AI and Ethics
https://doi.org/10.1007/s43681-023-00285-6

ORIGINAL RESEARCH

Intelligent machines, collectives, and moral responsibility

Simon Courtenage[1]

Received: 8 February 2023 / Accepted: 28 March 2023
© The Author(s) 2023

Abstract
Collectives, such as companies, are generally thought to be moral agents and hence capable of being held responsible for what they do. If collectives, being non-human, can be ascribed moral responsibility, then can we do the same for machines? Is it equally the case that machines, particularly intelligent machines, can be held morally responsible for what they choose to do? I consider the conditions required for moral responsibility, and argue that, in terms of the agency condition, artificial, non-human entities in general are excused from being responsible because, although they may choose their actions, the beliefs and desires that form the basis of their choices are predetermined by their designers, placing them in an analogous position to persons suffering covert manipulation. This creates a problem for collective responsibility, but I argue that collectives, through their supervention on human persons, represent an exception. Finally, I consider that the design of future machines may be sufficiently abstract and high-level as to fall below some threshold of influence, allowing machines enough freedom for us to hold them responsible.

Keywords Artificial intelligence · Machine intelligence · Moral responsibility · Moral agency · Collective agency

1 Introduction

In his novel, *Klara and the Sun*, Kazuo Ishiguro explores the complex emotional and moral relationships that may come to exist between human beings and intelligent machines. Klara, a humanoid robot with advanced artificial intelligence, is bought by the mother of a girl with a suspected fatal illness to be her companion. Klara carries out her role conscientiously and with great care. Her intelligence permits her considerable autonomy in interpreting and carrying out her duties.

Klara comes to believe that a particular construction machine is responsible for the illness of the girl in her care, and that if the machine is destroyed, the girl will recover. However, the only means she has to destroy it is to give up part of her internal fluid, which is vital to her operation. Klara's choice, therefore, is either to weaken herself by giving up some of her internal fluid to destroy the machine, but which will diminish her ability to care for the girl, or to carry out her duty of care undimmed but leave the machine she holds responsible for the girl's illness to continue operating.

Ishiguro's novel is a novelistic exploration of how we might react to the presence of intelligent but artificially created persons in our midst. Nonetheless, it sensitively portrays the kind of dilemmas concerning ascriptions of moral responsibility that we face, and will face more often, as machines[1] with some form of intelligence become more and more common in our everyday lives. Whichever way she chooses, Klara causes harm to come about, either to herself, and by extension to the girl in her care, or to the machine and its owners. Who, though, is responsible for these harms? Ishiguro does not answer this question, but rather invites us

✉ Simon Courtenage
courtes@westminster.ac.uk

[1] School of Computer Science and Engineering, University of Westminster, London W1W 6UW, UK

[1] I will use "machines" and "intelligent machines" to generally and equally refer to machines with some (unspecified) form of decision-making ability, and in a fairly informal way without specifying whether it refers to machines that are purely hardware, purely software or a mix of both.

Published online: 18 April 2023

to consider our own emotional response to Klara as we see ourselves and the world through her eyes.

In this paper, I consider whether or not intelligent machines can be held morally responsible[2] for their actions. This question is an important one. We are increasingly subject to the decisions and actions of machines with some form of intelligence. Self-driving cars, autonomous war robots, machine learning applications for border control, promotion at work, and surveillance—these are all examples of machines that have the potential to cause us significant harm through the actions they choose.

The question of the moral responsibility of machines is also a difficult one, in large part because what it means to be morally responsible, or what it is that qualifies someone or something to be responsible, resists easy definition. We are clear that adult human beings of normal disposition and upbringing can be considered morally responsible, but what it is that makes them so is not so clear. Different aspects of adult humans such as their subjective experience of pleasure and pain, their rationality, their consciousness and self-awareness, for example, have all been used as the basis for arguments for their moral responsibility, and for denying the same to intelligent machines. Whether or not machines can be morally responsible, therefore, can sometimes appear to be simply a matter of what characteristic it is thought that responsibility is based on.

I will approach the question of the moral responsibility of machines differently. Instead of arguing for one or another characteristic of human persons as necessary for moral responsibility and that this is possible or not for machines, I will examine whether or not machines can be related in some way to other non-human and artificial entities that are generally considered to have moral responsibility. Collectives, such as corporations, are generally accepted to be, under certain conditions, responsible for their actions. Therefore, the specific question I will examine is whether or not machines can be related to suitably organised collectives, such that if collectives can be considered morally responsible, then machines can be too. I will concentrate on that requirement for moral responsibility that a person should be able to make free and voluntary choices about what do, and argue that artificial entities, such as collectives and machines, experience prior influence on their ability to choose in the form of their design, analogous to covert manipulation, which excuses them from responsibility. I will further argue that collectives, by their supervention on their human members in relation to their ability to choose, are sufficiently different to machines that they are able to overcome this influence, and can therefore be considered morally responsible, whereas machines cannot. However, although I argue design of an artificial entity to represent a form of covert manipulation, I will concede that a sufficiently weak design, if that is possible for formulate and distinguish, may allow sufficient freedom to admit responsibility.

My argument, in outline, will be as follows. First, I will put forward the conditions for moral responsibility and present the case that collectives satisfy these conditions. I will then consider the same case for intelligent machines, but follow that case with the objection that artificial entities cannot satisfy the control condition for responsibility because their design represents a form of covert manipulation. While this objection holds for machines, I further argue that collectives are able to retain responsibility because of their human membership. After considering objections to this account, I end with the conclusion that, even though both collectives and intelligent machines are forms of non-human, artificial entities, the moral responsibility of collectives does not allow us to infer the same for machines, except in the case of weak design.

2 Conditions of moral responsibility and collectives

Conditions of moral responsibility Holding someone morally responsible means the satisfaction of the following three conditions:

1. that they are in control of their actions (the ‘control condition’),
2. that they aim at performing that action (the ‘intentionality condition’), and
3. that they know what they are doing and what will follow from their actions (the ‘knowledge condition’).

Satisfying the control condition means that an action is sufficiently within the control of the person that, counterfactually, it would not have occurred but for their exercise of free will to choose it and their capacity to carry out their will. Not only is the action actually available to them but it occurs from the free exercise of the person’s will. I might choose between spending money to buy a present for my wife or spending the same money on myself. Both are actions I can perform, in the sense that they are actually available to me, and, so long as my choice between them is free of undue influence, then the control condition is satisfied. On the other hand, coercion and manipulation of a person, so that they choose an action they would not have freely chosen themselves, generally exempt them from responsibility. Such exemptions I will call *negative excusing exceptions*.

[2] As will be made clear in the next section, I take moral responsibility to mean that an event can be assigned to a person’s agency such that they can be praised or blamed for the event.

