Artificial Intelligence Systems, Responsibility and Agential Self-Awareness

Lydia Farina

Abstract This paper investigates the claim that artificial Intelligence Systems cannot be held morally responsible because they do not have an ability for agential self-awareness e.g. they cannot be aware that they are the agents of an action. The main suggestion is that if agential self-awareness and related first person representations presuppose an awareness of a self, the possibility of responsible artificial intelligence systems cannot be evaluated independently of research conducted on the nature of the self. Focusing on a specific account of the self from the phenomenological tradition, this paper suggests that a minimal necessary condition that artificial intelligence systems must satisfy so that they have a capability for self-awareness, is having a minimal self defined as 'a sense of ownership'. As this sense of ownership is usually associated with having a living body, one suggestion is that artificial intelligence systems must have similar living bodies so they can have a sense of self. Discussing cases of robotic animals as examples of the possibility of artificial intelligence systems having a sense of self, the paper concludes that the possibility of artificial intelligence systems having a 'sense of ownership' or a sense of self may be a necessary condition for having responsibility.

Keywords AI responsibility · Artificial self · Agential self awareness · Personal identity

1 Introduction

The current debate on the possibility of attributing moral responsibility to artificial Intelligence Systems focuses on the concepts of autonomy or consciousness (Coeckelbergh, 2020; Müller, 2020). According to some views, moral responsibility for artificial intelligence systems is considered impossible because they lack either autonomy or consciousness and these are necessary (but not sufficient) for moral

L. Farina (✉)
Department of Philosophy, University of Nottingham, Nottingham NG7 2RD, UK
e-mail: lydia.farina@nottingham.ac.uk

V. C. Müller (ed.), *Philosophy and Theory of Artificial Intelligence 2021*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 63,
https://doi.org/10.1007/978-3-031-09153-7_2

responsibility (Haji, 1997; Peacocke, 2014). According to a recent view, what underlies this negative claim is the inability of artificial intelligence systems to have awareness of themselves as the agents of an action by having a capability for first person representations (Sebastian, 2021). Building on the claim that having a capability for first person representations is necessary for moral responsibility, I argue that another necessary but not sufficient criterion for moral responsibility is that artificial intelligence systems have an awareness of a self. It is through this awareness that they can then develop a capability for first person representations and an awareness of themselves as agents of actions. It is the lack of such an awareness of a self that underlies the lack of awareness of themselves as the agents of actions; this keeps them outside the scope of moral responsibility.

In other words, in this paper, I argue that agential self-awareness in the sense of having an awareness of oneself as the one performing an action presupposes an awareness of self. As the concept of a self is important for the main claim of this paper I suggest an account of the self as the 'minimal self'; borrowing from the phenomenological tradition (De Beauvoir, 1947; Husserl, 1952; Sartre, 1957) and a recent account of the self by Wallace (2019), I analyse the self as the subjective experience of having a self without entailing consciousness and without entailing an ontological claim for the self.[1] I argue that the ability to have first person representations presupposes the ability to have representations of a minimal self. If so, research on first person representations and agential self-awareness should be conducted in tandem with research on the nature of the self.

In the first section of the paper I show the relation between responsibility and agential self-awareness. It seems reasonable to hold a person accountable only for her own actions and to expect from a responsible agent to have awareness of herself as the one performing her actions (Aristotle, 1984; Coeckelbergh, 2020). If so, determining moral responsibility in human cases partly depends on the ability shown by agents to have awareness of themselves as being agents of their actions. This 'agential self-awareness' seems to be presupposed in human cases where we think it reasonable to appoint moral responsibility; if one is not capable of being aware that one is the agent of an action, people seem very reluctant to hold one responsible for these actions. Applying this reasoning in the case of artificial intelligence systems, if the capability for agential self-awareness is necessary for attributing responsibility in humans, it is likely that responsible artificial intelligence systems must also have a capability for agential self-awareness. This means that the possibility of responsible artificial intelligence systems largely depends on whether artificial intelligence systems can have agential self-awareness.

