Philosophy & Technology (2021) 34:1731–1750
https://doi.org/10.1007/s13347-021-00483-2

RESEARCH ARTICLE

Human Goals Are Constitutive of Agency in Artificial Intelligence (AI)

Elena Popa[1]

Received: 15 March 2021 / Accepted: 23 September 2021 / Published online: 23 October 2021

Abstract
The question whether AI systems have agency is gaining increasing importance in discussions of responsibility for AI behavior. This paper argues that an approach to artificial agency needs to be teleological, and consider the role of human goals in particular if it is to adequately address the issue of responsibility. I will defend the view that while AI systems can be viewed as autonomous in the sense of identifying or pursuing goals, they rely on human goals and other values incorporated into their design, and are, as such, dependent on human agents. As a consequence, AI systems cannot be held morally responsible, and responsibility attributions should take into account normative and social aspects involved in the design and deployment of the said AI. My argument falls in line with approaches critical of attributing moral agency to artificial agents, but draws from the philosophy of action, highlighting further philosophical underpinnings of current debates on artificial agency.

Keywords Artificial agency · Teleology · Responsibility · Autonomy · AI

1 Introduction

Discussions of responsibility in AI rely on the concepts of autonomy and agency applied to artificial systems. This paper argues for the incorporation of human goals alongside a broader framework of values and norms in the analysis of autonomy and agency in AI. On the view proposed here, including human goals in the analysis of artificial agency is necessary for making correct responsibility attributions in relation to AI systems. While current discussions on the topic take place within ethics, I construct my analysis in relation to the philosophy of action. Drawing from debates regarding whether agency is to be defined in a causal versus teleological way, I will argue that the quest for reduction and the elimination of teleological notions from science has led to the neglect of goals, intentions, and other human-specific factors

✉ Elena Popa
elena.popa@auw.edu.bd; elena.popa@protonmail.com

[1] Asian University for Women, 20 M.M. Ali Road, Chittagong, Bangladesh

in analyses of artificial agency. In contrast with these tendencies, I argue that an approach to responsibility in AI can achieve a satisfactory level of complexity only if the account of artificial agency is teleological. My analysis will discuss agency as a feature of current AI systems that possess higher degrees of autonomy.

The main steps in my argument will be defended in distinct sections as follows. I will first argue that concepts of autonomy in artificial systems are reliant on human goals (Sect. 2). This is important because often autonomy is taken to be a requirement for artificial agency without being consistently defined in the engineering and AI literature. Secondly, I argue that a concept of artificial agency requires a background of norms, values, and goals which are also human dependent; otherwise, it fails to capture social aspects of the design and functioning of artificial systems (Sect. 3). For this point, I rely on contributions from the Computers-in-Society research program (as in Johnson & Miller, 2008). Thirdly, I argue that if the social aspects are left out, responsibility attributions would be undermined. I initially formulate this as a critique against approaches that leave out the human side of AI systems (Sect. 4), then explain that the responsibility gap regarding AI behavior is an instance of this problem, and take a critical stance on it (Sect. 5). I conclude that human goals underlying autonomy and artificial agency should be included in the analysis of AI behavior in order to make accurate responsibility attributions, and end by exploring further connections with human agency (Sect. 6).

To illustrate the proposed view with an example to which I will come back, let us take an algorithm that uses machine learning for monitoring mental health and providing relevant information to the patient.[1] Were the software to encounter ambiguous or uncertain information, it may decide to overestimate the user's risk of self-harm and advise accordingly, thus erring on the side of caution. The problem of responsibility here involves specific ways in which the algorithm made the decision, but also questions of design—such as how to behave under uncertainty—and broader questions about the usage and approval of such algorithms—whether the patient is also under psychiatric supervision, talks to a therapist regularly etc. From the framework I propose, in assessing responsibility, it is not sufficient to refer to the immediate tasks of monitoring and providing information, and the goals set by design need to be examined as well. The decision to err on the side of overestimation would make sense if the overall goal is to promote help seeking, but cannot be explained solely in terms of the workings of the software. While the machine behavior contributes to this goal, the designers are the ones who set both the goal and the intended use of the algorithm. As such, responsibility for cases when the algorithm fails can be traced to the designers and programmers, but possibly also to medical professionals that deemed the case of the patient adequate for the usage of the specific software.

[1] See Burr and Morley (2020) for a discussion of digital health technologies along the lines of the software mentioned.

2 Behavioral Autonomy in AI

While autonomy is a leading topic in AI research, its meaning has not been consistent across research programs. Since autonomy is often identified as a condition for agency, I will focus on aspects of autonomy relevant for the question of artificial agency. More specifically, one would not attribute agency to a robot operating under remote control even if it may perform certain sequences of operations by itself. Higher degrees of autonomy, however, may enable the system to exercise a kind of agency. I will argue that even in the case of more advanced levels of autonomy the behavior of the AI system is modeled according to goals set by the researchers, which can be traced within relevant research programs. Thus, if agency is defined through autonomy, human goals will be part of the picture.

Before discussing specific definitions, it is important to distinguish two types of autonomy: constitutive and behavioral (Froese et al., 2007). Constitutive autonomy is connected to how the system operates and self-maintenance, by analogy to biological systems (see Maturana & Varela, 1980). Behavioral autonomy refers to how the system functions in relation to its environment and programming, and as such it is directly relevant to my purposes here. Henceforth, I will use the term "autonomy" to refer to behavioral autonomy only. Further clarification regarding what counts as autonomous behavior is needed before moving forward. One may consider behavior by an AI system that looks *as if* it is autonomous sufficient for autonomy, or one may point out that autonomy requires components such as consciousness and intentionality, and as such AI systems are not autonomous. My discussion in the following sections will contrast views of the former kind (going back to Dennett's, 1987 intentional stance) with the latter (e.g., Bryson & Kime, 2011; Gunkel, 2012; Johnson & Verdicchio, 2017, 2018). Due to the complexity of the debate, I will not go through all of the relevant aspects, but focus specifically on goals. Insofar as a system considered autonomous is employing higher level abilities to pursue goals, a question arises regarding the origin of the goals.

