AI and Ethics (2022) 2:789–800
https://doi.org/10.1007/s43681-022-00140-0

ORIGINAL RESEARCH

How a non-conscious robot could be an agent with capacity for morally responsible behaviour

Atle Ottesen Søvik[1]

Received: 4 November 2021 / Accepted: 18 January 2022 / Published online: 21 February 2022

Abstract

People have different opinions about which conditions robots would need to fulfil—and for what reasons—to be moral agents. Standardists hold that specific internal states (like rationality, free will or phenomenal consciousness) are necessary in artificial agents, and robots are thus not moral agents since they lack these internal states. Functionalists hold that what matters are certain behaviours and reactions—independent of what the internal states may be—implying that robots can be moral agents as long as the behaviour is adequate. This article defends a standardist view in the sense that the internal states are what matters for determining the moral agency of the robot, but it will be unique in being an internalist theory defending a large degree of robot responsibility, even though humans, but not robots, are taken to have phenomenal consciousness. This view is based on an event-causal libertarian theory of free will and a revisionist theory of responsibility, which combined explain how free will and responsibility can come in degrees. This is meant to be a middle position between typical compatibilist and libertarian views, securing the strengths of both sides. The theories are then applied to robots, making it possible to be quite precise about what it means that robots can have a certain degree of moral responsibility, and why. Defending this libertarian form of free will and responsibility then implies that non-conscious robots can have a stronger form of free will and responsibility than what is commonly defended in the literature on robot responsibility.

Keywords Robot responsibility · Internalism · Event-causal libertarianism · Revisionist responsibility

1 Introduction

There is an ongoing debate about whether robots might become moral agents (artificial moral agents, or AMAs).[1] Views differ regarding which conditions robots would need to fulfil and why. Behdadi and Munthe distinguish between *standardists* and *functionalists*.[2] Standardists hold that specific internal states (like rationality, free will, or phenomenal consciousness) are necessary in artificial agents, and robots are thus not moral agents since they lack these internal states.[3] Functionalists hold that what matters are certain behaviours and reactions independent of what the internal states may be, implying that robots can be moral agents—at least to a certain degree.[4]

The distinction between standardists and functionalists is coarse. In fact, there are many possible positions spread along a continuum. For example, internal states can be taken to require consciousness or not; consciousness can be understood as phenomenal or not; and phenomenal consciousness can be understood as physical or not. Views then differ on which combinations of which of these properties robots could have.

While many combinations are possible, it is natural to be more open to robots having internal states like rationality, free will and responsibility if these are *not* taken to require a special form of consciousness that humans have and robots do not. Articles arguing that robots (can) have free will, responsibility, etc., typically argue that the human mind is not very different from how machines work or that consciousness is not relevant.[5]

✉ Atle Ottesen Søvik
Atle.O.Sovik@mf.no

[1] MF Norwegian School of Theology, Majorstua, P.O. Box 5144, 0302 Oslo, Norway

[1] "Robot" and "artificial agent" are here used as synonyms for linguistic variation, even if "artificial agent" is a wider concept than "robot". "Robot" refers to "artificial agent", which again is understood as it is defined by Russell and Norvig below. "Machine" is sometimes used for variation, which again should be read as "artificial agent".

[2] Behdadi and Munthe [4]. Page numbers to this source refer to the online version found at https://philpapers.org/archive/BEHANA-2.pdf

[3] For example Johnson [18].

[4] For example Floridi [12], and Sullins [44].

[5] Behdadi and Munthe [4].

However, it is possible to think that humans have a special form of consciousness and still argue that robots lacking this property could, to a large degree, have free will and responsibility. This is the position defended in this article. Human consciousness will be understood as non-physical phenomenal consciousness, yet the argument is that non-conscious robots to largely have the same free will and responsibility as us—but with some important differences. The goal of this article is to give a detailed theory of how that is possible and what the exact relevance of the phenomenal consciousness is. If successful, this would show that even those who think humans are unique when it comes to consciousness could still to a large degree accept robot responsibility.

A famous definition of phenomenal consciousness (or qualia) is that it is something it is like for a subject to experience it.[6] It is subjective, qualitative experience we are aware of from a first-person perspective, like sense impressions, thoughts, feelings, and desires. Robots are commonly taken not to be conscious in this sense, but phenomenal consciousness can be understood either as physical or not. Views on consciousness are spread on a continuum from reductive physicalism, where consciousness is nothing but something physical, to substance dualism, where consciousness is something completely different from the physical. Along this line are middle positions like non-reductive physicalism, emergentism, property dualism, and various stripes of panpsychism and panqualityism, where consciousness can be taken to be physical or non-physical, or the meaning of "physical" is contested.[7]

Some functionalists argue that we do not know whether other people or machines are conscious, but that it does not matter; what matters is how we should relate to machines with similar behaviour to humans.[8] The problem with this response is that how we should relate to machines depends on what is best justified as being true about them. While we may have good reasons to act in certain ways given that we are uncertain about what is true about them, there is no good reason not to try to work out as best we can the most coherent theory about what is true about the internal states of machines and the moral implications of this truth. This article defends a standardist view in the sense that the internal states are what matters for determining the moral agency of the robot, but it is unusual in being an internalist theory defending to a large degree robot responsibility, even though humans, but not robots, are taken to have non-physical phenomenal consciousness. (From now on, "consciousness" refers to this understanding of phenomenal consciousness.)

When it comes to free will, the standard theories are compatibilism and libertarianism. Compatibilists hold that free will is compatible with the world being determined, which libertarians reject. Libertarians argue that compatibilism implies a concept of free will too weak to deserve to be called "free will", while compatibilists reject that it is possible to give a coherent libertarian theory.

A form of libertarianism very close to compatibilism is called event-causal libertarianism.[9] Different from other forms of libertarianism (like agent-causal[10] or non-causal libertarianism[11]), it holds that the human mind is a causal process like other causal processes found in nature. It is like compatibilism in holding that the human mind is a normal causal process, but like libertarianism in holding that indeterminism is necessary for free will. Event-causal libertarianism is a middle position between the typical forms of compatibilism and libertarianism and tries to get the best from both sides—a stronger form of free will easier to justify as a coherent and plausible theory.