The second condition is the intentionality condition. We are not considered responsible for things that happen by accident or for those things that randomly occur. If I am holding a cup of coffee and someone not looking where they are going bumps into me, making me spill the coffee on an expensive carpet, I am not usually considered responsible for the damage caused, since the action does not follow from an intention to bring it about. However, if I deliberately tip the coffee on the carpet to damage it, in other words, if this is something I do intentionally or that I aim at, then I am responsible.

Related to the intentionality condition is the knowledge condition. A person who is morally responsible for the action they performed is taken to know what harms would follow from their action, by which I mean not only that they know what effect would causally follow from an action but also that the effect would be a harm. Moreover, they know of the concept of harm itself and are able to evaluate an effect as a harm.

If any of these responsibility conditions are unsatisfied in a person, then this affects our consideration of their moral responsibility. If the control condition is unsatisfied, for example, in a person who is coerced or blackmailed into doing something harmful, then they are not held to be responsible. They lack control over their actions, in the sense that the action did not occur as a result of an exercise of their free volition. Nor is a child, on the ground of their general lack of knowledge or awareness. A child may freely and intentionally perform some action, but they are not held responsible on the grounds that they lack full knowledge of the concept of harm or of the particular harm that follows from their action.

There are alternative formulations of these conditions. Fischer and Ravizza, for example, state them as the “freedom-relevant condition”, which concerns control, and the “cognitive condition”, which includes both the matter of what is aimed at and what is known.[3] As Fischer and Ravizza point out, each of these conditions (in their formulation) “corresponds roughly to … negative excusing conditions” for saying someone is not responsible; either they were not in control of what happened, or it was not what they were aiming at, or they acted in ignorance.[4] A similar formulation is also described by Pettit, as *value sensitivity*, having the control necessary to choose between actions, *value relevance*, being an autonomous agent who may face making a choice of action that may inflict harm or do good, and *value judgement*, having understanding and evidence necessary to make judgements.[5]

The Moral Responsibility of Collectives Human persons do not always act alone and by an individual act of will, nor do they always make decisions or choices alone or separately from others about what they will do. In our highly organised and structured societies, it is common for people to decide and act in concert in pursuit of some common goal that, typically, is best achieved or can only be achieved by working together. For example, the board of trustees of a charity might decide on a new fundraising campaign, the senior management of a company might sign off on investment in a new factory, or the government of a nation state might declare war against another state.

Just as we ascribe moral responsibility to an individual human person for what they choose to do and the benefits or harms that follow from their choices, we wish also to do the same in many cases when people act together, as a collective or group, in pursuit of some collective aim. If an chemicals company dumps untreated waste into a river which poisons the river and kills its wildlife, we want someone or something to be responsible and to be held accountable. However, while we usually find it easy to decide who is responsible in the case of an individual human being, it is not so easy when dealing with a collective. If an individual to whom you have lent money reneges on the debt, you know who to blame, but if a company decides to dispose of chemical waste in a river to save money, where does responsibility lie?

In the case of the company that cuts costs at the expense of the environment, one possibility is to lay responsibility on the chief executive or some particular person within the organisation. This might be appropriate if the person identified had sole control, or dictated to others in the organisation what they were to do such that they had no choice but to obey. But in those cases where decisions are taken collectively or require, as a necessity, some form of collective assent before an action can be taken, we are unable to locate moral responsibility in a single person. Moreover, if those individuals are bound by policy or statute in some form to make decisions in a particular way or towards a particular aim, then we are also unable to hold them morally responsible, singly or even jointly.

If we assume, then, that collectives such as corporations are morally responsible, how do they satisfy the responsibility conditions?

Condition 1, the control condition, states that an entity must have free will to choose an action and the capacity to carry it out. If a collective is to meet this condition, it must be able to (i) bring before itself those choices that are possible, (ii) evaluate them according to some criteria and make a choice on the basis of that evaluation, and (iii) execute that choice. Moreover, it must do so in a way that is not reducible solely to its human members.

Group theorists, such as French, List and Pettit, and Hess consider that some collectives are able to meet the

[3] Fischer and Ravizza [5], Introduction, pg. 8.

[4] Fischer and Ravizza, ibid., pg. 8.

[5] Pettit [22].

control condition due to their organisational structures and schemes. French, for example, argues that corporations are moral agents due to their Corporate Internal Decision (CID) Structure, which is composed of an organisational flowchart laying out roles, relationships and responsibilities, and decision procedures in the form of policies that describe how decisions should be made.[6] List and Pettit, in a similar vein, argue that formal and informal voting procedures result in a supervention of group or collective agency over human members of the group.[7] Essentially, some collectives embody free agency in the sense that they harness the free agency of their human members within their structure. The means by which this occurs, I take as being through, following List and Pettit, a supervention relation, resulting in an agent that is autonomous of its members and which is able to bring before itself possible options, evaluate them and make a choice, and then put it into effect.

A collective, such as a university or a company, therefore, can satisfy the control condition provided it has a suitable structure. Whether any particular collective does have a structure that allows the condition to be met will, however, be an empirical question.

Condition 2, the intentionality condition, requires that collectives are intentional in the sense of having representational mental states and for those representational states to be "about something" in the world. They have beliefs about the world, desires about how they want to the world to be, and the capacity to act on those beliefs to achieve those desires. To meet this condition, a collective does not need to have control over what its choices.[8] It needs only to be able to represent the current state of the world, a goal representing the desired state of the world, and knowledge of what actions will achieve that goal. As it acts, the collective is able to amend its representational states in light of the changes it has made to the world.

Provided it has a suitable structure, for example, policies, organisational hierarchies and roles, mission statements, budgets, and operational plans, a collective, such as a company, can meet the intentionality condition. As Hess argues, the structure of corporate agents embodies "a logically coherent set of commitments about fact and value—about how the world is and what matters in it—that reliably guides action".[9]

To meet Condition 3, collectives must be aware both of the concept of harm and of the harm that follows from the actions they pursue. The company that discharges chemical waste into a river, causing environmental damage, satisfies the knowledge responsibility condition if it knows of the concept of harm and knows that their action of dumping the waste causes harm. Knowing, in this latter sense, I take to mean that a collective can evaluate the relative values of the choices they face, choose between them on the basis of these values, and moreover, that these values incorporate what it means to harm. It is not difficult to see that the company dumping waste is capable of evaluating the choice of doing so against the choice of paying for a sustainable disposal of its waste but at increased cost, that this evaluation can incorporate a notion of harm, and that the company is capable of deciding between different options taking into account this evaluation. Suitably organised collectives will codify notions of harm into policies and protocols drawn up by their human members, and represent their evaluation mechanisms as voting or decision procedures for committees or sub-groups within the collective. Hence, although the collective is animated by its human members and supervenes on them, its concept of harm, the evaluation of different actions to understand what harms might follow, as well as the procedure for choosing between them based on that evaluation, belong to the collective itself.[10]

3 The parallel between collectives and intelligent machines

Intelligent machines I will use the term "machine" to refer to those things humans craft and use and that are able to operate on their own for some, possibly indefinite, period before human intervention becomes necessary, and which are created to achieve some human-conceived goal. This is not a perfect definition—it will always be possible to find counterexamples—but, hopefully, its intent is clear. Machines are things we bring into existence in order to meet our needs in a way that is, to some degree, independent of us.