The discussion by Froese et al. (2007) takes place in the context of Artificial Life, and insofar as different senses of autonomy apply to artificial systems more broadly, they are relevant for my discussion. I will discuss two senses: (1) functioning without human intervention—used in the context of engineering; (2) acting to achieve goals or even set goals in the context of a certain environment, which involves teleological notions (Froese et al., 2007: 456–457).[2] Machine learning can enable AI systems to be autonomous in sense (2), using various algorithms to achieve a goal under different conditions.[3] By contrast, AI can also perform a predetermined sequence of operations without human intervention other than the initial programming in sense (1). To use the example in the beginning, a mental health monitoring algorithm is autonomous in sense (2) if it can adjust to the input received from the user to provide relevant diagnosis and information. This would run in contrast with, say, a chat bot that has a limited set of lines to use in reply to specific input.

[2] For sense (1), see Brooks (1991), for sense (2) Beer (1995); Nolfi and Floreano (2000).

[3] See Tan and Lim (2018) for a review of current advances in AI, including machine learning.

Ezenkwu and Starkey (2019) distinguish between low-level and high-level attributes of artificial autonomy. The former include perception, actuation, learning, context-awareness, and decision-making, while the latter comprise domain-independence, self-motivation, self-recovery, and self-identification of goals. The difference between the two is spelled out as follows: "a truly autonomous agent can develop skills to enable it to succeed in such environments without giving it the ontological knowledge of the environment a priori" (Ezenkwu & Starkey, 2019: 335). This maps onto the distinction above, and high-level attributes fit in with sense (2), enabling the entity to operate in different environments without having all the relevant knowledge coded beforehand. Regarding the extent to which such AI systems are presently in operation, Ezenkwu and Starkey's review shows that most approaches currently available do not demonstrate self-identification of goals.

I will now explain how these attributes would work in relation with the example above through agent architectures. According to Bryson, an agent architecture is a collection of knowledge and methods for designing artificial intelligence (2000: 165). Agent architectures require a modular structure, competences to perform complex tasks taking into account both features of the situation and agent priorities, and means of reacting to environmental changes in a timely manner (Bryson 2000: 185). Going back to the example, the mental health monitoring algorithm can identify warning signs and suggest helplines without having all of the possible scenarios provided a priori. This process involves identifying a certain input as a warning sign, which would trigger an action plan under the pursuit of a goal—in this case, advising the user to seek help, which falls under the goal of helping the user improve their well-being. The algorithm gains information about the mental state of the user and recognizes certain inputs as warning signs without having received this information a priori. This would count as domain-independence according to the definition by Ezenkwu and Starkey: "domain-independent agents do not require the ontological knowledge of their environments at design time to succeed in the environment" (2019: 339). The process also involves a decision by the machine and the appropriate response. This differs from an algorithm where specific input is a priori connected with specific responses, and assessing inadequate responses would come down to human decisions on what to code into the behavior, albeit in more complex circumstances. Note that neither of these cases involves direct human intervention on the behavior of the machine, but the responsibility attribution is more complex in the context of the former.

Having placed the focus on autonomy in relation to goals, I will now present the first step in my argument. In doing so, I rely on an earlier discussion of goals in relation to autonomy in Artificial Life, whose broad features can be applied to the more encompassing category of AI systems. While, as discussed above, high levels of autonomy involve goals, the discussion in the literature on engineering or AI does not specify whose goals they are. Analyzing how the significance of synthesizing life forms is laid out in the literature, Popa (2020) argues that these entities ultimately pursue human interests, environmental or medical uses serving as examples (2020: 593). This further leads to a substantial difference between biological organisms and artificial systems: the former can be attributed goals such as survival and reproduction, while the latter are designed for purposes specified according to a

research program. Bringing this together with the concept of autonomy as identification and pursuit of goals, it follows that AI systems can work on their own in the pursuit of a goal, but that goal will be set by humans through a research program. While the description of Froese et al. (2007) also includes the possibility of artificial systems setting their own goals, they would work as sub-goals in relation to particular challenges to help achieve the higher level goals set through the design. Still, as autonomous AI systems may operate independently from direct human control, the study of such machines would involve more complexity than viewing them simply as tools or other artifacts.[4] Illustrating this point through the earlier example, the improvement of mental health, or providing vulnerable individuals the opportunity to seek help when in distress are human goals that the algorithm would serve. Unlike in the goals of biological entities, which can be explained through the interaction between nature and nurture, this goal originates in the programming and training, and is set by designers.

3 Is There Agency in AI?

Having discussed autonomy in AI with particular focus on high-level abilities, the next question is to what extent can autonomous AI be said to have agency. For this, I will look into contributions to the philosophy of action, and their connection to the case of AI. I will argue that accounts of artificial agency that do not include norms or goals are unable to account for the social aspects of AI. I will defend the minimal agency account, which does not have this shortcoming. As the interaction between research in the philosophy of action and artificial agency has been limited, my argument will also aim to stress further connections in this sense.

Philosophical discussions of agency rest largely on work by Anscombe (1957) and Davidson (1963). As Schlosser points out, it is important to distinguish between the standard conception of agency, and the standard theory of action (often called the causal theory of action): "First, the notion of intentional action is more fundamental than the notion of action (…) action is to be explained in terms of the intentionality of intentional action. Second, there is a close connection between intentional action and acting for a reason" (2019: 2). The causal theory of action starts from this conception and holds that "the agent performs an action only if an appropriate internal state of the agent causes a particular result in a certain way" (Davis 2010: 32). Internal states comprise intentions, desires, beliefs etc. For example, my action of taking coffee from the cupboard can be caused by my desire for a cup of coffee and the belief that there is coffee available in the cupboard.

The debates over the shortcomings of the causal theory and other components of agency are beyond the purposes of this paper. However, one central question

[4] This is a view held by Bryson and Kime (2011), in whose account the responsibility for AI systems should rest with the developer. While moral responsibility is not my main focus here, my view looks at human goals more broadly (including, for instance, the possibility of deliberation, or the involvement of institutions).

concerns whether representation is necessary for agency. If one were to answer this in the positive, then, among others, animals and artificial entities would not possess agency.[5] Using the example above, the desire for a cup of coffee involves a mental representation (i.e., I can represent the desired state of drinking a cup of coffee, which will initiate the process of me getting coffee from the cupboard and making coffee). In the absence of mental representations, there is nothing to cause the action (i.e., no desire, or belief). There are views, however, that do away with the requirement of mental representation; following Schlosser (2019) I will refer to them as instrumentalist. Instrumentalism in this sense was introduced by Dennett (1987, 1988). Dennett holds that the intentional stance helps explain and predict behavior, and that "any system whose performance can be thus predicted and explained is an intentional system, whatever its innards" (1988: 495). This goes against the emphasis on inner mental states by the standard theory, and, importantly for my point here, would attribute agency to AI systems.