In the second section I argue that the ability to have first person representations and relatedly agential self-awareness is based on the ability to have and represent a self. In the case of artificial intelligence systems, if one accepts that artificial intelligence systems are capable of representations, one must then examine whether

[1] This lack of an ontological claim for the self leaves the door open to accounts according to which the self is a construction or an illusion. Here I stay neutral on the metaphysics of the self but assume that whatever the self turns out to be it is necessary for attributions of responsibility.

artificial intelligence systems are capable of representing a self. As there are different views on the nature of the self e.g. physical, psychological, narrative etc., depending on which view one favours, one can give a different answer to the possibility of artificial intelligence systems having agential self-awareness. In any case, to answer the question relating to agential self-awareness we first need to have a good account of what it means to have an artificial self which can be represented as such. Therefore, we first need to answer some very important questions on the nature of personal identity and the self e.g. 'what is the self' and 'in what sense can artificial intelligence systems have or represent a self' before we are able to address the moral responsibility question.

Depending on which account we accept on the nature of the self, we end up giving different answers to the possibility of responsible artificial intelligence systems. For example, phenomenological accounts take the self to refer to a 'minimal self' viewed as the subjective experience of having a self (Sartre, 1957; De Beauvoir, 1947). This minimal self has been analysed in different ways in the literature. Some take it to entail consciousness (Garcia-Carpintero, 2017; Recanati, 2007). Others explain it as a result of the fact that living organisms are self-maintaining systems where the ability of the system to represent itself is important for the overall maintenance of the system (Sebastian, 2018). Following the phenomenological tradition I approach the minimal self as entailing a minimal form of self-experience even if it does not entail consciousness (Gallagher & Zahavi, 2008). In this sense, lack of consciousness does not entail lack of an ability to have a minimal self. I conclude by suggesting that a possible hybrid account combining phenomenological accounts and Wallace's 'network self' gives a promising basis for designing artificial selves.

2 Agential Self-Awareness as a Necessary Condition for Responsibility

In this section I discuss the relation between moral responsibility and agential self-Awareness. In typical cases, we do not hold agents responsible for actions if these agents are not aware that they are the agents of those actions (Aristotle, 1984; Coeckelbergh, 2020). In such cases, to determine moral responsibility of an agent we typically look for a capability of being aware of one's own actions. This capacity refers to what is known in the literature as the 'epistemic condition' where one of the conditions of responsibility is that the agent knows or is aware of what she is doing (Rudy-Hiller, 2018). If this capacity does not exist, we are not likely to start a debate on whether the agent is morally responsible for an action. Here we can think of cases of some non-human animals or humans who do not have this capacity to be aware of themselves as agents of actions. Where this capacity is absent we are reluctant to appoint responsibility.

It is also important to differentiate between agents who lack this capacity outright from agents who have this capacity but do not exercise it. It seems that we would be

willing to appoint some responsibility to agents who have this capacity, even if we end up holding them partially responsible for an action. An example here is cases of 'temporary insanity' where agents declare that they were not aware of themselves as the agents of an action that they actually performed. It seems that in these cases questions of the type 'Why did you do action A?' become obsolete. The possible answers an agent can give in these cases are not answers involving giving reasons for their actions e.g. such an agent may respond 'I did not do A', or 'I did not know that I did A', or 'I did not know that I was doing A when I was doing A'. In these cases it seems to be the case that both the 'control' and the 'epistemic' conditions are not satisfied as the agent appears to lack control over her actions and is not aware that she is the one performing it (Rudy-Hiller, 2018).