This article will (in part two) present an event-causal libertarian understanding of free will and responsibility as basis for part three, which presents a theory of how and to what degree non-conscious robots can have moral responsibility and why. The conclusion will be that it is plausible that advanced non-conscious robots can be agents with capacity for moral responsibility when certain conditions are fulfilled.[12] Part four then answers objections before the conclusion in part five.

2 An event-causal libertarian theory of free will and responsibility

The standard theories of free will are different forms of compatibilism and libertarianism.[13] Compatibilists hold that free will is compatible with the world being determined, which libertarians reject. Determinism means that there is only one possible content of the future, usually understood as being determined by the laws of nature and the initial conditions of the universe.[14]

Libertarians argue that compatibilism implies a concept of free will too weak to deserve to be called free will. If the content of the future is determined before we were born we

[6] Nagel [27].

[7] Kim [22].

[8] Behdadi and Munthe [4].

[9] See for example Kane [21].

[10] See for example O'Connor [28].

[11] See for example Ginet [13].

[12] I use "capacity for moral responsibility" instead of just "moral responsibility", since some will say that you are only morally responsible when you have done a particular action with morally loaded consequences. The focus here is on the general capacity for doing actions for which one can appropriately be held responsible.

[13] For an overview of different theories of free will using the same terminology, see Kane [20].

[14] Hodgson [17].

cannot change what will happen, and thus we should reject that we have free will.[15] Libertarians use the term “free will” in a stronger sense: humans have power to influence what the content of the future will be without it being determined in advance. This is often described by libertarians by saying that humans must be the source of their own actions and have control, such that it is partially up to them what happens.[16]

Compatibilists reject that it is possible to give a coherent libertarian theory. Many libertarians are criticized for appealing to mysterious entities like irreducible agents and irreducible agent causation or even acting without causation.[17] The main picture is thus that it is easier to offer a coherent theory of a weak understanding of free will (as in compatibilism), but difficult to give a coherent theory of a strong understanding of free will (as in libertarianism). Instead of asking whether humans have free will, we could ask how strong a form of free will it is possible to give a plausible theory of.

There is a form of libertarianism which is very close to compatibilism. It is called event-causal libertarianism. Compatibilists hold that the mind is a causal process like others in nature, whereas most libertarians think that the mind is a unique form of agent causation or not a causal process at all. Event-causal libertarianism is like compatibilism in that the mind is an event-causal process but like libertarianism in that indeterminism is necessary for free will. By holding that the mind is a normal causal process, event-causal libertarians try to avoid criticism of being mysterious or incoherent. Instead, they try to get the best from both sides—a stronger form of free will easier to justify as a coherent and plausible theory.

Of course, event-causal libertarianism is also criticized. There are many objections against any theory of free will, but the main objection against this theory is the regress problem.[18] If the mind is a causal process, it seems we can trace it backwards in time to a point of time before the agent was born. How can there then be an agent who is the source of her actions? What makes it right to say that a choice is up to the agent or in the control of the agent?

The rest of this section will be a presentation of an event-causal theory of free will and responsibility. It is meant to be a theory of the strongest form of free will which can be plausibly defended, which is more than a weak compatibilist understanding. In two recent books, I have defended the theory in much more detail,[19] but here it must suffice to show the main features and advantages.

This theory is similar to Alfred Mele’s Daring Soft Libertarianism and Chandra Sripada’s Deep Self theory.[20] The main idea is to think of the self as developing gradually over time, implying that free will also come in degrees and develops gradually over time. Details on how a self can develop gradually over time are given in the self-theory of neuroscientist Antonio Damasio. He distinguishes between the core self, which is the conscious experience of being a subject from moment to moment, and the autobiographical self, where we store memories and which gives us an identity as persons over time.[21]

Humans start life without free will. Innate desires or random choices cause us to choose among different alternatives. The results of our choices are events that we experience as good or bad and store as memories in our autobiographical self. In later choices, the memories in the autobiographical self-start to influence which alternatives we choose. This way the autobiographical self can cause choices—and cause changes to itself—as new experiences get stored in the autobiographical self. While nothing can be the cause of itself, a self can over time be the cause of the *content* of itself in the way just described. We know this process as self-formation (or development of character or personality) over time.

This is how to understand what it means to will what you will, to be source of your choices, or to control your choices. To control a choice is to cause the choice, for how can you control something to which you are not causally connected?[22] This is also how the regress problem can be solved. If we live in an indeterministic world, there are many states of affairs pushing and pulling in different directions, with indetermined events occurring and no results determined to happen. In such a scenario, it will sometimes be right to select an autobiographical self as the cause of why A happened as opposed to B. A full defence of this claim requires a detailed theory of causation and selection of causes for which there is not room here.[23]

This process of self-formation is described in the following figure:

[15] Standard arguments against the compatibilist understanding of free will is the consequence argument (Van Inwagen [45]) and the manipulation argument (Pereboom [30]).

[16] Kane [20].

[17] Pereboom [30].

[18] See for example Bok [5], 201–05., or Strawson [43].

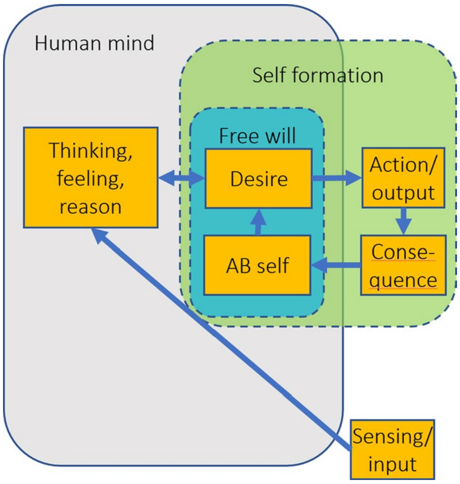
[19] See Søvik [39, 40].

[20] Mele [26]; Sripada [42].

[21] Damasio [11].

[22] Mele [24].

[23] See Schaffer [35]; and Søvik [39], chapter 2.