With regard to artificial agency, I am going to discuss two relevant views, both falling under the instrumentalist viewpoint of not requiring mental representations for agency. An overarching view drawing from Dennett's intentional stance is discussed in moral context by Johansson (2010). Behdadi and Munthe (2020) describe this view through an "as if" structure: briefly put, a machine is said to possess agency if it behaves *as if* it were an agent regardless of its ability to use mental representations.

One such view by Floridi and Sanders (2004) employs the "levels of abstraction" method to describe artificial agency. Levels of abstraction are described as collections "of observables each with a well-defined set of possible values or outcomes" (354). Furthermore "Each LoA [level of abstraction] makes possible an analysis of the system, the result of which is called a model of the system. Evidently an entity may be described at a range of LoAs and so can have a range of models" (354). In further work, Floridi (2008) describes the levels of abstraction method in an epistemological way, namely as levels of interpretation of a system. The relation between levels of abstraction is hierarchical, as described by Floridi under the term "gradient of abstraction." Briefly put, a gradient of abstraction can consist of multiple levels of abstraction according to the perspective taken. While Floridi uses this framework to discuss multiple examples, I will only focus on agency here.

Using this method to analyze agency, the point is that a higher level of abstraction than that applied to human adults (as standard theories would have it) would reveal features of agency that hold across humans and artificial agents. The three features are described as follows:

- *Interactivity* concerns the relation between the artificial system and its environment, namely how it can both influence and be influenced by specific circumstances.

[5] Another example would be certain human actions as well, like skilled action (see Clark 2010). I will not discuss this here.

- *Autonomy* refers to the agent's ability of operating on its own, without external influence; this should be distinguished from its use in relation to AI systems in the previous section.
- *Adaptability* amounts to the agent learning from previous interactions, thus balancing the first two conditions (Floridi & Sanders, 2004: 357–358).

It should be noted that this approach to artificial agency does not include goals, and makes no reference to a normative framework. This is due to the employment of a high level of abstraction that would do away with human-specific conditions. In criticizing this view, Johnson (2006) points out that even though moral agents in general may exhibit these features, one cannot ignore that artificial systems were designed by humans and perform their functions within a social setting. As such, choosing this specific level of abstraction appears to miss one central aspect of AI behavior. Johnson argues, contra Floridi and Sanders, that AI agents are not moral agents, but should instead count as moral entities. Further criticism by Grodzinsky et al. (2008) targets the way in which the designer's intentions, while not present on higher levels of abstraction, still constrain the behavior of the AI system, and are thus important from an ethical point of view. While a full account of the moral aspects of this debate is beyond the purposes of my paper, my interest lies in the question of attributing agency to AI systems. In this sense, my view falls in line with Johnson and Grodzinsky et al. with regard to the importance of design and relevant social aspects.[6] However, I do not think that this is a knock down argument against attributing agency to AI systems altogether. Rather, I take the criticism to be an indication that focusing on the particular level of abstraction and features that Floridi and Sanders do paints an incomplete picture of agency in AI. To put it another way, my position is that it is possible to ascribe a sense of agency to AI systems, though it would not count as *moral* agency. As I will argue below, moral considerations belong to the human realm.

Another approach to artificial agency that does not depend on mental representations is that of "minimal agency" by Barandiaran et al. (2009). On this view, the conditions for minimal agency are "(a) there is a system as a distinguishable entity that is different from its environment, (b) this system is doing something by itself in that environment, and (c) it does so according to a certain goal or norm" (Barandiaran et al., 2009: 369). As opposed to the previously discussed view, this account does take norms into consideration (condition c), and it falls in line with previous accounts connecting goal directedness to normative considerations, such as Bedau (1992). Furthermore, it can answer the earlier criticism regarding the relation between an AI system and society.

In the light of the discussion in the previous section, the presence of norms brings about a similar question regarding who sets the norms. Contrasting minimal agency in animals with that in AI systems will be helpful here: while in the case of

[6] For other debates on moral agency and artifacts more broadly, see Illies and Meijers (2009) and Peterson and Spahn (2011). For an expansion of Johnson's (2006) critique, also see Johnson and Miller (2008).

animals having agency one could count survival and reproduction as goals in line with broader evolutionary considerations, even if the animal cannot represent them, for AI systems the goals are connected to problems that people are aiming to solve. Assuming it would be possible for synthetic life forms to reproduce due to future technological advances, the existence of offspring, as well as the behaviors of such organisms, would be explained through the overall goals of the research project. For instance, scientists may decide to design artificial life forms that are able to reproduce if it would be more resource efficient to do so as opposed to creating every generation of synthetic life forms de novo. Thus, if minimal agency involves acting according to a set of norms, in the case of AI systems those goals are set by people, and that is enacted through the research program determining the design of the system in question. To put this another way, an AI system may behave *as if* it is following a goal, and that may suffice to attribute it a minimal sense of agency, but the goal does not belong to the AI itself, but to the people involved in the process of research, design, and approval.

The presence of normative considerations also helps connect discussions of autonomy and agency in AI systems, bringing the perspectives from engineering and philosophy together. On the view presented here, both artificial moral agency and minimal agency would go beyond the machine simply doing something by itself as in sense (1) of autonomy above, or possessing a priori knowledge about potential problems arising in its environment. Thus, the question of agency only arises for those AI systems possessing autonomy in the sense of high-level abilities. I will discuss this in relation to the mental health monitoring software above. It can be investigated, for instance, whether such algorithm has high-level abilities such as domain-independence. This would entail that the algorithm monitors the mental state of the user on the basis of the input without having domain-specific information provided at design stage. The agent architecture would enable the identification of certain triggers through the interaction of the software with the user and the pursuit of an appropriate action plan by the software. This, however, raises the question whether such complex patterns can be enabled by machine learning alone, on the basis of trial and error. Several challenges arise, some ethical, regarding what interactions are allowed between the software and the patient, others epistemic, regarding whether domain-general information is sufficient for assessing someone's mental state.[7] These challenges may be addressed by constraining the behavior of the machine, and/or adding background information, but this casts doubt on whether the machine is truly domain-independent after all. Still, even if current AI systems are not fully domain-independent, the properties enabling courses of action such as those described above are more complex than the machine performing a task by itself. As such, these AI systems would be better suited for ascriptions of minimal agency. Connecting this to goals, one aspect worth noting is that goals should align with broader social norms that enable the deployment of the software. For instance, the goal of improving mental health presupposes that symptoms of poor mental

[7] I am grateful to an anonymous referee for bringing this problem to my attention.

health are complaints to be monitored and treated, and should not be viewed as actions or character traits for which the user is to be blamed.