These cases highlight the important links between internal states of the agent whilst the agent is deliberating to do A and between the internal states of the agent whilst the agent is doing A (Mele, 2019). More specifically it seems that the link between deliberating whether to do A and being aware that I am now doing A breaks down. I argue that the awareness by the agent that they are now doing A is a necessary but not sufficient criterion for determining moral responsibility. In other words I argue that it is necessary that the agent is aware of themselves as the agents of the action when the action is being performed what I call 'agential self-awareness' in the remaining of this paper.[2] To give an example, Eva is throwing a cricket ball at the window. If Eva is not aware that she is the one throwing the cricket ball at the shop window, Eva does not have this agential self-awareness. In the absence of such self-awareness, determining moral responsibility becomes very difficult if not impossible because, as a general rule, we do not hold agents morally responsible for actions they commit unless they are aware that they are the agents of these actions when these are being performed.

A relevant consideration here is that in the absence of this agential self-awareness, the debate on the causal roles of internal states e.g. desires, goals, intentions of the agent before the performance of the action which may have brought about the performance of the action become obsolete. It does not matter what internal states are causally responsible for Eva's action if Eva is not aware of being the agent of the action because the links between deliberation and action break down. These internal states would explain the action only when Eva is aware that she is the agent of the action performed. Eva may be deliberating whether to throw the ball at the shop window but unless she is aware that she is the one lifting her arm and releasing the ball at the direction of the window, the relevance of the internal states during deliberation are not important. If one turns to Eva right after she threw the ball and broke the shop window to ask 'Why did you throw the ball towards the window?' Eva is likely to answer 'I did not throw the ball to the shop window' rather than 'I had my reasons for intending to break the shop window' or 'I did not think it would break'.

[2] For other types of awareness that also appear to be necessary so that the 'epistemic' condition is satisfied see Rudy-Hiller (2018), Coeckelbergh (2020).

The example above can be used to show that legal responsibility and moral responsibility occupy different places on the general debate on agent responsibility. To establish legal responsibility all we have to do is to show a recording of the event to Eva and to a jury which clearly shows Eva throwing the cricket ball at the shop window. In the case where Eva is aware of herself as the agent of the action she is responsible for her action both legally and morally. As it happens she decided to throw it because the Mafia is holding her brother and threaten to hurt him if she does not throw the ball. However, in the case where Eva is not aware of throwing the ball at the window or even of being in front of the window holding a cricket ball and aiming at the window, it seems that we would not think it reasonable to hold her morally responsible for this action even if we hold her partially legally responsible for the action.

In a recent article, Coeckelbergh (2020) argues that the debate on responsibility attribution in the case of artificial intelligence systems should include an understanding of responsibility as answerability in the sense of moving from a discussion of control and knowledge to a more relational view where moral patients deserve answers and explanations from moral agents concerning what is done with them and why. My view is compatible with this account as I take the moral agent to be aware of herself as agent so that she can be in the position to give appropriate answers, reasons or explanations for her actions. Applying this reasoning in the case of artificial intelligence systems, if the capability for agential self-awareness is necessary for attributing responsibility in humans, it is likely that responsible artificial intelligence systems must also have a capability for agential self-awareness. This means that the possibility of responsible artificial intelligence systems largely depends on whether artificial intelligence systems can have agential self-awareness. In the next section I discuss the possibility of artificial shelves in the sense of artificial intelligence systems with an ability to have agential self-awareness.

3 The Minimal Self as a Necessary Condition for Agential Self-Awareness

In Sect. 2 I argue that in order for agents to be able to give reasons or explanations for their actions they need to have the ability to be aware of themselves as agents of those actions. In this section I argue that the ability to have this agential self-awareness is based on the ability to have and represent a self. What this means for the possibility of responsible artificial intelligence systems, if one accepts that artificial intelligence systems are capable of representations, one must then examine whether artificial intelligence systems are capable of agential self-awareness. In other words, the capability for agential self-awareness presupposes the ability to represent a self. If so, before we answer the question of AI self-representation in the sense of de se representation we need to answer the question of the possibility of artificial selves.