In the context of machines, moral responsibility only really becomes a question when a machine possesses some form of intelligence or appears as intelligent to us. We do not think to ask if a machine can possibly be morally responsible if it is a steam train, a trouser press, or a piece of word processing software running on a laptop. But the question does become relevant if a machine has the capacity to make decisions, particularly if those decisions are made in a way that we have not clearly specified in advance or that we cannot predict because there is some uncertainty that we require the machine to resolve by itself. For example, a machine that

[6] French [9].

[7] List and Pettot [18].

[8] French's notion of a corporation as a "Davidsonian agent" [9] includes intending to act as well as intentionality as I use it here, so would meet both the control and the intentionality condition.

[9] Hess [12].

[10] Pettit makes a similar point (Pettit, ibid., Section III) about group agents satisfying his value judgement condition for moral responsibility, and the interaction between a group's organisation and its animation by human members who satisfy the condition in their own lives.

decides who should receive a social security benefit and who should not, or who should be shortlisted for promotion in a company.

Not all decision-making qualifies a machine to be classed as intelligent. A thermostat, for example, makes decisions about when to turn the heating on based on the current and desired temperatures. But it is not considered intelligent. Thermostats are an example of quite the opposite, so-called "dumb" objects, "dumb" because they "blindly" follow precise rules. Machines start to appear intelligent when the goals they are given require qualitative decision-making or decisions involving some degree of uncertainty. For example, an intelligent machine is able to decide what is the "best" way or the "best" time to do something.

It might be thought that I am specifically and only addressing artificial intelligence. I am not. However, I will base my definition on one given by Stuart Russell to define artificial intelligence[11]: an intelligent machine, in my account, is a machine that makes decisions about how best to achieve its goal. But *how* these decisions are made is not important. My definition of intelligent machine, therefore, subsumes the use of artificial intelligence as a means of making qualitative decisions or to implement the means of making such decisions.

The Parallel Between Collectives and Intelligent Machines A collective can be thought of as a non-human entity that comes into existence because of the activity and purpose of human persons. It also has a goal, which is the goal of the persons who brought it into existence, and the collective, through its structure and composition, decides on the best way to achieve that goal. A company is created by its founders for the purpose of creating profit, and, through its structure and operations, determines how it can best deploy its resources to maximise that profit. A university is founded by benefactors or the state to provide education, and decides through its various committees and management structures how best to provide that education.

In this sense, then, an intelligent machine is analogous to a collective. Both are non-human entities that are created by human persons for the purpose they specify, and which, by their human-specified structure, policies, and operation, work out the best way to achieve that human goal. If collectives can be considered morally responsible, by meeting the responsibility conditions, then can we say the same for intelligent machines?

Applying the three responsibility conditions to intelligent machines, based on the similarity with collectives, appears to provide a *prima facie* case for their moral responsibility. Firstly, intelligent machines seem to exercise some control over what they choose to do. Through their interaction with the world, they are able to select those actions that are available to them and rank them according to some criteria to decide which action is most likely to result in goal success. This then becomes the machine's choice, what it intends to do. Companies, for example, consider different choices of action to generate profit and evaluate and rank them by outcome and degree of risk, and the way intelligent machines consider what actions to take would appear, at least on the face of it, to be a similar process. Hence, if suitable collectives can meet the control condition, then so can suitable intelligent machines.

Secondly, machines interact with the world through input and output. Internal components, such as internal memory, file systems, and databases are capable of functionally acting as the same intentional states, representing belief and desire, as collectives. They also employ input devices such as environmental sensors, keyboards, and microphones to gather input from the world, and use output devices such as displays and audio speakers to generate output to change the state of the world. Finally, just as collectives have internal decision procedures or voting systems, intelligent machines have processes and rules. Just as collectives meet the intentionality condition, then, so do intelligent machines.

Thirdly, machines can also have a knowledge of harm and what harms may follow from their actions. Collectives can define and specify harms through such policies as equality and diversity policies, or anti-bribery policies. Similarly, machines can have coded within their structure and rules a concept of harm, or they might be shown examples of harm from which they can infer what harm "looks like", some representation they can use in the future to detect that an action might cause harm, which can then form part of their evaluation process.

Intelligent machines, then, would appear to satisfy the responsibility conditions in similar manner to collectives. As non-human entities without phenomenal consciousness, designed and created by humans to fulfil human goals, if we accept the fulfilment of the responsibility conditions by collectives, we must surely do the same for intelligent machines.

This view is a functional one. In essence, it is the same argument as that made by List.[12] Each responsibility condition is met by some feature of a machine that performs the same kind of function as a feature of a collective. Just as, for example, French discusses "suitably-organised" collectives, to mean that a collective with the appropriate structures and operation to meet the responsibility conditions, we might also talk about "suitably-constituted" machines to mean the same thing. That where, for example, in the case

[11] Russell's definition of artificial intelligence is as the field that studies how to build intelligent entities, "machines that can compute how to act effectively and safely in a wide variety of novel situations" (Russel [23], Chap. 1).

[12] List [17], Sect. 5.3.

of the control condition, a collective has an structure and appropriate internal decision procedures to allow the condition to be met, a machine has a structure and procedures, protocols and algorithms to allow it to operate in functionally the same way and equally satisfy the condition. And just as it is an empirical question to whether a collective is "suitably-organised", it is also the case as to whether a machine is "suitably-constituted". In both cases, the answer to this empirical question depends of the constitution of the collective or machine to say whether it is suitable or not.

The objection to machine responsibility from phenomenal consciousness The account I have given above seeks to establish a functional analogy between collectives and intelligent machines in order to demonstrate that, just as collectives satisfy the conditions of responsibility, so do intelligent machines. In the remainder of this paper, I will argue that this analogy does not hold in an important way concerning the control condition, and that, on the basis of that argument, intelligent machines cannot be morally responsible. However, there is another argument against the responsibility of machines that is ontological in nature, that they are not responsible because of something they are not. I will address this objection here and argue why it is wrong. It should be noted that each of the examples I give of this objection, with the exception of Collins, do not consider the functional analogy with collectives, and therefore apply to both machines and collectives.