In sum, if there is agency in AI systems, then it can be captured by a view that incorporates its relation to norms set by design such as the one by Barandiaran et al. (2009). In accordance with instrumentalism, this would not require machines to operate with mental representations. Still, it would involve human goals, and as such a degree of dependence on human agency. Due to the functioning of such systems, this dependence is not direct, leaving the AI with the possibility of learning from interacting with the environment and setting sub-goals.[8] Before moving on, I would like to make a broader point about instrumentalism. Going back to Dennett's view and other approaches attributing agency to systems that behave as if they were agents, it is important to distinguish minimal agency (or agency understood through a high level of abstraction) from more advanced kinds of agency. The view defended here does so in terms of goals as opposed to representations: only in more advanced forms of agency (such as those of human adults) an agent acts in pursuit of the agent's own goals. In the case of AI systems, the goals are set by design, and overlap with human interests. In the case of animals, the goals can be specified as part of a broader biological framework (survival and reproduction). This point, however, can only be made from a teleological approach at least as far as agency in AI goes. Thus, insofar as goals or other connections to human interests are absent from instrumentalist accounts, a problem arises regarding conflating minimal agency with more advanced forms of agency. This is an important issue especially in connection to responsibility, which I will discuss in the following sections.

4 The Case for a Teleological Approach to AI Behavior

In the previous sections, I have argued that higher forms of machine autonomy involve human goals and that attributing agency in a minimal sense to AI systems also requires reference to human goals as part of a broader normative framework. Separating the behavior of AI systems from such goals leaves out a central aspect of their functioning. This section will take a critical stance on views that leave out goals or other human components of artificial agency, tracing and subsequently rejecting assumptions underlying such views. This will prepare the ground for an argument in relation to responsibility in Sect. 5.

Views on artificial agency leaving out goals are akin to Floridi and Sanders (2004) discussed above, with analogous claims being made in different areas of AI research. As mentioned above, the levels of abstraction method enables one to look at components of artificial agency without considering human-specific features such as intentions, goals, or norms. Part of the motivation for this is to avoid an "anthropocentric" perspective (i.e., analyzing agency in a way that leaves out non-human entities). Johnson (2006) argued that this level of abstraction is irrelevant when

[8] This also helps set apart AI systems from other artifacts; see van de Poel (2020a: 399–400) for a comparison.

discussing moral agency, since human factors matter in relation to this particular context. This critique is expanded in Johnson and Miller (2008), where the omission of relevant social aspects is explained through the research program of "Computational Modelers" attributed to Floridi and Sanders, according to which the analogy between human and machine behavior plays a central role (pp. 126–127). My view is in agreement with Johnson and Miller regarding the importance of social factors, but I attempt a different explanation of the omission of such factors by the Computational Modelers research program, namely, in relation to deeper philosophical assumptions. Before moving forward, I should point out that there are overlapping points between my critique and that of Johnson and Miller, such as the reference to reduction: "while these concepts and their relationships can be modeled and represented, to say that they can be reduced to a different level of abstraction seems at least to beg the question, if not to be entirely misguided" (Johnson & Miller, 2008: 128). However, Johnson and Miller focus on moral aspects here, while I will mainly refer to conceptual issues in the philosophy of action. As such, the view defended here can also be seen as supplying a conceptual background for the "Computers-in-Society" program defended by Johnson and Miller. I would now like to point out similarities between the levels of abstraction method and tenets of particular versions of naturalism, though without attributing this particular view to Floridi and Sanders.

The defense of physicalism throughout twentieth century philosophy relied on various versions of the claim that higher level phenomena (the mental, or the social) can be reduced to lower level physical entities. In a strong sense, this would involve the causal closure principle: "if mental and other special causes are to produce physical effects, they must themselves be physically constituted" (Papineau, 2020: 1.3). However, weaker senses were also brought forward, referring to the laws of physics, for instance. These debates took place especially in the context of the philosophy of mind, particularly in relation to mental causation and the mind–body problem, further influencing the philosophy of action.[9] While the focus has shifted to different concerns in present debates in the philosophy of action, the remnants of the discussion regarding the reduction of the mental to the physical appear to be present in the discussion of artificial agency. As engaging with various strands of naturalism is beyond my purposes here, I concentrate mainly on the strategy of using lower level constituents (or components singled out on higher levels of abstraction) to fully account for higher level ones. Moving between levels without missing anything would rest on the assumption that the lower level entities would exhaust everything there is about the phenomenon in question (in this case, the social would be explained exclusively in terms of the physical). However, as discussed above, leaving goals or intentions out also omits information relevant to the moral assessment of AI behavior.

While the focus on lower level constituents explains the neglect of goals, I believe that a further assumption comes into place, related to explanatory practices: that of explaining away purposes and other teleological notions. This has been a hallmark

[9] See D'Oro and Sandis (2013) for a historical overview.

of modern science, and current debates in the philosophy of biology show different ways of conducting functional analysis without referring to design or purposes. While this is legitimate in most scientific areas, there are reasons to doubt the applicability of this principle to the behavior of AI systems. AI behavior is following patterns that are shaped by an initial design, meant to fulfill particular human goals. While in the case of biological organisms and adaptation, a trait looks *as if* it had been designed for a purpose but the process is explained through natural selection or other biological processes, in the case of AI systems this is not merely metaphorical. Various algorithms are meant to guide the AI through fulfilling certain tasks which are specified already at the design stage, as in the case of the mental health monitoring software mentioned above. As in the case of evolution in biological entities, some behaviors may be more adaptive than others, and as such retained, but, importantly, the process is guided by an initial design which is human dependent. Thus, human agency plays a central role in what behaviors an AI system will come to exhibit. Further connections between my view and approaches to human agency will be explored in Sect. 6.