There are different views on the nature of the self in the literature with dominant views taking the concept of the self to be synonymous with personal identity (Kind, 2015). Accounts focusing on the reidentification issue of personal identity e.g. physical/bodily or psychological views focus on the question of identifying the same self over time (see for example Parfit (1984), Olson (1997)). On the other hand other accounts, such as Schechtman's narrative account, focus on the characterisation issue of personal identity or on the identity of a self e.g. on what makes us who we really are (Schechtman, 2014; Ricoeur, 1991). To answer the question of whether it would be possible for artificial intelligence systems to have a self so as to represent it I argue that we need to focus on identity as self rather than identity as sameness.[3] In other words, to answer the question relating to AI agential self-awareness, we first need to answer some very important questions on the nature of personal identity e.g. 'what is the self' and 'in what sense can artificial intelligence systems have or represent a self' before we are able to address the moral responsibility question.

3.1 What Is the Self?

One way to give an answer to this question is by focusing on phenomenological accounts which take the self to refer to a 'minimal self' viewed as the subjective experience of having a self (Sartre, 1943; De Beauvoir, 1947). This minimal self has been analysed in different ways in the literature. Some take it to entail consciousness (Garcia-Carpintero, 2017; Recanati, 2007). Others explain it as a result of the fact that living organisms are self-maintaining systems where the ability of the system to represent itself is important for the overall maintenance of the system (Sebastian, 2018). On the other hand, other accounts take the self to refer to a person which is more than simply the subjective experience or the 'I' or to self-consciousness. For example Wallace's recent account takes the self to be a cumulative network of traits e.g. physical, psychological, biological traits and relations e.g. social, familial, ethnic, cultural (Wallace, 2019, 1). As these traits and relations change over time this cumulative network is temporal and it has a synchronic and diachronic unity.[4]

Depending on whether we choose more or less demanding accounts on the nature of the self, we end up giving different answers to the possibility of responsible artificial intelligence systems. However, accounts which take the self to refer to a person presuppose the existence of a minimal self or an 'I'. As in this paper I am interested in bringing to light what would be the minimal requirements that artificial

[3] I am borrowing these descriptions from Kind (2015).

[4] Wallace's (2019) account unites two different traditions trying to explain the nature of the self. From the feminist tradition she borrows the idea of a social or relational self conceptualised partly in terms of relationships and social relations. From philosophical theories on the nature of time she borrows the idea that a self has temporal features to claim that the self is a process rather than an object persisting in time. She defines it as 'a relational, plurally constituted complex, that is, a network of interrelated biological, genetic, physical, social, psychosocial, linguistic, semantic, and so on traits' (Wallace, 2019, 8).

intelligence systems must satisfy so that they have a capability for agential self-awareness, I focus on the self as a minimal self rather than the self as a person. Following the phenomenological tradition I approach the minimal self as entailing a minimal form of self-experience even if it does not entail consciousness in the sense of reflective consciousness. In this sense, lack of consciousness does not entail lack of an ability to have a minimal self. This would allow several entities to have minimal selves even if these do not have a capacity to develop into full blown persons.

Gallagher and Zahavi define this minimal self as 'the immediate and first-personal givenness of experience' which is explained in terms of a pre-reflective self-consciousness (Gallagher & Zahavi, 2021). They give examples of when we are experiencing this pre-reflective consciousness: 'the pre-reflective self-consciousness which is present whenever I am living through or undergoing an experience, e.g., whenever I am consciously perceiving the world, remembering a past event, imagining a future event, thinking an occurrent thought, or feeling sad or happy, thirsty or in pain, and so forth' (Gallagher & Zahavi, 2021).