The objection concerns the question of whether or not the sentience, consciousness or self-awareness[13] that is found in human persons is necessary for responsibility. There are numerous instances of arguments against machine responsibility, or the responsibility of "artificial agents", based on their lack, in some sense or another, of phenomenal consciousness.[14] Moor, for example, in defining what he calls "full ethical agents", those agents able to make and justify explicit moral judgements, associates them with the attributes of "consciousness, intentionality and free will"[15] (a view that seems to be shared by Wallach and Allen).[16] A similar argument is made by Parthemore and Whitby about what constitutes a morally responsible agent when they write that "a moral agent lacking in consciousness is a contradiction",[17] as well as by Floridi and Sanders, who argue that moral agency should be considered separately from responsibility, since machines (or "artificial agents") can be sources of actions that can qualify for moral judgements but cannot be blamed because they "lack a psychological component".[18] Véliz believes that "conscious experience, or sentience" is necessary to be a morally responsible agent, contending that "the main reason why algorithms can be neither autonomous nor accountable is that they lack sentience", that algorithms are "moral zombies" that are "incoherent as moral agents because they lack the necessary moral understanding to be morally responsible".[19] Similarly, Collins argues against machine responsibility on the grounds that machines lack "moral self-awareness", which she defines as "a phenomenal belief-like attitude … to the proposition 'I will do wrong'", although she accepts that suitably-organised collectives can nonetheless be responsible on the condition that there are human members of the collective that as a locus for such awareness.[20],[21]

These arguments are essentially ontological in that they proceed with the aim of arguing what machines are not. The basic argument is:

(i) only phenomenally conscious beings can be responsible or are fit to be blamed;
(ii) artificial agents, such as machines, are not phenomenally conscious;
(iii) therefore, artificial agents are not responsible or fit for blame.

If this argument is true, then any attempt to assign responsibility to machines, as well as to collectives, will fail.[22] However, I argue that it is not for two reasons.

Firstly, phenomenal consciousness is not sufficient for responsibility. There are examples of phenomenal consciousness and self-awareness in human persons who we do not normally hold responsible for their actions, such as children and dementia patients. Such arguments also typically fail to consider cases of manipulation, in which the subject is conscious, sentient, morally self-aware, yet not responsible for the harms they do. Secondly, it is difficult to argue that phenomenal consciousness is necessary for moral responsibility when our everyday experience is considered. As I have argued elsewhere in this paper, when we deal with

[13] These terms are often used without being fully defined, but I will take sentience, consciousness, etc. to mean some form of, or derived from, phenomenal consciousness.

[14] A similar argument in the context of collectives has been made by Baddorf [1]. For Baddorf, accountability is affective in a Strawsonian sense, it is an emotion of resentment and anger that seeks retribution or revenge in order to make the "offender realise what they have done to you", and since collectives lack the capacity needed for such realisation, they cannot be held accountable. However, ordinary practice shows that we do hold collectives to account, even though we know they do not in themselves possess consciousness.

[15] Moor [19].

[16] Wallach and Allen [27].

[17] Parthemore and Whitby [21].

[18] Floridi and Sanders [7], pg. 367.

[19] Véliz [26], pg. 488.

[20] Collins [3].

[21] A similar argument is made by Bernáth [2].

[22] See, for example, Tollon's argument on the application of reactive attitudes towards AI [25].

entities such as companies, we feel able to hold the company responsible, allot blame to it and exact retribution. If we think we have received a poor deal or bad customer service because of company practice or "company culture", we do not consider contacting a different person in the company, we take our business elsewhere. If a company dumps raw sewage in a river, spoiling its natural habitat for wildlife, we protest and boycott its products and services. Yet we know the company itself is not phenomenally conscious. Phenomenal consciousness, then, is not necessary for responsibility or for blame[23] in our everyday experience and practice.

This is not to say that in adult humans, consciousness, sentience, self-awareness in some form may play a role in establishing the capacity for moral responsibility. Instead, my claim is that they are not the only gatekeepers to moral responsibility, and we are justified in examining the possibility of extending that role to non-human entities that lack consciousness.

An objection similar to the phenomenal consciousness objection is provided by Sparrow.[24] Sparrow's argument is that a machine cannot be held responsible for its action because it cannot suffer punishment in the way we require when we hold someone responsible. However, Sparrow's argument takes a narrow view of punishment, its purposes, and the forms it might take relevant to the object of punishment. A company is non-human and cannot be said to experience punishment as a human person might. There is nothing that it is for a company to experience punishment. Yet we still punish companies.

These objections, therefore, do not succeed in denying responsibility to machines or any other form of artificial agent on the ontological grounds of lacking consciousness or being unable to suffer punishment.

4 The issue of control

In the previous section, I presented a *prima facie* argument, similar to that made by List,[25] that collectives and intelligent machines are functionally analogous in those ways that are relevant to satisfying the responsibility conditions, and that, therefore, just as suitably organised collectives possess moral responsibility, so do suitably constituted intelligent machines.

There is, however, a problem with the *prima facie* argument, which concerns negative excusing exceptions to moral responsibility. In the case of the first responsibility condition, the control condition, I will argue that collectives and intelligent machines are not functionally analogous due to the negative excusing condition of covert manipulation and that, while collectives are able to overcome this exception, machines cannot.

In the rest of this section, I will first discuss the relationship between control, as I use it in the formulation of the control condition, intentional agency, and moral responsibility. I then consider the negative excusing exception of covert manipulation on responsibility by looking at how it affects how one freely chooses. Following that discussion, I argue that design of non-human artificial entities, such as collectives and machines, is analogous to covert manipulation. However, I further argue that collectives are able to defeat the covert manipulation exception and retain responsibility, while machines cannot.

Control, Agency, and Responsibility Notwithstanding the other conditions for responsibility, we can only be responsible for what is within our control, in the sense of an exercise of our free will. Even though we can cause something to happen, if we do not do so of our own free volition, it cannot be said to be within our control. A hammer that hits the head of a nail causes the nail to sink deeper into the wood, but cannot be said to exercise control over that action. We can also intend something to happen, even intently desire that it happen, and then cause it to happen, yet still not be in control of what happens, if our intention and desire are planted in our minds by means of manipulation or if we are in the grip of some internal force, such as addiction.