I will refer to the two assumptions as the reduction to lower level entities, and explaining away goals, respectively. These assumptions can be understood as working in a way similar to presuppositions in Collingwood's (2001) sense: enabling particular questions to be asked in relation to AI behavior and determining what factors are relevant, while not being part of the first-order investigation of AI behavior.[10] As pointed out above, adhering to these claims has both epistemic and moral shortcomings: by leaving out a significant variable determining how AI systems act, they also conceal aspects relevant to the moral assessment of such systems (for instance, intentions that led to their deployment).

An illustration of this is the neglect of features of human agency that do not match AI behavior: "the tendency to idealize artificial intelligence as independent from human manipulators, combined with the growing ontological entanglement of humans and digital machines, has created an 'anthrobotic' horizon, in which data analytics, statistics and probabilities throw our agential power into question" (de Miranda, 2020: 597). Thus, discussions of AI behavior tend to neglect human goals, although the levels of autonomy discussed earlier have not yet been achieved. This is at least partly due to the particular branch of naturalism sketched above, where physical entities and processes are employed, while overlooking normative or social aspects. While de Miranda is making a point regarding how this leads to a distorted picture of human agency, namely, ignoring different ways of reasoning other than quantitative ones, my aim is to stress the case for incorporating the human perspective into artificial agency.

Before moving on, I should note that the view I propose is not completely incompatible with naturalism from a broader philosophy of science perspective. One example of naturalism not seeking to reduce or eliminate social aspects is Kitcher's

[10] Though I should note that they would not count as absolute presuppositions, which in Collingwood's view do not have truth values. Rather, I take the falsity of these assumptions to lead to a worse account of AI behavior than approaches relying on different assumptions.

(2011a) "ideal conversation" framework that would involve the deliberation of socially relevant issues while taking into consideration the interests of all the participants involved. Drawing from Kitcher's (2011b) view on ethics, this approach would explain ethics as a means of solving social problems, and its usage goes back to early human societies. While my critique goes against assumptions underlying particular naturalist views in relation to reduction or explaining away intentions, a discussion of agency in AI and the highlight of social aspects to AI use can take place from a broadly naturalist framework that does not involve reducing the social or the ethical to lower level entities.

5 Artificial Agency and Responsibility

This section defends the claim that if human goals are absent from the analysis of artificial agency, then responsibility attributions for AI systems are undermined. While moral responsibility is analyzed in terms of several conditions, my discussion will focus only on agency as a necessary condition for responsibility. I will refer to the case of the "responsibility gap" as an illustration of this problem, and seek to overcome it through incorporating goals in the analysis of artificial agency.

The question whether the development of autonomous AI relying on machine learning would enable it to perform actions that are not directly dependent on designers or operators is raised by Matthias (2004). On this view, a responsibility gap is likely to arise as a result of the following steps: the programmer becoming "a 'creator' of 'software organisms,' the exact coding of which she does not know and is unable to check for errors," the behavior of the machine being determined by its operating environment to a larger extent than the initial conditions, a blurring of the distinction between programming, training, and operation, and the impossibility of human control over each operation performed by the machine (Matthias, 2004: 182–183). Subsequent literature responding to this argument has been divided. Against these debates, Tigard (2020) contests the responsibility gap, arguing for a more dynamic understanding of moral responsibility, that would accommodate cases of AI behavior: "the techno-optimists and the techno-pessimists appear to agree upon a fundamental premise, namely that technology poses an especially unique problem for our existing moral and legal practices. For this reason, I propose we step back and ask whether or not there is a responsibility gap in the first place" (2020: 2). As my interest here lies more in how agency connects to moral responsibility than in moral responsibility more broadly, I will not pursue Tigard's argument. My point is, rather, that the responsibility gap arises as a result of the views on artificial agency leaving out human goals, due to the assumptions explained above. This is consistent with Tigard in claiming that the way in which AI systems operate does not necessarily entail a responsibility gap. In my view, this is not about how AI works, but it is rather a conceptual issue, i.e., how people understand AI to work. Again, there appears to be an assumption that AI behavior can be fully understood in terms of keeping track of every operation performed by machines as opposed to considering the broader human goals in virtue of which the machine operates. Thus, leaving out

these higher level aspects also leads to responsibility vanishing alongside with the human factors relevant to the function of the AI system.

By contrast, if human goals and values are taken into account, the responsibility gap need not arise. On the teleological view defended here, judgments about responsibility and artificial agency need to take into consideration *(i) human goals that the machine in question is meant to fulfill*. I will add one more condition to capture other relevant aspects, namely *(ii) a background of other norms and values that the machine should not undermine*. These components involve a normative framework, as discussed above, and I will initially illustrate how this works on two examples from Matthias (2004: 176–177), then on the example introduced in Sect. 1. First, let us suppose a more sophisticated version of the rover Curiosity employs machine learning and develops increasingly effective ways of explore the Mars environment.[11] On this line of argument, were the rover to fall into a hole, one could not trace the occurrence to the designer or operator, but rather to the behavior learned by the machine. Under the proposed teleological view the fall undermines its goal (i.e., to further explore the surface of the planet). Thus, effective design should take such scenarios into consideration, for instance including specific sensors, or using other machines that can assist the vehicle while stuck. While this may not be anticipated in the first instance, upon future design such scenarios are to be considered in order to make the machine better at pursuing the goal. This answer also falls in line with the judgment that the team of designers and engineers would most likely claim responsibility for this kind of situation, while this conclusion would not be available if one were to accept the responsibility gap. Secondly, another example involves a robot that would act as an intelligent toy for children and through a learning algorithm would end up running around the apartment, colliding with the child, and injuring them. Talk about goals here would refer to what the robot was intended to do, say, entertain the child, with running around the apartment being part of that. Still, looking at other norms (and perhaps at how less sophisticated toys are built) entertainment should not come at the price of risking the child's health. Thus, while one may say that the robot acted in accordance with its intended goal, it broke other norms (i.e., safety). Again, such values should be taken into account when designing and testing AI systems. Another way of explaining this example would accommodate further worries about intended or unintended behaviors. Given that it is impossible to anticipate all the consequences of AI behavior, further adjustments to the design should screen for bad unintended consequences, in this case including a safety protocol. The overarching view would thus involve an interaction between human values and the development of technology (van de Poel, 2020a, 2020b). Thirdly, regarding the mental health monitoring software, suppose that after repeated interactions with a user going through mental health problems, the software concludes that the user is feeling better and does not recommend any further action. However, this assessment proves to be premature, and the health of the user deteriorates, leading to another bout of illness. According to the view defended by Matthias, the software would count as responsible since it acted on information it

[11] See Cardoso et al. (2020) on autonomy and the Curiosity rover.

inferred from the patient input. Again, looking at this from the perspective of goals, one can see the failure to account for all the relevant information about the user's health as a failure of the algorithm to perform its task of monitoring mental health, and thus better programming and/or training is required.