The minimal self according to this view, is explained as pre-reflective self-awareness and not as an object of some other act of consciousness e.g. another mental state. It is 'lived through' rather than appear in an objectified manner (Husserl, 1952; Sartre, 1943).[5] This pre-reflective self-awareness is necessary for the ability to represent a self. It is explained as a sense of ownership, or sense of for-me-ness. Only beings with this sense of ownership can form the concept of a self and employ goal directed behaviour and execute actions for which they will take responsibility (Gallagher & Zahavi, 2021). As Smith (2020) explains according to its proponents, this self-awareness is pre-reflective 'in the sense that it does not require one to explicitly reflect one's own mental states, or to otherwise take them as objects of attention' (§3.2). The question I focus on for the remaining of this paper is whether artificial intelligence systems can have this pre-reflective sense of ownership.[6]

Although the account of the minimal self discussed above is far less demanding than accounts taking the self to be synonymous to a person, a difficulty here still remains in that the sense of ownership presupposes the ability to 'sense'. This makes it the case that even if reflective consciousness is not necessary for the minimal self, the ability to sense appears to be necessary. This sense appears to be a prerequisite for other types of sensing e.g. a sense of agency where one is the author of a mental state (Bayne, 2008), a sense of ownership where one is the owner of a mental state or a sense of location where one senses that the mental state is located in one's own mind (Smith, 2020). If so, the question of artificial selves and related questions on the possibility of AI agential self-awareness depends on whether artificial intelligence

[5] Gallagher and Zahavi note that this does not mean that this pre-reflective self-awareness cannot become the object of another mental state; instead, they interpret this to mean that 'To be self-aware is not to capture a pure self or self-object that exists separately from the stream of experience, rather it is to be conscious of one's experience in its intrinsic first-person mode of givenness' (Gallagher & Zahavi, 2021).

[6] See Smith (2020) (§3.2 and §3.3) for an interesting discussion of the literature on the minimal self and the sense of ownership.

systems can have phenomenal consciousness.[7] Although the debate on this issue is still ongoing, the relation between this debate and the debate on AI responsibility are typically conducted independently from one another. However as the discussion in this chapter shows, the question of AI agential self-awareness, AI responsibility and the possibility of AI having phenomenal consciousness are intimately related and should not be conducted independently from one another. In the next chapter I examine whether recent experiments using synthetic cells can provide support for the claim that artificial intelligence systems could have this sense of ownership.

3.2 Towards Artificial Selves

In 2016 scientists at the University of Harvard created an artificial stingray described as 'an artificial animal—a tissue engineered ray—who could swim and phototactically follow a light cue' (Park et al., 2016). In more simple words, they created a swimming robot that mimics a ray fish and can be guided by light. This robotic ray was a combination of tissue from the heart of rats and a micro fabricated gold skeleton. This is an example of a hybrid 'anibot' a robot mimicking an animal in this case a ray fish in appearance and behaviour. This robotic ray has a number of living cells which enable it to move following light cues.

In 2021, Blackiston et al. constructed 'synthetic living machines' out of frog skin cells (Blackiston et al., 2021). These robots, named *xenobots* after the frog species *Xenopus laevis* from where the skin cells originated, can swim, and have the ability to sense their environment on the basis of being sensitive to light. These xenobots operate in swarms working together to complete specific tasks. As they are created from cells, these xenobots are eventually biodegradable e.g. they eventually break apart. The ongoing research on xenobots is fascinating in many different respects e.g. one of the possible applications could be to use this research to restore damaged organs in human animals (Ebhrahimkhani & Levin, 2021). In this paper, I focus on the light this research can shed on the possibility of artificial intelligence systems having the ability to sense. As Ebhrahimkhani and Levin claim, xenobots 'blur traditional definitions attempting to cleanly demarcate categories of living beings, robots and machines' (2021, 14). As these xenobots are based on living cells, have an ability to

[7] In the literature this is also described as phenomenal consciousness. Higher order theories of consciousness pose no real problem to the possibility of a minimal self for artificial agents. For the agent to be aware that they are performing *A* all we need is two mental states or mental representations on representational accounts. The first mental state *A* is a pre-reflective type of consciousness and the second mental state *B* a reflective type of consciousness looking back at mental state *A* or having as its object mental state *A*. In this paper I do not focus on higher order theories of consciousness. I agree with phenomenologists (e.g. Gallagher and Zahavi (2008) or Kriegel (2009)) that simply because one state has another state as object is not a good explanation of reflective consciousness or phenomenal consciousness. I focus on phenomenological accounts which do not explain reflective or phenomenal consciousness in this way.

sense their environment and eventually perish, one could claim that they do have an ability to sense their environment and more importantly their bodies.

