagreed (e.g., Köhler et al. 2017). In any case, the difference between discursive agents and non-discursive agents turns out to be more profound than the other distinctions. This is consistent with the classical view that persons, i.e., discursive agents, must never be used by other persons merely as a means (Kant 2016). This does not imply that the treatment of all other agents is morally neutral or unimportant. Animals, as sentient beings, are widely believed to have a moral status such that we may not treat them cruelly. What would have to be fulfilled for artificial agents to also acquire such a moral status is currently the subject of intense debate. For the time being, however, we as discursive agents need to think first and foremost about the requirements we place on agents we use as tools for specific tasks. The typology developed earlier may be particularly relevant in this regard.

1.7 Outlook

The considerations I have presented in this chapter show that the term "decision making" can be understood quite broadly and can encompass quite different types of processes. The unifying element is that it is always a selection of an action from a set of alternatives, based on observations of the environment. Systems that are able to do this are called agents. However, a broad understanding of terms can easily lead to misunderstandings, especially when different scientific disciplines work together. The proposed typology can help to avoid such misunderstandings by offering conceptual differentiations. If there are any uncertainties, it is always worth asking what kind of decision-making one actually has in mind. The most profound distinction is between non-discursive and discursive agents. So far, only humans fall into the latter category. Since discursivity and responsibility are necessarily linked, so far only humans can be assigned responsibility for decisions. We must, therefore, carefully consider how we deal with other types of agents and for what and how we employ them in our own decision-making.

[7] According to this distinction, non-human animals are also means or tools, since they cannot be responsible agents either. Even if we grant moral status to non-human animals, we still do not treat them as full members of the discursive community in our moral and legal practice. For example, we do not hold them responsible for damage they cause. In the case of pets or livestock, it is rather the owners who must assume this responsibility.

Acknowledgements I would like to thank Ulrich Ettinger and Carsten Murawski as well as the members of my working group "Neuroethics and Ethics of AI" for providing invaluable comments on an earlier version of this chapter.

References

Anscombe GEM (2000) Intention. Harvard University Press, Cambridge
Bostrom N (2014) Superintelligence: paths, strategies and dangers. Oxford University Press, Oxford
Brandom R (1994) Making it explicit. Reasoning, representing, and discursive commitment. Harvard University Press, Cambridge
Brandom R (1995) Knowledge and the social articulation of the space of reasons. Philos Phenomenol Res 55(4):895–908. https://doi.org/10.2307/2108339
Christian B (2020) The alignment problem. Machine learning and human values. Norton, New York
Davidson D (2006) The essential Davidson. Oxford University Press, Oxford
Dreyfus HL (1992) What computers still can't do. A critique of artificial reason. MIT Press, Cambridge
Flavell JH (1979) Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. Am Psychol 34(10):906–911
Floridi L (2013) The ethics of information. Oxford University Press, Oxford
Heidbrink L, Langbehn C, Loh J (eds) (2017) Handbuch Verantwortung. Springer, Wiesbaden
Holzinger A, Saranti A, Molnar C, Biecek P, Samek W (2022) Explainable AI methods: a brief overview. In: Holzinger A, Goebel R, Fong R, Moon T, Müller K-R, Samek W (eds) xxAI—beyond explainable AI. xxAI 2020. Lecture notes in computer science (LNCS, vol 13200). Springer, Cham, p 13–38. https://doi.org/10.1007/978-3-031-04083-2_2
Kant I (2016) Grundlegung zur Metaphysik der Sitten. In: Kraft B, Schönecker D (eds) Meiner, Hamburg
Kochenderfer MJ, Wheeler TA, Wray KH (2022) Algorithms for decision-making. MIT Press, Cambridge
Köhler S, Roughley N, Sauer H (2017) Technologically blurred accountability? Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In: Ulbert C, Finkenbusch P, Sondermann E, Debiel T (eds) Moral agency and the politics of responsibility. Routledge, London, pp 51–67
Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6:175–183. https://doi.org/10.1007/s10676-004-3422-1
Müller VC (2023) Ethics of artificial intelligence and robotics. In: Zalta EN (ed) The stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/
Russell SJ, Norvig P (2020) Artificial intelligence: a modern approach, 4th edn. Pearson, Hoboken, NJ
Salhi S, Boylan J (eds) (2022) The Palgrave handbook of operations research. Palgrave Macmillan, Cham
Sellars W (1997) Empiricism and the philosophy of mind. Harvard University Press, Cambridge
Setiya K (2022) Intention. In: Zalta EN (ed) The Stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/win2022/entries/intention/
Steele K, Stefánsson HO (2020) Decision theory. In: Zalta EN (ed) The Stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/win2020/entries/decision-theory/.
Wittgenstein L (1984) Philosophische Untersuchungen (Werkausgabe, Bd. 1). Suhrkamp: Frankfurt am Main