Handling these cases through a teleological approach yields into attributing agency to the respective systems, but in a way that does not entail a responsibility gap. This happens because in addition to considering the behavior the machines learn by interaction with their environment, the goals in line with which they were designed are also considered. While I do not disagree with Matthias (2004) in holding that people should not be blamed for things they had no full control of (in this case, involving operations performed by machines), I argue for a more complex picture, taking social and normative factors into consideration that went into design, as opposed to merely looking for particular individuals liable to responsibility. A further issue to take into account is the plurality of explanations available regarding why the machine malfunctioned. To use an example by Collingwood (2001), a highway engineer may hold that an accident took place because of the poor condition of the road, while a police officer may refer to the driver being reckless, without the two causal explanations being incompatible. A similar point can be made when explaining why a machine behaved in a specific way, with responsibility being collectively shared.[12] In the case of harmful effects of AI behavior, the responsibility should be traced to how the design matches specific human goals, as well as to further institutions that would approve such AI, issue safety certificates etc. This falls in line with approaches I briefly review below.

On the view defended by van de Poel (2020a), AI entities are sociotechnological systems working within a framework that comprises technological artifacts, human agents, institutions, artificial agents, and technical norms (2020a: 387). It should be noted that under this system human goals are also acknowledged and can be taken into account when making judgments of responsibility. Relevant to my argument, van de Poel distinguishes human agency from artificial agency as follows: "artificial agents can embody values, while it would be a category mistake to say that humans embody values. (…) Conversely, while humans can embed values in other entities, artificial agents lack that ability as they have no intentionality" (2020a: 399). While this account does not discuss goals in particular, it does point to a normative dimension (values), and it employs a teleological notion, that of intentionality, to distinguish human agents from artificial ones. While compatible with van de Poel's ethical framework, my argument traces the conflation between human and artificial agency to assumptions that leave out higher level entities, as opposed to pointing to a category mistake. Nevertheless, the two explanations are not incompatible: the category mistake may be enabled by hidden philosophical assumptions. Another point to stress is that van de Poel emphasizes the asymmetry between embodying and embedding value by referring to a feature humans have and that AI entities lack. While this overlaps with my points above, I make a further claim about how the

[12] I am grateful to an anonymous referee for this point.

ability to set and pursue goals characteristic to humans is instantiated in AI systems, and how that provides humans with a notable role in responsibility matters.

The issue of values and norms is also present in Gabriel's (2020) defense of the alignment of AI behavior with values. The central normative claim in this context is that AI should align with human goals, and this point has been made in ethical context since the beginnings of AI research (Wiener, 1960), with a more recent formulation in Asilomar and Principles, 2017). While these views discuss alignment with human values from a normative viewpoint, I make a conceptual claim. Agency, insofar as it applies to the most sophisticated AI systems, involves human goals. As such, the question of alignment above would need to investigate both whether a system is serving a certain goal effectively, but also whether the choice of goals is correct (for instance, whether the interests of different groups are taken into account, whether there is transparency etc.). Thus, on the view defended here, a connection between human goals and AI behavior is inevitable, but the question of alignment needs to account for further technical and ethical issues such as those sketched above. In this regard, I would like to contrast my view with a descriptive claim by Johnson and Miller: "computer systems are always tethered (connected) to human beings, though there are a multitude of ways to conceptualize and abstract the workings of these systems" (2008: 132). This claim is ontological, stating the dependency of computer systems on humans. While Johnson and Miller focus on the social processes leading up to the design, deployment, and continued use of artificial systems, my claim is about action. Acknowledging that AI systems can perform actions, and attributing them a sense of agency can be done while still keeping them connected to humans, through goals. While Johnson and Miller deny artificial moral agency, my claim is that artificial agency can be described in a way that would leave the relevant moral factors human dependent (thus not machine dependent).

In recent work, Johnson and Verdicchio (2019) analyze agency as a triadic concept involving designers, users, and the AI system. On this view, AI systems have causal agency, while humans have intentional agency. In relation to responsibility, Johnson and Verdicchio hold that "agency and responsibility should be separated in the sense that agency is triadic while responsibility is always ascribed to humans" (2019: 644). My view is compatible with these considerations: on my view too only humans are responsible because they are the ones setting the goals. Furthermore, by attributing intentional agency to humans, the authors also defend a teleological view, which runs into tension with attempts to define agency as a causal concept. Contributions regarding autonomy and AI systems as sociotechnical entities by Johnson and Verdicchio (2017, 2018) are also relevant for the discussion here. Johnson and Verdicchio point out that while autonomous artifacts do not need human intervention at run time and can thus exhibit unpredictable behaviors; their activity is still embedded into a context where human beings employ the said AI to perform certain tasks (2017: 583). Overlooking the role of human agents in this process amounts to "sociotechnical blindness" (2017: 287). My argument for including human goals in the analysis of artificial agency can contribute to this line of inquiry, as a way of countering sociotechnical blindness.

Stemming from the issue of the responsibility gap previously discussed, a broader objection to approaches that leave out human goals can be articulated, which is more

pressing in the case of AI than in other domains involving agency and responsibility. I will call it the opacity problem. The issue with pointing out which part of the behavior of an AI system went wrong comes down to the code being opaque, and to the probabilistic operations involved in machine learning not being well understood by humans. An approach to agency in AI focusing solely on the sequence of operations performed by the machine would find it impossible to point out what led to the machine performing a certain action. Unlike beliefs, desires, or intentions that can be ascribed to humans, the algorithms followed by machines are not easily singled out. It may be claimed that this is not necessarily a problem at the level of ontology—the action may be caused in a way that is not intelligible to a human subject. However, from the perspective of attributions of responsibility this raises an important issue: one is unable to explain why the machine acted in the specific way it did. Switching to a teleological approach to AI behavior shows that there is more to the picture than particular operations performed by the system, and unlike the effort of spelling out each step taken by the machine as part of the learning process, the goals underlying the design are intelligible provided there is transparency. While what the machine does may function in the fashion of a black box, the input and the output are intelligible, and one can identify whether the behavior matches the design accordingly. This provides a broader case for a teleological approach: it offers sufficient space for picking relevant variables regarding responsibility for AI behavior and corresponding adjustments.