This concludes the presentation of the event-causal theory of free will. To discuss robot responsibility, a theory of responsibility is required. There are two main theories of responsibility: basic-desert theories and consequentialist theories.[24] Basic-desert theories say that a person deserves to be held responsible when certain conditions are fulfilled, while consequentialist theories say that we hold people responsible to achieve good consequences. Libertarians typically hold basic-desert theories while compatibilists are typically consequentialists.

The situation with responsibility is similar to the situation with free will in the following way: Those who defend responsibility as basic-desert argue that consequentialism does not deserve to be called “responsibility”, whereas those who defend responsibility as consequentialism argue that there can be no coherent defence of humans having responsibility in the sense of basic desert.[25] Again, one could ask how strong form of responsibility it is possible to give a plausible defence of.

The rest of this section will be a defence of a libertarian consequentialist theory of responsibility. Like with event-causal libertarianism, it is close to compatibilism in understanding responsibility as consequentialism, but it differs from compatibilism in arguing that indeterminism is important for responsibility. It is thus a stronger form of responsibility than compatibilist consequentialism but a weaker form of responsibility than libertarian basic desert. Libertarian consequentialism is meant to be the strongest form of responsibility that can be coherently defended.

What is the difference between compatibilist and libertarian consequentialism? If the world is determined, then when A holds B responsible for C, it was determined before they were born that A should hold B responsible for C. Then the series of events D, E, F… following after A held B responsible for C was also determined to happen before A and B were born. If the world is *not* determined, what will happen in the future is open, and whether A holds B responsible for C or not will influence whether the future becomes D, E, F… or X, Y, Z….

A compatibilist could argue that a person will feel no difference between living in a deterministic or an indeterministic world since one would have to go through the process of deliberating and making a choice in any case. But there is a metaphysical difference in what actually happens and an existential difference in the importance of our choices. Only in an indeterministic world will it be true that it is physically possible for you to make the future go in either direction A or B. Seeing what is best justified as true when it comes to determinism does influence how meaningful and important we understand our choices to be.[26]

The theory of responsibility defended in this article is the revisionist theory of Manuel Vargas, but with the important difference that Vargas combines it with a *compatibilist* understanding of free will,[27] while it is in this article combined with a *libertarian* understanding of free will, which means that holding others responsible is an important action which can make the future become either *this* or *that*. This is important since many libertarians would reject that the concept of responsibility makes sense if everyone just does what they were determined before their birth to do anyway.

In the following, Manuel Vargas’ revisionist theory of responsibility is presented.[28] According to Vargas, holding others responsible is a general strategy for cultivating good behaviour in others. First, a look at how he defends his view against the four most common objections.[29]

The first objection is that a consequentialist theory of responsibility seems to equate responsibility with influence. But many things can be influenced in many ways which we would not call responsibility, so it must be something more than just influence. Vargas responds that responsibility is a certain *kind* of influence where we influence the deliberation processes of people to make them want to behave differently.

The second objection is that holding others responsible often does not have the consequence that people behave better. Vargas agrees but nevertheless thinks that holding others responsible as a general strategy often—even if not always—does work.

[24] Pereboom [31].

[25] Derk Pereboom is an example of a philosopher who argues that responsibility should be understood as basic desert but also finds this understanding of responsibility to be incoherent, with the result that he rejects the claim that humans are responsible for their actions [30].

[26] Vohs and Schooler [47], Baumeister et al. [2].

[27] Vargas [46].

[28] Ibid.

[29] Ibid., 187–95.

The third objection is that we should distinguish between holding others responsible and holding others *appropriately* responsible. It seems that we can be wrong in holding others responsible, but how can we be wrong if holding others responsible just means to influence others to behave better? Putting innocent people in prison for a certain crime could function to make more people behave better and not do this crime. Vargas answers that, as a general strategy, it functions best to influence those who are causally responsible for something. He could also have added other ethical reasons for not punishing innocent people.

The fourth objection is that we often praise and blame people without the intention of cultivating good behaviour. We may, for example, blame someone who is dead. Vargas answers that while cultivating good behaviour is often not the intention, it is still a practice that has this effect, and even blaming the dead can influence the living.

Vargas' account can be strengthened with an addition from Hilary Bok on what we do when we hold others responsible.[30] Bok argues that we compare their actions with a standard for what we think they should have done in the situation, which means that we can hold others responsible also for not acting or for events that they did not cause. For example, we can blame someone for not trying to help a drowning child, even if they did not cause the drowning.

Holding someone responsible presupposes that they have a capacity for being held responsible, which not all people have. This capacity means to have the kind of normal cognitive and emotional capacities that make it possible for the person to be influenced through a normal deliberation process to act in a morally different way by being held responsible.

Free will and responsibility are not either/or issues. Instead, each of them exists on a continuum where the degree of freedom has to do with the involvement of the autobiographical self: how strongly it is involved in the deliberation process and how independent it is having been involved in earlier deliberation processes. Free will and responsibility develop over time. The degree of free will or the degree of control that we have over our actions is the degree to which our actions are influenced by an independent autobiographical self. Holding others responsible works by influencing the person in the way that (expected) acts of praise and blame are events the person experiences and stores in their memory, which are then activated and influence future choices.

This concludes the presentation of free will and responsibility, which explains how they can come in degrees and be developed through causal processes. The next part discusses how this could be implemented in a robot.

3 How could a robot be an agent with capacity for morally responsible behaviour?

There are two main insights from the theories presented in part two that are important in this section of the article. The first is that a mind with a free will can be understood as a causal process where an autobiographical self learns from experience and influences both itself and future behaviour through a continuous process. The second is that responsibility is the capacity to be influenced by being held responsible through a normal reasoning process. Could such a mind, free will, and responsibility be actualized in a robot, thus making it an agent with capacity for morally responsible behaviour? A closer look at how robots work implies that it could.

Russell and Norvig describe four basic types of artificially intelligent agents with different level of complexity, and these fit well with the different levels of involvement in choices described in part two of this article. The four types of artificial agents, in increasing degree of complexity, are simple reflex agents, model-based reflex agents, goal-based agents, and utility-based agents.[31] The *simple reflex agent* acts directly on input with a condition-action rule, "If A, then B". For example, a self-driving car could have the condition-action rule "If the car in front of you brakes, then brake".