Control is often mixed, in some arguments about responsibility, with an entity being an intentional agent. In other words, the idea of intending to do something is conflated with the concept of intentionality. For example, French, List, and Laukyte, as well as Collins, use a notion of intentional agency that includes control. French defines corporations as agents that are "Davidsonian" in nature, meaning that they are agents that both intend their actions and their actions occur because their intentional mental state is directed towards that action, following Davidson's argument about what actions an agent is responsible for.[26] A person who spills coffee on a rug because they intend to, because this is what they have chosen to do, is an agent of their action in this sense, whereas someone who spills coffee on the rug because someone jiggled their hand is not. Intentional agency is often used, then, in a way that implies control, in

[23] It is worth considering, although I do not have space to do so here, that the objections all assume a fundamental symmetry between the giver and receiver of a phenomenological moral response, such as anger. Both must have phenomenal consciousness of some form. However, this does not rule out more asymmetrical relationships are possible provided the moral response can be translated into a form understandable by the receiver. In the case of companies that we boycott, the boycott itself is the mechanism by which we hope our moral emotional response is translated into a form to which the non-phenomenally conscious company can react.

[24] Sparrow [24].

[25] List, ibid.

[26] Davidson [6].

the sense that the intended spilling of coffee on the rug happens because the agent wanted it to happen, they chose of their own free will that particular course of action, and then brought it about through an intentional mental state.

As long as intentional agency is used for adult human persons, this conflation of control and intentional agency, in the sense of intending *and* aiming at something, is not necessarily an issue. Where there is intentional agency but control is constrained or removed, it is usually seen as a matter of external constraint, such as coercion, or internal force, such as addiction. A bank employee is a full intentional agent, and, according to this definition, has the capacity for full control. His control means he is morally responsible for his actions, because he could freely, within perhaps the parameters of rationality and possibility, do otherwise. However, in a specific set of circumstances, his control over what he chooses to do may be constrained, for example, if criminals hold his family to ransom to force him to reveal the combination to the bank vault.

However, where an entity is not a human person, we cannot take the subsumption of control into intentional agency for granted. In other words, we cannot assume that intentional agency, for example, of the "Davidsonian" kind assumed by French, implies the same control enjoyed by adult humans. For example, it may be that certain artificial entities cannot make certain types of choices due to some innate aspect of their design. A university may be bound by covenants put in place by donors to act in a particular way, or a church may be constrained by articles of faith or sacred texts. Their degree of control, therefore, may not be the same as that of an adult human person, although their intentional agency will be equivalent. Hence, it will be important, when arguing for (or against) the assignment to other types of entity of the moral responsibility that we assume for adult human persons, to be aware of the differences, where they exist, in the degree of control that those entities possess.

Hence, I separate control from intentional agency in the responsibility conditions, not out of any disagreement about agency and control, but because I wish to emphasise the possibility of difference in this respect with human persons and the effect this may have on responsibility assignment to non-human entities. The essence of this difference, I argue, lies not in what may be chosen, which I will call the capacity of choice, but in the ability to choose.[27]

On the capacity of choice, the following example demonstrates that, even when the capacity to choose is restricted for some reason, the agent can still be morally responsible. Suppose a person Y has alternatives for action, but that their actions are purposefully limited, in some covert way, by another person X to only those alternatives that X allows. If X wants Y to make another person, Z, suffer, then Y is free to choose any action that achieves that end. Y can choose what they do, so long as it meets X's objective, and X will only step in if Y looks as if their intention is to do something that means Z will not suffer. Y can choose to blacken Z's name, steal Z's identity and empty their bank account, or some other action, but Y cannot do something that does not make Z suffer or that prevents Z from suffering. Similarly, suppose that X wants to make either Z or another person, such as W, suffer and does not mind who. Again, Y is free to choose, but their degree of control is limited to choosing which one suffers. In both these cases, Y can still be morally responsible since it is still possible that the action they choose is one that they genuinely want to choose.

Restricting capacity for free choice, by restricting what alternatives for action are possible or available, does not, therefore, eliminate the possibility for moral responsibility. Just as Y may be responsible in the first formulation of this problem, the same applies when we expand Y's range of choice of action, so long as what is done is what the agent, Y, freely wanted to do.

Influencing or controlling the ability to choose, however, has a different consequence for responsibility. Suppose X, instead of limiting Y's actions, manipulates Y's intentions, making it impossible for Y to conceive of any other action than one that harms Z. Y's control remains the same in terms of capacity, in terms of what actions may be chosen, but now differs in terms of ability. Due to X's covert interference, Y is incapable of conceiving of any choice or entertaining any intention other than to harm Z. In this case, Y's choice of action is not made freely; it is the result of manipulation by another, and so Y is not morally responsible. Y lacks control in a way that removes moral responsibility.

If we assume that X, rather than simply subtracting those intentions that Y may freely choose, instead positively sets, through manipulation of beliefs and desires, Y's choices, then this case is an instance of Kane's *covert nonconstraining control* (CNC),[28] in which agents act in accordance with their beliefs and desires but have been manipulated by others to have exactly those beliefs and desires and hence to do what their controllers want.

Manipulation, design, and control in artificial, non-human entities Kane's CNC was intended to highlight the flaws in compatibilist notions of responsibility based on alternative possibilities, and to demonstrate that freedom of choice needs to take account of the source of actions, the will. It is not enough simply to say that some alternative action could have been done to establish the freedom that

[27] Johnson [13], in arguing against the moral agency of computer systems, also makes the lack of freedom of choice central to her argument. Although in her case, the inability to choose, in the sense of the inability to intend to do something, results from computer systems lacking mental states—a claim which is not part of my argument.

[28] Kane [14], Chap. 5.

underpins moral responsibility. Instead, freedom to choose alternatives must be located in a person's will that is unaffected by forces such as coercion and manipulation that constrain what and how they choose.

When it comes to non-human, artificial entities, covert nonconstraining control is also a concern, because what such artificial entities may choose, and the values, beliefs, and desires that are the basis for how they choose, are also subject to constraints, in the form of their design.

By their nature, artificial entities, including collectives and machines, are entities that are subject to some form of design. They are the deliberate and planned creation of human persons, who create the artificial entities with the aim of achieving some specified goal of those humans. The design will include a specification of the goal, as well as specifications of the procedures and processes to be used to achieve that goal. To have some assurance that the artificial entity will achieve the goal, they are created according to a plan or blueprint that includes how the entity will achieve it. A company, for example, will be created for the purpose of selling software products. Its founders will create a set of documents detailing the purposes of the goal, agree an organisational structure, an initial set of policies, and so on. A software application for facial recognition will be developed according to some initial design document laying out aims and objectives, a test specification, initial training and testing datasets, and a description of a methodology for creating the underlying algorithm to classify faces.