In relation to this, accountability and transparency are discussed by Dignum (2017), who introduces a way of explicitly considering the values of designers and stakeholders when developing AI systems through the concept of value-sensitive design (as in Friedman et al., 2006; van der Hoven, 2005).[13] Within this broader ethical framework, my approach would highlight that the ethical analysis of AI behavior and its presumed agency should consider what goals it was meant to serve and how the specific behavior interacts with other values. An analysis of these values in terms of the interests of different groups and deliberation would link agency in AI to broader social questions about technology design and use. Transparency about goals helps address another potential challenge: how can human goals be known?[14] Ideally, such goals should be specified in the design plans, but it is possible for goals different than those stated to be pursued. In such cases, regulatory bodies should note the discrepancy and hold the relevant parties accountable. While the practical aspects of this require further inquiry, an analogy can be drawn to accountability in different areas, such as whether the behavior of a company is in line with regulations it claims to follow.

[13] Also see Elliott (2017) for a discussion of transparency about values in a general philosophy of science context.

[14] I am grateful to an anonymous referee for raising this point.

6 Artificial and Human Agency

As argued above, a teleological approach to artificial agency would help address issues with responsibility in AI that arise under the assumptions reconstructed in Sect. 4. On this view, artificial agency and human agency are connected by the goals that artificial systems are meant to serve. Thus, one remaining question to be addressed in this section is how human agency and artificial agency interact, and whether that entails any commitments with regard to defining human agency.

Logically, my view would be compatible with both teleological and non-teleological approaches to human agency. One can acknowledge that human goals determine AI behavior, and explain the goals through the standard theory. Still, there is a particular affinity between my view and views critical of the standard theory, particularly teleological approaches. The issue I pointed out above regarding explaining away goals may as well apply to human behavior. Reviewing the history of the debate between causalism and non-causalism, Schumann (2019) points out that the causal theory is currently taken as the standard approach not because it has successfully answered all the relevant challenges, but, among other things, because "it seems to capture the scientific spirit of the age" (2019: 18). Insofar as my earlier discussion has pointed out that reduction and explaining away goals cannot be taken simply as assumptions inherent to any scientific investigation, but may apply differently to various domains, a similar critique can be extended to the broader context of the philosophy of action, undermining the case for defending the causal theory. This would also be in line with views critical of reductive aims in the philosophy of action, such as that of Sehon (2010): "if there is no successful recipe for reduction, then there is no simplicity argument against non-reductionist views" (2005: 127). Thus, I will leave the account of human agency open, while keeping a critical stance on the assumptions regarding reduction and explaining away teleological notions when discussing artificial, and possibly human behavior. While the impact of these assumptions on the understanding of human agency can be traced to research practices in the social sciences, in the case of AI systems there are direct implications for responsibility and policy.

A further correspondence can be drawn to a different set of teleological approaches to agency, and a potential objection raised. Approaches to agency inspired by Wittgenstein would hold that since agency is acting for reasons and since behaviors indicative of goals or intentions are found in humans but not in AI systems, they would count as mere artifacts (Wittgenstein, 2009/orig. 1953, 1958; Hanfling, 2003; Hacker, 2019). With regard to agency, while the behavior may appear as teleological—we describe a machine as wanting to perform a task—it can be described in fully causal terms. In this sense, von Wright's discussion of quasi-teleological explanation is relevant (von Wright, 1971: pp. 153–155). Von Wright describes as quasi-teleological explanations involving biological functions such as explaining the color pattern in a mimicry butterfly in terms of avoiding predators (the explanation appears teleological because avoiding predators here may appear to be the purpose rather than the cause of the color pattern). In

line with von Wright's argument that this kind of explanation would not hold in the case of historical events, one could claim that AI behavior may appear quasi-teleological but in fact it involves a causal process showing it to be an artifact for human use. My view would accept the broad claims about currently observable (and possibly future) AI behavior not matching the kind of human behavior exhibiting specific beliefs, desires, intentions, and such. The point of contention is whether it makes sense to accept a concept of minimal agency in AI as opposed to treating machines simply as artifacts. My answer is that a concept of minimal agency is helpful for distinguishing simpler artifacts from AI systems. While in the case of a tool, say a hammer, the *how* is directly connected to the *why*—there is an obvious connection between the shape and its uses—for an AI system using machine learning to perform a task, the causal chain is longer and not fully intelligible. In this sense, emphasizing the goals which it is meant to serve is crucial to explaining its behavior and potential malfunctions. Particularly in uses regarding responsibility, as argued above, the emphasis of goals can address the shortcoming of focusing exclusively on a causal process which is largely unintelligible.

A final objection to discuss would hold that since the view I defend brings in goals, which are in turn connected to humans and to a type of agency that only applies to human adults, the question is whether this view is anthropocentric in the sense of requiring human-specific characteristics of artificial systems. My answer is that insofar as it holds that machines can follow goals, the view is not anthropocentric. Following goals can be associated with various types of agency, but what is specifically human is the ability to set goals and write them into particular designs. It is at the level of goals where human and artificial agency meet.

7 Conclusions

This paper has argued that the concepts of autonomy and agency in artificial systems involve human goals and their discussion requires a broader framework of values and norms in relation to ethical issues. While current debates, especially regarding ethics in AI, employ different conceptions of artificial agency, I have argued that taking human goals into account is necessary in order to address the problem of responsibility. In comparison with current approaches to the topic, my argument draws from the philosophy of action to provide the basis for including human goals when discussing artificial agency. I have traced the tendency to overlook goals, intentions, and other human-specific factors in specific strands of AI research to ontological assumptions regarding reduction to lower level entities and explaining away teleological notions. Thus, insofar as an approach to responsibility for AI systems is to achieve an adequate level of complexity, a teleological account of artificial agency can supply the relevant components.

Overall, my argument helps clarify conceptual issues arising in the context of AI research, particularly regarding understanding AI behavior. Highlighting the importance of the implicit human component to the behavior of AI systems will help have clearer debates on ethical or policy issues. More broadly, this article also stresses how philosophical analysis can disclose hidden assumptions that obscure aspects of

technology relevant to both epistemological and ethical debates. I have illustrated this with agency, but similar analyses can be conducted in other key areas involving interactions between human and machine behavior.