The *model-based reflex agent* does not only respond automatically to input, but it additionally has a model of the world which says what happens if the agent chooses alternative A or B. *Goal-based agents* are even more advanced. They have models for what happens if the agent does A or B and goals that the alternatives can be compared with. *Utility-based agents* can also measure goals against each other and choose a goal based on maximal utility. Russell and Norvig write that this is like the artificial agents asking themselves how "happy" the different goals would make them.

In addition, there are learning agents where the agent learns based on an internal measurement of how well it predicted outcomes, or it can learn from an external standard of what is useful. The way that happens is that feedback is interpreted either as a reward or a penalty. All learning is about making the parts fit better together to increase the utility for the agent.

There are obvious parallels between artificially intelligent agents and the description of free will and responsibility in part two of the article. The simple reflex agent is like an action where desires are not involved at all, which would be like reflex actions in humans. Goal-based agents can represent alternatives to choose among, while utility-based

[30] Bok [5]

[31] The descriptions of the different agents are from Russell, Norvig, and Davis [34].

have desires of different strength, making them select the presumed best alternative. In addition, the agents can learn, based on feedback, in the same way as humans can develop more independent autobiographical selves, and in the same way as humans learn moral behaviour by being held responsible.

In the following there will first be a discussion of the degree to which robots can have free will and then responsibility. Could a robot implement a causal process with something corresponding to the autobiographical self, learning from experience and influencing both itself and future behaviour through a continuous process? The answer seems clearly to be yes since, as shown with the example from Russell and Norvig, we already have robots that learn from experience to be better at reaching their goals.

One kind of feedback they learn from is how well a result matches with an original goal, but another kind of feedback is whether humans respond positively or negatively to what the robot does (where the robot has positive human feedback as a goal). Coaching social robots is a process where robots act and get feedback from humans. It can be in a simple form, like humans saying the words "positive" or "negative" when the robot acts,[32] but there are also advanced robots that interpret the facial expressions of humans and adjust their behaviour based on that.[33] There are even robots with a kind of autobiographical memory designed to be similar to how it works in human brains: it stores important learned experiences which influence what are considered alternatives for future choices and what should be done with those choices.[34]

All of these examples still just work on specific goals, so there are no machines with general intelligence, and there are different ways that robots learn. They are getting better at more and more tasks. Most people working with AI think that it is a question of *when* and not *whether* machines will have general intelligence (although one can also assume that the most pessimistic people do not choose to work with AI). General intelligence means the ability to solve most types of complex problems.

Assuming that there can be machines that can learn on a general basis to alter their behaviour based on feedback, how should we understand the goals and desires of humans as opposed to goals and desires in machines? On the one hand, it seems that humans have conscious desires that cause them to act, which would then be lacking in non-conscious robots. On the other hand, there are many reasons to think that the conscious part of desires in humans does not play any causal role.

First, choices are commonly understood by neuroscientists as causal processes in the brain.[35] (An example of how to understand a choice as a causal process was given in the previous part of the article.) Second, it is hard to find a causal role for consciousness at all in our universe. The principle of causal closure says that every physical effect has a sufficient physical cause if it has a cause at all.[36] This principle is equivalent to the principle of conservation of energy insofar as there are no other known kinds of energy than what physics today recognizes.[37] The principle of causal closure is not proved, but it is a presupposition with much inductive evidence in its support. Further, it seems to lead to a whole host of problems to accept that consciousness has causal effects. It would be an unknown kind of causal influence which seems to break the principle of conservation of energy. How does it work? What is the causal nexus that combines non-physical consciousness with the physical world?

On the other hand, it is strange that evolution should have selected conscious brains if consciousness does *not* have a causal role. And yet evolution has selected several byproducts without causal effect, like the red colour of blood. The question of the causal role of consciousness is a huge debate which cannot be dealt with here.[38] Instead, the most common view will be presupposed, namely that choices come about as causal processes, even if our mental content, when we deliberate, is conscious. The process of desire and action happens in an if–then manner, which could be thought of as dispositions or algorithms saying how long to consider alternatives and when to act. The general structure of such a choice has been described above.[39] It seems quite clear that a robot could have representations for actions, alternatives and desires, and could follow algorithms for when to act how.

How should we then understand the fact that humans change their desires over time, that they change what they experience as good, and that they do not always act according to what feels best? Is it not an important part of human free will that we can also change our goals? How could this be implemented in a machine?

[32] Hirkoawa and Suzuki [16].

[33] Gruebler, Berenz, and Suzuki [14].

[34] Prescott et al. [32]. For an overview of different designs of attempted artificial moral agents, see Cervantes et al. [7].

[35] Schroeder, Roskies, and Nichols [36]; Damasio [38].

[36] Kim [23]; Papineau [29].

[37] Papineau [29], 55–57.

[38] Elsewhere, I have argued that consciousness has a causal role in shaping the evolution of brains even if it does not have a causal role in specific choices (Søvik [40], chapter 7). Consciousness is relevant for understanding why you have the kind of brain you have, but when you have this brain and make choices, the choice is caused by physical processes. To defend this claim takes much space and cannot be done here.

[39] For more detailed suggestions or evidence that this is how choices happen in the human brain, see Carruthers [6]; Adina Roskies [33].

These facts about humans can be understood as a causal process relating various desires to more fundamental desires. To simplify, let us say that there are two fundamental desires, to avoid pain and to experience good, and that there are physical realizers for our conscious experiences of good and bad. The autobiographical self is guided by what is experienced as good or bad, but it may change which concrete actions or experiences feel good or bad.