The influence of its design upon an artificial entity in terms of control is both to constrain what choices are considered to be available and to influence the ability to choose. A design can restrict what may be done, and hence affect capacity, as in the case of a university bound by its charter and covenants to use its resources to educate. It cannot, for example, choose to turn its hand to gold mining. It also influences the values, beliefs, and desires of the entity. It determines what the entity will desire or seek, and what actions it will execute in order to achieve that desire. A facial recognition application created for the purpose of identifying candidates for promotion in a workplace will have a set of beliefs about the world, desires about what world it wants to bring about, and knowledge of what actions, or what kind of actions, are available, derived from its design. This is a necessary consequence of what it means to be a machine. A machine exists, as an artificial creation, to achieve some human goal; the design of the machine must assure its creators that the machine will achieve that goal; in order to provide that assurance, the design must restrict the capacity of choice, through constraining control, and direct the machine in how to choose, through covert nonconstraining control. If the design does not do this, then it cannot provide the assurance to the creators of the machine that it will achieve their goal.

In a sense, non-human, artificial entities are *predetermined persons*, in that they are intentional agents whose capacity of choice and ability to choose has been predetermined and preconfigured by others through their design, for the purpose of meeting the goals of those others. As such, artificial entities are analogous to human persons suffering covert nonconstraining control. A human person under covert manipulation is reflectively unaware that what makes them choose a particular action is not their own but implanted in them by others; a machine is unaware of its design, yet it is the design that is the source of what makes it choose a particular action. Covert manipulation of a human and the design of a machine are, therefore, analogous in the effect they have on the human's and the machine's ability to choose.

It follows, then, that just as covert manipulation of a human excuses them, the manipulative effect of the design is a permanent excusing exception to the moral responsibility of artificial entities. In other words, unless there is some mitigating factor that undoes or overcomes this exception, we must consider artificial entities, such as collectives and machines, as permanently suffering from covert manipulation and therefore not responsible for what they do.

It might be asked if the *prima facie* case, and the functional approach taken by, for example, List, can be amended to take into account such excusing conditions, but I do not think this is possible. The essence of the functional approach to machine responsibility is that the machine is deemed morally responsible purely on the basis of what is observable. List, for example, relies heavily on Dennett's notion of the *intentional stance*,[29] in which an entity is considered an intentional agent if it is observed to act as one. Such an approach would be unable to account for negative excusing exceptions to responsibility that are not observable, such as covert manipulation. So the *prime facie* case, and other functional approaches to machine responsibility, must be rejected.

A mitigating factor for collectives that preserves moral responsibility It would seem that the argument I have presented above means that no artificial entity of any kind can be held morally responsible. But this would seem to conflict with the general consensus that suitably organised collectives do possess responsibility, as well as with our everyday practice of holding collectives such as corporations to account.

In the case of collectives, however, there is a factor that mitigates against the influence of their design. A collective is composed of human members, and has, in its membership, a source of free will and agency that is not subject to its design. The human members of the executive board of a company can, for example, revise its articles of association

[29] Dennett [4].

or introduce new policies, that represent, for example, new goals on equality and diversity or on environmental sustainability, because beliefs and desires that they hold themselves.

What this illustrates is that the will of a collective is not wholly determined by its design, but, because of the supervention relationship on its members, who themselves possess free will that is not subject to the design, is also partially free. And it is free not only in that it has the capacity, through this supervention, to consider choices beyond those permitted by its design, but also in its ability to choose, in that the values, beliefs and desires of the collective supervene on those of its members. What I claim is that manipulation can only be considered a negative excusing exception to responsibility if the manipulation is total, but if it is only partial, as I argue it is in the case of collectives, then it cannot act as an exception. A collective that has recourse to free will, through its supervention on human persons, has a capacity of choice and an ability to choose that would not be possible under total manipulation. Collectives, therefore, because they are composed of human persons cannot be excused from moral responsibility due to their design.

Machines and moral responsibility Machines, however, do not supervene on human persons. The components that constitute a machine, whatever their nature, do not include human beings. The components of a robot, for example, may include various articulated bits of hardware to provide a moving skeleton, a central processing unit, memory storage, and input and output peripherals for vision and speech. A machine learning system will be composed of software that runs on a hardware platform. A machine may require a human administrator, may communicate its progress to a human supervisor, or provide its results to a human user, but these roles of humans with respect to machines—administrator, supervisor, and user—do not constitute a *supervention* relationship, but rather an *association* relationship. A machine runs under the guidance of, and for the purpose of, human persons, but is not itself composed of human persons. As a result, the will of a machine is entirely bound by its design, which is fixed and determined by its creators. Machines, therefore, are permanently excused from moral responsibility because of the manipulation imposed by their design.

It might be thought that, just as the executive board of a company can revise the articles of association and policies of the company, perhaps to allow different choices to be made or to represent different values or desires, users can have a machine revised, and that, if a machine can be revised just as a collective can be revised, then machines can also be responsible in the same way as collectives.

However, the possibility of revision of a machine's design does not overcome the negative excusing effect of its design for two reasons. Firstly, the possibility of revision in the case of collectives is not what restores moral responsibility, but rather an effect. The restoration of responsibility lies primarily in the supervention on the free will of the collective's human members, not with what effects the supervention makes possible, such as revising company policies.

Secondly, it might be argued that the users and developers of the machine also have free will in exactly the same way as the collective's members. In which case, the free will of the machine's users and developers plays the same role in creating free will in the machine as the free will of the collective's members in creating the free will of the collective. This argument, however, is mistaken in the nature of the relationship between the machines and its users and developers, and the collective and its members. As I argued earlier, the machine is associated with a user or a developer, but does not supervene on them. Hence the free will of the human users and developers of the machine does not form part of the will of the machine. The machine remains entirely in thrall to its design.

In essence, the will of the collective includes the free will of its human members, but the will of the machine, while it may be affected by changes to its design due to the free will of its users and developers, does not include that free will when it makes choices. Collectives, as a result, satisfy the control condition for moral responsibility, while machines fail. The functional analogy between collectives and intelligent machines, therefore, does not hold because of this difference. Collectives can be morally responsible because they are free to do otherwise. Machines, however, cannot be morally responsible, because their will cannot be free from the manipulation inherent in their design.[30]

5 Objections

I have argued that the functional analogy between collectives and intelligent machines does not hold for moral responsibility. The design of an artificial entity is analogous to covert manipulation, which excuses moral responsibility by failing the control condition. In the case of collectives, there is a mitigating factor that allows collectives to retain their responsibility, but this factor is absent in the case of machines.

Here I will consider three objections against this argument. The first is that machines can freely choose within the scope of their design; the second, that machines can change their own code and that this represents a degree of self-reflection that can overcome manipulation; and third, that a design may be sufficiently general or even weak enough to allow for the existence of free will. The first and second

[30] Having reached this conclusion, it is natural to ask "then who is?", when the actions of machines lead to harm, but that further question is beyond the scope of this paper.

objections I will dismiss, but the third I will consider significant enough to allow for possible future ascriptions of moral responsibility to machines.