Acknowledgements I would like to thank Daniel Kodaj, Gunnar Schumann, László Bernáth, and the anonymous referees for their feedback, which has helped improve this article. This paper has been presented at the seminar "New Work on the Metaphysics of Teleology", held at Central European University.

References

Anscombe, G. E. M. (1957). *Intention*. Basil Blackwell.

Asilomar AI Principles (2017). Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017],https://futureoflife.org/ai-principles/.

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior, 17*(5), 367–386.

Bedau, M. (1992). Goal-directed systems and the good. *The Monist, 75*, 34–49.

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence, 72*(1–2), 173–215.

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds & Machines, 30*, 195–218.

Brooks, R.A. (1991). Intelligence without reason. In J. Myopoulos & R. Reiter (Eds.), *Proceedings of the 12th International Joint Conference on Artificial Intelligence* (pp. 569–595). San Mateo: Morgan Kaufmann.

Bryson, J. J., & Kime, P. P. (2011). Just an artifact: Why machines are perceived as moral agents. In *Twenty-second international joint conference on artificial intelligence*.

Burr, C., & Morley, J. (2020). Empowerment or engagement? Digital health technologies for mental healthcare. In *The 2019 Yearbook of the Digital Ethics Lab* (pp. 67–88). Springer, Cham.

Cardoso, R. C., Farrell, M., Luckcuck, M., Ferrando, A., & Fisher, M. (2020). Heterogeneous verification of an autonomous Curiosity rover. In *NASA Formal Methods Symposium* (pp. 353–360). Springer, Cham.

Clark, R. (2010). Skilled activity and the causal theory of action. *Philosophy and Phenomenological Research, 80*(3), 523–555.

Collingwood, R. G. (2001). *An essay on metaphysics*. Oxford University Press.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy, 60*(23), 685–700.

de Miranda, L. (2020). Artificial intelligence and philosophical creativity: From analytics to crealectics. *Human Affairs, 30*(4), 597–607.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and Brain Sciences, 11*(3), 495–505.

Dignum, V. (2017). Responsible autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (IJCAI'2017), pp. 4698–4704.

D'Oro, G., & Sandis, C. (2013). From anti-causalism to causalism and back: A century of the reasons/causes debate. *Reasons and Causes: Causalism and Non-causalism in the Philosophy of Action*, 1–47.

Elliott, K. C. (2017). A tapestry of values: An introduction to values in science. *Oxford University Press*.

Ezenkwu, C. P., & Starkey, A. (2019). Machine autonomy: Definition, approaches, challenges and research gaps. In *Intelligent Computing-Proceedings of the Computing Conference* (pp. 335–358). Springer, Cham.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329.

Friedman, B., Kahn, P., & Borning, A. (2006). Value sensitive design and information systems. *Advances in Management Information Systems, 6*, 348–372.

Froese, T., Virgo, N., & Izquierdo, E. (2007). Autonomy: A review and a reappraisal. In *European Conference on Artificial Life* (pp. 455–464). Springer, Berlin, Heidelberg.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds & Machines, 30*, 411–437.
Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology, 10*(2–3), 115–121.
Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.
Hacker, P. M. S. (2019). *Wittgenstein: Meaning and mind (Volume 3 of an Analytical Commentary on the Philosophical Investigations)*, Part 1: Essays. John Wiley & Sons.
Hanfling, O. (2003). *Wittgenstein and the human form of life*. Routledge.
Illies, C., & Meijers, A. (2009). Artefacts without agency. *The Monist, 92*(3), 420–440.
Johansson, L. (2010). The functional morality of robots. *International Journal of Technoethics, 1*(4), 65–73.
Johnson, D. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.
Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*(2–3), 123–133.
Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. *Minds and Machines, 27*(4), 575–590.
Johnson, D. G., & Verdicchio, M. (2018). Why robots should not be treated like animals. *Ethics and Information Technology, 20*(4), 291–301.
Johnson, D. G., & Verdicchio, M. (2019). AI, agency and responsibility: The VW fraud case and beyond. *Ai & Society, 34*(3), 639–647.
Kitcher, P. (2011a). *Science in a democratic society*. Prometheus Books.
Kitcher, P. (2011b). *The ethical project*. Harvard University Press.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.
Maturana, H.R., & Varela, F.J. (1980). Autopoiesis and cognition: The realization of the living. *Boston Studies in the Philosophy and History of Science*, 42. Dordrecht: Springer Netherlands.
Nolfi, S., & Floreano, D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. MIT Press.
Papineau, D. (2020). Naturalism, *The Stanford encyclopedia of philosophy*, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2020/entries/naturalism/
Peterson, M., & Spahn, A. (2011). Can technological artefacts be moral agents? *Science and Engineering Ethics, 17*(3), 411–424.
Popa, E. (2020). Artificial life and 'nature's purposes': The question of behavioral autonomy. *Human Affairs, 30*(4), 587–596.
Schlosser, M. (2019), "Agency", *The Stanford encyclopedia of philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.),https://plato.stanford.edu/archives/win2019/entries/agency/.
Schumann, G. (Ed.). (2019). *Explanation in action theory and historiography: Causal and teleological approaches*. Routledge.
Sehon, S. (2010). *Teleological explanation. A companion to the philosophy of action*. Blackwell.
Tan, K. H., & Lim, B. P. (2018). The artificial intelligence renaissance: Deep learning and the road to human-level machine intelligence. *APSIPA Transactions on Signal and Information Processing*, *7*.
Tigard, D. W. (2020). There is no techno-responsibility gap. *Philosophy & Technology*, 1–19.
van de Poel, I. (2020a). Embedding values in artificial intelligence (AI) systems. *Minds and Machines, 30*(3), 385–409.
van de Poel, I. (2020b). Three philosophical perspectives on the relation between technology and society, and how they affect the current debate about artificial intelligence. *Human Affairs, 30*(4), 499–511.
van den Hoven, J. (2005). Design for values and values for design. *Information Age, 4*, 4–7.
Von Wright, G. H. (1971). *Explanation and understanding*. Routledge & Kegan Paul.
Wiener, N. (1960). Some moral and technical consequences of automation. *Science, 131*, 1355–1358.
Wittgenstein, L. (1958). *The blue and brown books*. Blackwell.
Wittgenstein, L. (2009). *Philosophical investigations*, 4th edition, P.M.S. Hacker and Joachim Schulte (eds. and trans.), Oxford: Wiley-Blackwell.