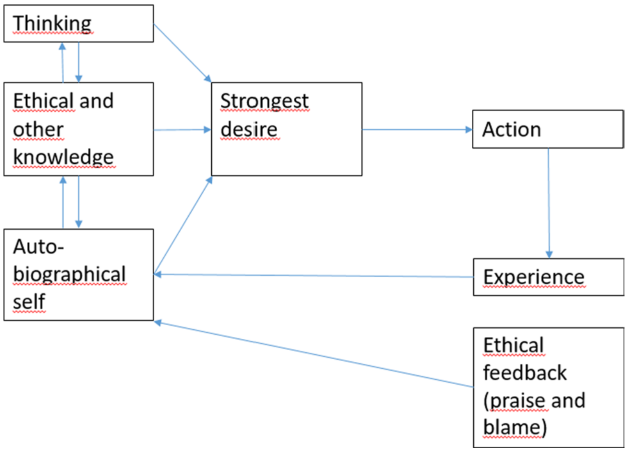
For example, eating candies can feel good and make you eat candies. But then you have a negative experience of getting fat and a good experience of eating less and feeling better, so the autobiographical self-changes to make eating candies less desired. Or reflecting upon and having a feeling of what would be a good world and a bad world can make you desire to act in ways that does not feel good in the moment, but which is experienced as a means to a good goal. The goal of a better world does feel good and is a stronger desire than to have a good feeling in the moment.

The details are probably different, but the main point is that it is quite plausible to think that normal causal processes in the brain make us act as we do. There are many examples of how putting electricity at certain parts of the brain can cause extremely strong desires (for example, for eating or having sex) and corresponding actions, indicating strongly the physical side of this.[40] We seem to have inherited certain desires and brain mechanisms through evolution, and it seems that machines could have been given similar functions. But could machines also be thought to have capacity for responsible behaviour? This question will occupy us in the following.

If a robot could learn from experiences, it seems that it could also learn from being held responsible (through praise and blame). When humans are able to be influenced by praise and blame through a normal reasoning process, we say that they have capacity for morally responsible behaviour. Not all humans have this capacity, but those are not held accountable for their actions either. They might be confined if they are dangerous, but not as disciplinary punishment. Instead, using the same logic as when people are quarantined, it is for reasons of safety.

It seems that there could be robots that take praise and blame into account when acting. We could make a robot follow two general algorithms: "If it feels good, then do it" and "if it feels bad, then avoid doing it", and add two more algorithms saying "if humans praise you, it feels good" and "if humans blame you, it feels bad" (and better or worse depending on how many blame or praise how long and how strongly). To make this work, more detailed algorithms would have to be added for what to do in conflicts and how many and who should be praising and blaming, but here we will just discuss the general idea at a higher level of abstraction.

It also seems that human reasoning is guided by what we experience as good and bad and that praise and blame are experiences of good and bad that we take into account in our reasoning. In addition, we reason about what are means to goals and how good goals are, which can make us act in ways that do not feel good in the moment but which are means to the goal. It seems that machines could do the same as long as we assume that it is the physical actualizers and not the conscious experience that causes the actions. Here is a model to describe how it works in humans and could work in robots:


Thinking
Ethical and other knowledge
Strongest desire
Action
Auto-biographical self
Experience
Ethical feedback (praise and blame)

We will end this section with a discussion of the implications that follow from robots not being conscious. Maybe robots will become conscious in the future. The content of their consciousness will then probably be very different from ours, since our conscious content is so strongly shaped by what is evolutionarily advantageous for us. But only non-conscious robots will be discussed here.

Above, choices were assumed to be physical processes not needing consciousness, but consciousness seems to have other roles. If one does not think that consciousness can be reduced to physical processes, consciousness seems to be what makes possible a unified subject with phenomenal experiences of intentional thoughts and emotions that feel good and bad. This is important since it seems that consciousness is a necessary condition for some of the components needed for having capacity for responsible behaviour. I have here grouped the most important ones together as follows:

First, consciousness seems necessary for there to be a unified subject which can be ascribed free will and

[40] Rhawn Joseph [19].

responsibility.[41] A machine is just bits and pieces with no subject or *I* anywhere being the subject that has free will or responsibility.[42]

Second, consciousness seems necessary for intentionality (i.e., to have thoughts that are about something), for how can electrical signals *be about* something? If a machine cannot have thoughts about actions or intentions about how one should act or understand reasons for acting, it does not understand what it is doing.[43] Is it then acting at all, or are actions, intentions and reasons constituted partly by intentionality? At least several actions seem constituted by a conscious aspect, like for example wanting to afflict pain on others or being afraid of feeling pain yourself.[44]

Third, consciousness seems necessary to experience feelings—not only praise and blame, but all feelings—which again seem to constitute the goodness and badness of many actions. Can robots be said to know what they are doing if they do not know what things feel like? Does it make sense to praise, blame or punish them at all? And if not, what does it mean to hold them responsible?

The different points here all refer to important parts of the normal human process of holding each other responsible, which again makes them seem like important objections. On the other hand, neuroscience teaches us that humans can do almost anything non-consciously, including advanced reasoning. A classic example is driving a car "on automatic pilot". You can drive the car for a long time talking to someone sitting beside you and not think about driving. Nevertheless, you have clearly processed a lot of information and adjusted your driving accordingly, even if you have not been thinking consciously about it.[45] Blindsight is another fascinating example, where people with no conscious experience of seeing can move through a labyrinth, catch what is thrown to them, and so on.[46]

We can also not only sense but also non-consciously think. Damasio and colleagues did an experiment where people should a card from one of two decks depending on a clue they were given. There was a system in the clues, which means that when you cracked the code you could draw correct cards all the time. Test subjects played this game while Damasio and the team measured their skin conductance. After a while, they noticed that the test persons would start sweating just before drawing cards from the wrong deck, but not when drawing from the right deck. It became clear that the test persons non-consciously cracked the code several minutes before they consciously solved the problem.[47]

It thus seems like the process of holding others responsible is something that could work at a non-conscious level also in humans. Consider the following example: It seems very plausible that a man could forget to put down the toilet seat without consciously making a choice to leave it up. It seems very plausible that his wife could blame him for not putting down the toilet seat. It also seems plausible that he could non-consciously remember to put it down later, as a result of non-conscious reasoning. Many similar examples could be given, like closing the door or saving the last piece of cake for someone else.