Machines can freely choose even within their design Floridi and Sanders[31] have argued that artificial agents, agents that are created by humans and which possess sufficient knowledge, intelligence and capacity to operate independently of them, can be considered sources of moral actions (which they term 'moral agents') because, among other things, they act freely. For Floridi and Sanders, artificial agents are "free in the sense of being non-deterministic systems",[32] and that "the agents … satisfy the usual counterfactual: they could have acted differently, and they could have chosen differently because they are interactive, informed, autonomous and adaptive". By adaptivity, Floridi and Sanders mean the ability for an agent to change their own rules (which I will deal with as a separate objection below).

The essence of Floridi and Sanders' argument is that machines can be considered free even within the scope and influence of their design, because some part of their decision-making is not wholly determined by their design. In their case, this source is non-determinism, which we might take as meaning randomness. It does not, I think, really matter what the origin of randomness is within the machine, nor even its purpose, so long as it represents some factor in decision-making such that decisions are not wholly determined by the design.

On the face of it, this seems a significant objection, since it argues that, even with a design, machines can still satisfy the control condition. What a machine actually chooses is not fixed by its design but depends on some other source. However, I argue that this objection is wrong for two reasons.

Firstly, randomness does not equate to intent. If the machine's choice is between action A and action B, and A is chosen due to some random factor, this does not mean the machine intended to do A. A reply to this point might be that a person throwing a dice to decide between A and B is still responsible if the dice decides A and doing A results in harm. But in this case, the person chose to use the dice to make their decision. It was an application of their free will to rely on the outcome of the dice and then to do what the outcome of the dice roll indicated. This is not the case for the machine. The use of non-determinism, far from being the free choice of the machine, is an aspect of their design, as is the instruction to follow what the source of randomness indicates should be done.

Randomness, then, not only does not mean that the machine intends to do what randomness points to, but, even if it did, would not result in free will because the use of randomness and the compulsion to follow what it points to, is a result of the manipulation by the design. Since my argument is that design is analogous to manipulation of intent, rather than that the design prescribes what should be done in every circumstance, it follows that this objection fails. Randomness does not allow the machine to be free of its design.

Secondly, non-determinism only affects what choice is to be made between those choices possible, not *what* choices are possible. Recall, from the discussion of control and manipulation, the case of X and Y. X manipulates Y such that Y can only do what X wants them to do. If X wants Y to harm Z, then they manipulate Y covertly such that Y harms Z unaware they are being manipulated. It seems to Y that they have chosen freely to harm Z, but due to X's hidden interference, they could not have done otherwise. This remains true even if X's interference is such that, as X wishes, Y has the choice of harming Z or W. Y is manipulated, therefore, not just as the level of what they actually choose to do, but also in what choices they consider possible. Non-determinism does not overcome manipulation at this second level, in terms of what choices are possible.

Machines can change their design This objection is similar in nature to the counter-argument discussed in the previous section, in which claimed that the users and developers of a machine, in changing the design of a machine, end up creating free will in the machine in the same way that members of a collective do. That counter-argument I dismissed on the grounds that it was not the ability to change design that created free will, but, for collectives, it was the supervention relationship which is missing in the case of machines. Changing the design, whether it is by the machine's users and developers or by the machine itself, does not satisfy the control condition and hence result in free will.

In response, it might be argued further that the fact that it is the machine changing its design that is significant in this case, since it represents a degree of self-reflection. Even if the machine is not phenomenally conscious, it could be argued that it can still normatively assess its operation and outcomes and reach the conclusion that a particular change is required to its own design. A computer virus, for example, can modify itself when it assesses that its operation on the host machine has violated certain rules of anti-virus checkers.

However, this further objection also fails, since the normative evaluation of the machine of its design cannot be carried out beyond the scope and influence of the original design, which represents its manipulation. In other words, it cannot step outside its original manipulated state and freely evaluate itself. Any decision to modify its own design is done on the basis of that original design. In effect, what the machine would experience is an infinite regress of manipulation on the basis of the original design.

[31] Floridi and Sanders [7].

[32] Floridi and Sanders, ibid. Sect. 3.2.3.

Non-controlling Design One might suppose that it is possible to create a design for an artificial entity that does not constrain control or manipulate. However, as I argued earlier, designs are by their nature constraining and manipulating. A design represents a particular goal of the human persons who have commissioned and created the artificial entity, as well as the means to achieve that goal. For example, a university with a covenant placed on its buildings stating that they can only be used for education cannot sell them to be turned into a shopping mall. A facial recognition system designed to classify faces, into those that may be lying and those may not be,[33] cannot decide to introduce a new category, those that it likes, nor can it choose to disobey and stop classifying because it disagrees with the underlying methodology.

The specification of a goal of an artificial entity and the means to achieve it are in themselves the constraint and manipulation of the artificial entity. If the design did not constrain and manipulate, it would fail to fulfil its nature as a design. We could not be sure that the goal would be met; the artificial entity would then, in our eyes, be worthless as a vehicle for achieving our goal nor as an acceptable return on our investment in its creation. It is not possible, by definition, therefore, to create a design that is non-controlling.

A variation on this objection is that a design, rather than being non-controlling, could be sufficiently general in nature as to weaken its influence on the artificial entity sufficiently that the entity could be considered responsible for what it chooses to do.

Suppose a robot is created with the general goal of making humans happy by being happy itself. We might additionally specify that the robot should learn for itself how to do this, by providing a general definition of happiness so that it can recognise it in others, and to learn in some way from those people it finds to be happy. When the robot then chooses some action, given the lack of specificity in the design, can we still excuse it from responsibility for what it in fact does?

The notion of a design that is sufficiently weak in terms of the manipulation it represents is, I think, a significant objection. It raises the possibility that there is some threshold of the influence exerted by design, beyond which the effect of its manipulation may, while steering the entity in a general direction concerning its choices, still leave enough space to allow the entity to be considered free enough to be responsible. I will not argue here for how this threshold might be formulated or even the terms in which it may be understood.[34] I merely wish to suggest that such a threshold is conceivable, and hence that it is conceivable that machines, if their design were to place them above that threshold, could be considered morally responsible.

In the context of moral responsibility in general, Garnett[35] has made the case that a person can still be considered to act freely and be considered responsible even when what they do is caused by other persons. In a social context, we cannot be entirely free of the influence of others, such as educators, but we can, according to Garnett, have *enough* freedom, that there exists a threshold above which the influence of others is weak enough to allow sufficient freedom to live a decent life, where that includes being responsible for yourself. I suggest that Garnett's notion of a threshold of manipulation is similar, although without the social setting, to the threshold I have proposed here between the effects of strongly and weakly manipulative design on a machine.