In the previous chapter I showed that a minimal requirement for the possibility that artificial intelligence systems can have an artificial self is that they have this sense of ownership that human and some non-human animals seem to have. Our ability to sense our body is intimately related to having a sense of ownership of that body and to the phenomenology that usually accompanies having a body. The cases of robotic rays and xenobots discussed above, show that it is in principle possible for us to design artificial intelligence systems who use living cells to perform various functions. If human and non-human animals have cellular bodies and the ability to sense is associated with cellular organisms, giving similar bodies to artificial intelligence systems could be one way of giving them the ability for a sense of ownership of those bodies.

To sum up, in this section I discuss an account of the minimal self which explains the self as having a sense of ownership usually cached out as having a sense of ownership of a living body. Such an account presupposes that the minimal self is intimately related to our ability to sense or have phenomenal consciousness. If having this sense is a necessary condition in human animals for having agential self-awareness and responsibility, this sense may also be necessary if we wish to create responsible artificial agents. I discuss recent research in developmental biology where artificial intelligence systems are paired with or based on living cells and appear to interact with their environment. According to the researchers conducting these experiments, it is becoming more difficult to maintain strict boundaries between living organisms and machines. This research also allows for the possibility of hybrid organisms e.g. organisms combining artificial intelligence with living cells. As the debate on explaining this sense of ownership in the wider literature is still ongoing (Smith, 2020), it is an open question whether these experiments show artificial intelligence systems with a sense of ownership in the required sense. I hope the discussion shows that this debate is very important for questions on the possibility of AI responsibility.

4 Conclusion

In Sect. 2 I argue that the concept of responsibility is intimately related to research conducted on the nature of the self and personal identity. Focusing on the concept of the minimal self I show that a good analysis of the minimal self and of the ability to have agential self-awareness is needed if we are discussing the possibility of AI responsibility. The main claim of this paper is that a necessary condition for the possibility of holding artificial intelligence systems responsible for any actions conducted autonomously, is that they have a minimal self which can then be represented. Assuming that artificial intelligence systems are in principle capable of representations, if they do have a minimal self, then it is an open question whether they can have a capability for agential self-awareness.

The discussion in Sect. 3 focused on the possibility of artificial intelligence systems having a minimal self defined as a sense of ownership. As argued earlier in this paper, if this sense of ownership is necessary for agential self-awareness, artificial intelligence systems must have a capacity for it. As the focus of this paper is to highlight a minimal necessary condition so that artificial intelligence systems have a minimal self, I do not discuss several significant ethical considerations attaching to the possibility of designing robotic animals by using living cells. In a similar way, I do not discuss a related epistemological question of whether we would be in a position to know whether or not artificial intelligence systems have this sense of ownership. This epistemological concern is important in order for us to know whether artificial intelligence systems are capable for agential self-awareness and whether they can be considered morally responsible agents.

As my focus in this paper is to establish a minimal necessary condition for responsibility without claiming that this condition is sufficient, the epistemological concern remains relevant and will need to be addressed before we reach a conclusion on the possibility of responsible artificial intelligence systems. However, the epistemological concern does not seem to be incompatible with the main claim of this paper; even if we were in a position to know that artificial intelligence systems have this sense of ownership, on its own this would not entail that they are responsible agents. In a similar way, although we do know that human and some non-human animals have this sense of ownership, we do not automatically hold them responsible for their actions on this basis. To conclude, by discussing two recent examples where artificial intelligence systems are paired with synthetic cells, I suggest that some questions relating to the possibility of artificial life or the possibility of artificial intelligence systems having the ability for a sense of ownership may not be as easily dismissed now as they have been in past years.

References

Aristotle. (1984). Nicomachean ethics. In J. Barnes (Ed.), *The complete works of Aristotle* (Vol. 2, pp. 1729–1867). Princeton University Press.