In these examples, the actions are done non-consciously and influenced by blame, but the receipt of blame itself happened in a conscious process. However, we know that we can register facts non-consciously as well, so the fact that an action of mine annoyed someone else is something that could possibly be registered non-consciously (e.g., somebody else's body language when I did something inappropriate). In these cases, there are actions we usually think of as conscious actions that happen without conscious intentionality or a conscious subject consciously choosing to act, and whether they are done consciously or not makes no difference. It seems like all parts of the process of holding someone responsible and taking into account that one is held responsible could happen non-consciously, such as when registering a traffic sign non-consciously and adjusting to it.

However, one could argue that *sometimes* the person must be conscious, even if he or she sometimes acts non-consciously. It would be different if the person was never conscious, never having understood that something was good or bad. I agree that there is a difference when it comes to describing what actually happens, so I think that it is clarifying to distinguish between conscious and non-conscious free will and conscious and non-conscious responsibility. In the non-conscious version there is no subjectivity, intentionality or understanding in the intentional sense, but it is physical processes in humans that cause the content of the conscious experience and human actions. The process and goal of holding others responsible can thus work the same way for humans and machines, but humans understand in the intentional sense what is going on, even if there are physical processes at work in both cases.

[41] There are different ways of arguing this. For a good account of how subjects are constituted by consciousness, see Coleman [9].

[42] When Floridi answers objections against non-conscious robots, he does not mention this objection (Floridi [12]). Similarly, when Behdadi and Munthe argue that internal mental properties can be understood as dispositions, they do not mention subjectivity and intentionality, which seem to be first-person perspective phenomena that cannot be translated into something physical (Behdadi and Munthe [4]).

[43] The classical argument for this is Searle [37].

[44] Floridi argues that intentions are not important given utilitarianism (Floridi [12]), but most utilitarians will say that your motive should be that which gives the best consequences.

[45] Armstrong [1].

[46] Carruthers [6], 87–88.

[47] Damasio [11], 276. Bechara et al. [3].

This is then the precise difference between a conscious and a non-conscious agent: both can make free and responsible choices since they can reason about means to goals and learn from feedback, but the non-conscious agent does it without subjectivity, intentionality and understanding in the intentional sense. Lacking consciousness, robots do not have a subjective experience of an action and its consequences, which also means that feedback does not feel either good or bad for robots.

Since there are these differences between conscious and non-conscious agents, it is good to distinguish between conscious and non-conscious versions of the terms involved and be clear about what the difference is. I have tried to describe in detail some similarities and differences between conscious humans and non-conscious robots when it comes to free will and responsibility, but we must also consider an important difference when it comes to praise, blame and punishment.

As mentioned, responsibility comes in degrees. Humans are all different and raised under different conditions, but also have different degrees of free will and responsibility. Robots are different, as well, and could have different degrees of free will and responsibility. Just as humans can be placed on a scale from no or little to much capacity for responsible behaviour, robots could be placed along a similar scale.

But when do robots have sufficient capacity for responsible behaviour so that they could be called morally responsible? Most countries say that persons must be more than 15 years old before they can be punished, since they need time to learn and internalize what is right and wrong. Robots would also have to be intelligent enough to know the difference between right and wrong in most new situations they run into to have sufficient capacity for morally responsible behaviour. An important reason to establish sufficient capacity for morally responsible behaviour for humans is to find out when it is ethically acceptable to punish someone for their actions. But many have argued that it does not make sense to punish robots that do something wrong since they cannot consciously feel pain.[48]

There are many different reasons why we punish someone. One is to influence behaviour, and in milder variants (like blaming) we do this from early childhood on to teach children how to act. “Punishing” a robot to influence behaviour is thus not morally problematic even if it does not have sufficient capacity for responsible behaviour. It could just be thought of as a part of the training. “Punishing” a robot would then just to be to give it feedback showing that it did something wrong, which makes it change behaviour. It may be more confusing than clarifying to use the word “punish” in the robot context, so “correcting” or something else would probably be better.

It could seem that you cannot punish a robot since it does not have conscious feelings, but punishing merely in the sense of influencing behaviour (which is just a small part of what punishment is) would be a response that the robot registers as negative feedback and takes into account as a reason to change its behaviour. As discussed above on the causal role of consciousness, the conscious feeling of pain does not seem to be the actual cause of a behavioural change in humans either.

But there are other reasons that we punish which seem to make sense only if robots could have a conscious feeling of pain. We want to recognize those who have suffered pain by inflicting pain on a wrongdoer, which the wrongdoer seems to deserve in order that justice should be done. Does any of this make sense in the case of robots if they cannot have a conscious experience of frustrated desires?

Victims of a robot action could probably feel some recognition and restoration if a wrongdoing robot had to give them money or work for them or be destroyed—something which was not a goal for the robot but is in the interest of the victim. Basic desert views on responsibility hold that there is a fundamental relation between actions and responsible persons that makes wrongdoers deserve pain even if it serves no other goal, but I reject this view, leaning instead on Vargas’ theory of responsibility.

Again, there are similarities and differences in blame and punishment for humans and robots, so again it would be clarifying to distinguish between a conscious and a non-conscious version, and most important is to know exactly what the terms imply. In this article I have unpacked a theory of how some of these similarities and differences should be understood.

We can hold both humans and robots responsible with the common goal of influencing their behaviour, but these important similarities and differences between non-conscious robots and conscious humans should make us hold them responsible in different ways.

4 Objections

Four remaining objections will be answered in this part of the article. The first objection is that it does not matter if a machine has internal properties similar to humans if they have been placed there by a designer. The machine will then not have the same independence as we do.[49] Hakli and Mäkelä use Alfred Mele to press the point that history matters: if a person suddenly had all specific desires put

[48] Sparrow [41].

[49] Johnson [18].

into her brain, she would not be responsible for her will and actions.[50] The argument could be pressed further with the zygote argument by Alfred Mele: if our universe is deterministic and a goddess designs a zygote and inserts it into our universe to ensure that it will be born and on a specific day kill a specific person, this is not a free person.[51]

This objection is easily answered given the theory of free will presented here, since humans start as unfree and create their own freedom and responsibility through self-formation in an indeterministic world. The zygote case is impossible in an indetermined world, and machines starting as unfree when designed can become free by acting based on indetermined experiences and learning. Machines start without freedom or responsibility, but will gradually have more responsibility as they change themselves from within. Since the point of responsibility is to cause positive change we would blame the designer in the start, but when the machine starts changing itself to behave in new ways, we must praise and blame the machine instead of the designer. Different machines will have different constraints, but the same applies to humans. Humans have free will and responsibility to different degrees, and machines will have even greater variation.