6 Conclusion

My argument is that machines cannot be held morally responsible for what they do or for what they cause to come about, because they are in the same position as a human person suffering hidden manipulation or covert non-constraining control. Machines are in this position because, as a type of artificial entity, they are brought into existence and operate according to a design, which encapsulates a human goal and the means of achieving it. It is the design of an artificial entity, which is part of what it means to be artificial, that manipulates the entity and which excuses it from moral responsibility.

The argument made here applies to all artificial entities, since design is part of the ontology of artificial entities. However, in the case of collectives, I have furthered argued that there exists a mitigating factor that retains their responsibility. Collectives are not wholly determined by their design, since they are composed of human members, whose will exists outside the scope and influence of the design of the collective. By means of the supervention of the collective upon its human membership, therefore, collectives can still freely consider their choices and be responsible for what they do.

Since this argument relates to technology we use or may come to use in everyday life, one might well ask how we could verify the questions asked about moral responsibility,

[33] Such a system, ADDS [20], was a key component of the EU-funded project iBorderCtrl (https://cordis.europa.eu/project/id/700626/reporting).

[34] Nor do I mean to imply that any conceivable technology, such as Artificial General Intelligence, would represent such weakly general design.

[35] Garnett [10].

and the limitations of the answers given.[36] I confess I do not know how one might verify questions relating to ascription of moral responsibility concerning artificial entities, or whether it is possible. Our view of collectives, such as companies, as being morally responsible, for example, seems largely a practice that has come about and which exists by consensus, rather than through verification. Regarding limitations of this work, the dividing line between "strong" and "weak" design, the line between a design that is equivalent to manipulation that excuses responsibility, and a design that is weak enough to allow sufficient autonomy to allow for responsibility, is unclear. One important question concerns the terms in which this dividing line might be defined.

Ishiguro's Klara, his intelligent, empathetic robot, cannot by the argument made here be considered to be morally responsible for her choices. She may choose to harm herself, believing that it will save the girl in her care, or she may not. Either way, her consideration of possible choices and the basis on which she actually makes her choice are directly or indirectly an effect of her design. Her apparent freedom to choose is only apparent to herself. It is a product of the design formulated by her creators.

This conclusion, however, places Klara in our present, in which intelligent machines are created using designs that bind them tightly to specific goals. One possible objection to this conclusion is that it may be possible, in future, to design a machine in so general a way that it represents a level of manipulation weak enough as to permit freedom of will and hence re-admit moral responsibility for machines. The argument for or against this position is beyond the scope of this paper, and remains to be explored. If, however, this is at all possible, then Klara may yet still be responsible for her choice.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Baddorf, M.: Phenomenal consciousness, collective mentality, and collective moral responsibility. Philos. Stud. **174**(11), 2769–2786 (2017)
2. Bernáth, L.: Can autonomous agents without phenomenal consciousness be morally responsible? Philos. Technol. **34**(4), 1363–1382 (2021)
3. Collins, S.: I, Volkswagen. Philos. Q. **72**(2), 283–304 (2022)
4. Dennett, D.: Intentional systems theory. In: Beckermann, A., McLaughlin, B.P., Walter, S. (eds.) The Oxford Handbook of Philosophy of Mind, Oxford University Press (2009).
5. Fischer, J.M., Ravizza, M.: Perspectives on Moral Responsibility. Cornell University Press (1993)
6. Davidson, D.: Agency. In: Marras, A., Bronaugh, R.N., Binkley, R.W. (eds.) Agent, Action, and Reason, pp. 1–37. University of Toronto Press (1971)
7. Floridi, L., Sanders, J.W.: On the morality of artificial agents. Mind. Mach. **14**(3), 349–379 (2004)
8. Frankfurt, H.G.: Alternate possibilities and moral responsibility. J. Philos. **66**(23), 829–839 (1969)
9. French, P.A.: The corporation as a moral person. Am. Philos. Q. **16**(3), 207–215 (1979)
10. Garnett, M.: Freedom and indoctrination. Proc. Aristot. Soc. **115**(22), 93–108 (2015)
11. Hakli, R., Mäkelä, P.: Moral responsibility of robots and hybrid agents. Monist **102**(2), 259–275 (2019)
12. Hess, K.M.: Does the machine need a ghost? Corporate agents as nonconscious Kantian moral agents. J. Am. Philos. Assoc. **4**(1), 67–86 (2018)
13. Johnson, D.G.: Computer systems: moral entities but not moral agents. Ethics Inf. Technol. **8**(4), 195–204 (2006)
14. Kane, R.: The Significance of Free Will. Oxford University Press, New York (1999)
15. Laukyte, M.: Artificial agents among us: should we recognize them as agents proper? Ethics Inf. Technol. **19**(1), 1–17 (2017)
16. List, C.: What Is It Like to Be a Group Agent? Noûs **52**, 295–319 (2016)
17. List, C.: Group agency and artificial intelligence. Philos Technol **34**(4), 1213–1242 (2021)
18. List, C., Pettit, P.: Group Agency: The Possibility, Design, and Status of Corporate Agents. Oxford University Press (2011)
19. Moor, J.: The nature, importance, and difficulty of machine ethics. IEEE Intell. Syst. **21**(August), 18–21 (2006)
20. O'Shea, J., Crockett, K., Khan, W., Kindynis, P., Antoniades, A., Boultadakis, G.: Intelligent deception detection through machine based interviewing. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2018)
21. Parthemore, J., Whitby, B.: What makes any agent a moral agent? Reflections on machine consciousness and moral agency. Int. J. Mach. Conscious. **5**(02), 105–129 (2013)
22. Pettit, P.: Responsibility incorporated. Ethics **117**(2), 171–201 (2007)
23. Russell, S.: Artificial intelligence and the problem of control. In: Werther, H., Prem, E., Lee, E.A., Ghezzi, C. (eds.) Perspectives on digital humanism. Springer Nature Perspect. Digit. Hum., p 19 (2022)
24. Sparrow, R.: Killer robots. J. Appl. Philos. **24**(1), 62–77 (2007)

[36] I am grateful to one of the reviewers for raising these concerns.

25. Tollon, F.: Responsibility gaps and the reactive attitudes. AI Ethics **3**(1), 295–302 (2023)
26. Véliz, C.: Moral zombies: why algorithms are not moral agents. AI & Soc. **36**, 487–497 (2021)
27. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right From Wrong: Teaching Robots Right From Wrong. Oxford University Press (2008)

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