Bayne, T. (2008). The phenomenology of agency. *Philosophy Compass, 3*(1), 182–202.

Blackiston, D., Lederer, E., Kriegman, S., Garnier, S., Bongard, J., & Levin, M. (2021). A cellular platform for the development of synthetic living machines. *Science Robotics*, *6*, eabf1571.

Coeckelbergh, M. (2020). Intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics, 26*, 2051–2068.

De Beauvoir, S. (1947). *The ethics of ambiguity* (B. Frechtman, Trans.). Open Road Integrated Media. (2018).

Ebhrahimkhani, R., & Levin, M. (2021). Synthetic living machines: A new window on life. *iScience*, *24*, 102505.

Gallagher, S., & Zahavi, D. (2008). *The phenomenological mind*. Routledge.

Gallagher, S., & Zahavi, D. (2021). Phenomenological approaches to self-consciousness. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021 ed.). https://plato.stanford.edu/archives/spr2021/entries/self-consciousness-phenomenological/.

Garcia-Carpintero, M. (2017). The philosophical significance of the De Se. *Inquiry, 60*(3), 253–276.

Haji, I. (1997). An epistemic dimension of blameworthiness. *Philosophy and Phenomenological Research, 57*(3), 523–544.
Husserl, E. (1952). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy* (Second Book) (R. Rojcewicz & A. Schuwer, Trans.). Kluwer. (1989).
Kind, A. (2015). *Persons and personal identity*. Polity Press.
Kriegel, U. (2009). *Subjective consciousness: A self-representational theory*. Oxford University Press.
Mele, A. (2019). *Manipulated agents: A window to moral responsibility*. Oxford University Press.
Müller, V. (2020). Ethics of artificial intelligence and robotics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 ed.). https:// plato.stanford.edu/archives/ fall2020/entries/ethics-ai/.
Olson, E. (1997). *The Human Animal: Personal Identity Without Psychology*. Oxford University Press.
Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.
Park, S.-J., Gazzola, M., Park, K. S., Park, S., Di Santo, V., Blevins, E. L., Lind, J. U., Campbell, P. H., Dauth, S., Capulli, A. K., Pasqualini, F. S., Ahn, S., Cho, A., Yuan, H., Maoz, B. M., Vijaykumar, R., Choi, J. W., Deisseroth, K., Lauder, G. V., … Parker, K. K. (2016). Phototactic guidance of a tissue-engineered soft-robotic ray. *Science, 353*, 158–162.
Peacocke, C. (2014). *The mirror of the world: Subjects, consciousness, and self-consciousness*. Oxford University Press.
Recanati, F. (2007). *Perspectival thought: A plea for (moderate) relativism*. Oxford University Press.
Ricoeur, P. (1991). 'Narrative identity' translated by Mark S. Muldoon. *Philosophy Today*, *35*, 73–81.
Rudy-Hiller, F. (2018). The epistemic condition for moral responsibility. In *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/moral-responsibility-epistemic/.
Sartre, J.P. (1943). *Being and Nothingness; An Essay on Phenomenological Ontology*. Translated by Hazel. E. Barnes, London: Routledge, 1998.
Sartre, J. P. (1957). *The transcendence of the ego* (F. Williams & R. Kirkpatrick, Trans.). Noonday Press.
Schechtman, M. (2014). *Staying alive: Personal identity*. Oxford University Press.
Sebastian, M. A. (2018). Embodied appearance properties and subjectivity. *Adaptative Behaviour, 26*(5), 199–210.
Sebastian, M. A. (2021). First person representations and responsible agency in AI. *Synthese*. https://doi.org/10.1007/s11229-021-03105-8.
Smith, J. (2020). Self-consciousness, In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2020 ed.), https://plato.stanford.edu/archives/sum2020/entries/self-consciousness/.
Wallace, K. (2019). *The network self: Relation, process and personal identity*. Routledge studies in American philosophy. Taylor and Francis.