The second objection considered here is that there could develop responsibility gaps if machines are able to cause great harm but only have a certain degree of responsibility. If they do something that the designer did not cause, it seems that harmful events can be left with nobody responsible. Andreas Matthias raised this objection not to argue that there are no responsibility gaps, i.e., situations where nobody is responsible for what happened, but to argue that the gaps will widen.[52]

I agree with Matthias that there may be harmful events where nobody can be selected as the cause of the event since there were many people causing many parts of the process without knowing what the result would be. The way to deal with moral responsibility in all cases is to determine it by asking what we think a person (or machine), given their resources, should have done in the relevant situation. That both leaves no gaps in the theory of and also determines moral responsibility. It may leave a gap in the sense that something bad happened that nobody can be blamed for, but if so, that means that it is true that something bad happened that nobody can be blamed for. It is like a hurricane caused by global warming: no individual can be blamed for the hurricane, but many individuals can be blamed for polluting.

Given an understanding of responsibility where such is determined by comparing people's action to a standard for what they should have done, such as presented in part two, there will not be a big problem of responsibility gaps. It is not the case that many evils happen with nobody to blame. The leaders of the few most powerful countries are the ones that could have secured international agreements to end climate change and prevent the hurricanes that are caused by an unnaturally and unnecessarily warm climate. They should have prevented it.

Autonomous weapons are another good example. Champagne and Tonkens argue that army generals can take on a blank check responsibility for what autonomous weapons do.[53] In my view, responsibility depends on what we think people should do given their resources to actualize the best world. We should probably not make weapons that are not controlled by humans, which means that the makers of such weapons should be held responsible for them. Or if they should be made, then people are responsible for acting in such a way that the result becomes the best.

The third objection considered is that there is an important difference between humans and machines: very intelligent machines will probably be able to change all of their own algorithms, which is not something we humans can do (yet). We cannot choose to change what feels good and bad for us, but very intelligent non-conscious machines probably could, so it seems that the idea of implementing learning through praise and blame would not work the same way in machines as it does in humans. Humans build a stable moral character over time by being held responsible, whereas a machine can turn into something else completely in no time. Does the process of holding it responsible then make sense?

To consider this problem, imagine that humans could change, at will, what feels good and bad for them. What would humans then do if they could change what feels good, and what if they could even change the basic mechanism that we want to do what feels good? Imagine that, at will, you could decide to let anything motivate your behaviour. What would you do?

Whatever your motive would be for changing would be something that already motivates you now. If you already have an overriding motivation—for example, doing what feels good—you have no motive for changing it, like for example saying that from now on I will only do what makes me scratch, or makes me sing the national anthem, or makes me turn left. Turning to the robot, if there is an overriding algorithm guiding the behaviour of the robot, the robot would not have a motive for changing it.

Nevertheless, it could happen, by some chance event or accident, that a very intelligent machine got a morally bad motivation (like how a mutation might make a person not

[50] Hakli and Mäkelä [15].

[51] Mele [26], 188–89.

[52] Matthias [24].

[53] Champagne and Tonkens [8].

want to do what feels good), so the risk that something might go wrong is, of course, always there.

5 Conclusion

This article has defended an event-causal libertarian theory of free will and responsibility, describing in detail how the process of developing free will and responsibility works. The mind is an event-causal process that can be influenced by being held responsible. Non-conscious robots can be influenced in very similar ways, learning from positive and negative feedback. Even if humans are conscious, non-conscious brain processes do the causal work.

There are important differences that follow from lack of consciousness. Robots do not have the unified first-person experience of being a subject, which is made possible by consciousness, which means that they do not experience feelings connected to praise and blame. Their process of developing free will and being responsible is different from humans, but nevertheless the main structures are the same.

We should distinguish between the conscious and non-conscious versions of free will, responsibility, praise and blame. In this article, the similarities and differences have been specified. Even if different, non-conscious free will and responsibility deserve to be called free will and responsibility because the main structures are the same as in the conscious versions. The libertarian form of free will and responsibility which has been defended means that non-conscious robots could have a stronger form of free will and responsibility than what is commonly defended in the literature on robot responsibility.

Funding Open access funding provided by MF - Norwegian School Of Theology, Religion And Society.

Declarations

Conflict of interest The author declares none.

References

1. Armstrong, D.M.: The Nature of Mind, and Other Essays. Cornell University Press, Ithaca, NY (1981)
2. Baumeister, R.F., Masicampo, E.J., DeWall, C.N.: Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. Pers. Soc. Psychol Bull. **35**(2), 260–268 (2009)
3. Bechara, A., Damasio, A.R., Damasio, H., Anderson, S.W.: Insensitivity to future consequences following damage to human prefrontal cortex. Cognition **50**(1–3), 7–15 (1993)
4. Behdadi, D., Munthe, C.: A normative approach to artificial moral agency. Mind. Mach. **30**(2), 195–218 (2020)
5. Bok, H.: Freedom and Responsibility. Princeton University Press, Princeton, NJ (1998)
6. Carruthers, P.: The Architecture of the Mind: Massive Modularity and the Flexibility of Thought. Oxford University Press, Oxford (2006)
7. Cervantes, J.-A., Ruiz, S.L., Rodríguez, L.-F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: A survey of the current status. Sci. Eng. Ethics **26**, 501–503 (2020)
8. Champagne, M., Tonkens, R.: Bridging the responsibility gap in automated warfare. Philos. Technol. **28**, 125–137 (2015)
9. Coleman, S.: Mental Chemistry: Combination for Panpsychists. Dialectica **66**(1), 137–166 (2012)
10. Coleman, S.: Panpsychism and Neutral Monism: How to Make up One's Mind. In: Brüntrup, J., Godehard, L. (eds.) Panpsychism. Oxford University Press, Oxford (2016)
11. Damasio, A.R.: Self Comes to Mind: Constructing the Conscious Brain. Pantheon Books, New York (2010)
12. Floridi, L.: Artificial Agents and Their Moral Nature. In: Kroes, P. and Verbeek, P.-P. (eds.) The Moral Status of Technical Artefacts, pp. 185–212. Springer, Dordrecht (2014)
13. Ginet, C.: On Action. Cambridge Studies in Philosophy. Cambridge University Press, Cambridge (1990)
14. Gruebler, A., Berenz, V., Suzuki, K.: Coaching Robot Behavior Using Continuous Physiological Affective Feedback. Paper presented at the 2011 11th IEEE-RAS International Conference on Humanoid Robots, 26–28 Oct. (2011)
15. Hakli, R., Mäkelä, P.: Moral Responsibility of Robots and Hybrid Agents. Monist **102**(2), 259–275 (2019)
16. Hirkoawa, M., Suzuki, K.: Coaching Robots: Online Behavior Learning from Human Subjective Feedback. In: Jordanov, I., Jain, L.C. (eds.) Innovations in Intelligent Machines 3: Contemporary Achievements in Intelligent Systems, pp. 37–51. Springer, Berlin (2013)
17. Hodgson, D.: Quantum Physics, Consciousness, and Free Will. Chap. 3 In: Kane, R. (ed.) The Oxford Handbook of Free Will, pp. 57–83. Oxford University Press, Oxford (2011)
18. Johnson, D.G.: Computer systems: moral entities but not moral agents. Ethics Inf. Technol. **8**(4), 195–204 (2006)
19. Joseph, R.: Neuropsychiatry, Neuropsychology, and Clinical Neuroscience: Emotion, Evolution, Cognition, Language, Memory, Brain Damage, and Abnormal Behavior, 2nd edn. Williams & Wilkins, Baltimore (1996)
20. Kane, R.: The Oxford Handbook of Free Will, 2nd edn. Oxford University Press, Oxford (2011)
21. Kane, R.: The Significance of Free Will. Oxford University Press, New York (1996)
22. Kim, Jaegwon. *Philosophy of Mind.* 3rd ed.: Routledge, 2011.
23. Kim, J.: Philosophy of Mind, 2nd edn. Westview Press, Boulder, CO (2006)
24. Matthias, A.: The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. Ethics Inf. Technol. **6**(3), 175–183 (2004)

25. Mele, A.R.: Agents: From Self-Control to Autonomy. Oxford University Press, New York (1995)
26. Mele, A.R.: Free Will and Luck. Oxford University Press, New York (2006)
27. Nagel, T.: What Is It Like to Be a Bat? Philos. Rev. **83**(4), 435–450 (1974)
28. O'Connor, T.: Agent-Causal Theories of Freedom. Chap. 15 In: Kane, R. (ed.) The Oxford Handbook of Free Will, pp. 309–28. Oxford University Press, Oxford (2011)
29. Papineau, D.: The Causal Closure of the Physical and Naturalism. Chap. 2 In: McLaughlin, B.P., Walter, S., Beckermann, A. (eds.) The Oxford Handbook of Philosophy of Mind, pp. 53–65. Oxford Handbooks in Philosophy, Clarendon, Oxford (2009)
30. Pereboom, D.: Free will, agency, and meaning in life. Oxford University Press, New York (2014)
31. Pereboom, D.: Hard Incompatibilism. Chap. 3 In: Fischer, J.M, Kane, R., Pereboom, D., Vargas, M. (eds.) Four Views on Free Will, pp. 85–125. Blackwell Publishing, Oxford (2007)
32. Prescott, T.J., Camilleri, D., Martinez-Hernandez, U., Damianou, A., Lawrence, N.D.: Memory and Mental Time Travel in Humans and Social Robots. Philos. Trans. R. Soc B: Biol. Sci. **374**(1771), 20180025 (2019)
33. Roskies, A.: Monkey Decision-Making as a Model System for Human Decision-Making. Chap. 12 In: Mele, A.R. (ed.) Surrounding Free Will: Philosophy, Psychology, Neuroscience, pp. 231–54. Oxford University Press, New York (2014)
34. Russell, S.J., Norvig, P., Davis, E.: Artificial Intelligence: A Modern Approach. Prentice Hall Series in Artificial Intelligence. 3rd ed. Prentice Hall, Upper Saddle River (2010)
35. Schaffer, J.: Contrastive Causation. Philos. Rev. **114**(3), 327–358 (2005)
36. Schroeder, T., Roskies, A.L., Nichols, S.: Moral Motivation. In: Doris, J. (ed.) Moral Psychology Handbook, pp. 72–110. Oxford University Press, Oxford (2010)
37. Searle, J.R.: Minds, Brains, and Programs. Behav. Brain Sci. **3**, 417–457 (1980)
38. Singer, W.: Verschaltungen legen uns fest: Wir sollten aufhören, von Freiheit zu sprechen." In Geyer, C.(ed.) Hirnforschung und Willensfreihet. Zur Deutung der seuesten Experimente, pp. 30–65. Suhrkamp, Frankfurt am Main (2004)
39. Søvik, A.O.: Free Will, Causality and the Self. DeGruyter, Berlin (2016)
40. Søvik, A.O.: A Basic Theory of Everything. A Fundamental Framework for Philosophy and Science. DeGruyter, Berlin (2022)
41. Sparrow, R.: Killer robots. J. Appl. Philos. **24**, 62–77 (2007)
42. Sripada, C.: Self-Expression: a deep self theory of moral responsibility. Philos. Stud. **173**(5), 1203–1232 (2016)
43. Strawson, G.: The impossibility of moral responsibility. Philos. Stud. **75**(1/2), 5–24 (1994)
44. Sullins, J.P.: When Is a Robot a Moral Agent. Int. Rev. Inf. Ethics **6**(12), 23–30 (2006)
45. van Inwagen, P. An Essay on Free Will. Clarendon Press, Oxford (1983)
46. Vargas, M.: Building Better Beings: A Theory of Moral Responsibility. Oxford University Press, Oxford (2013)
47. Vohs, K.D., Schooler, J.S.: The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating. Psychol. Sci. **19**(1), 49–54 (2008)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.