AI and Ethics (2022) 2:523–532
https://doi.org/10.1007/s43681-021-00109-5

ORIGINAL RESEARCH

Distributed responsibility in human–machine interactions

Anna Strasser[1]

Received: 9 August 2021 / Accepted: 4 October 2021 / Published online: 18 October 2021

Abstract
Artificial agents have become increasingly prevalent in human social life. In light of the diversity of new human–machine interactions, we face renewed questions about the distribution of moral responsibility. Besides positions denying the mere possibility of attributing moral responsibility to artificial systems, recent approaches discuss the circumstances under which artificial agents may qualify as moral agents. This paper revisits the discussion of how responsibility might be distributed between artificial agents and human interaction partners (including producers of artificial agents) and raises the question of whether attributions of responsibility should remain entirely on the human side. While acknowledging a crucial difference between living human beings and artificial systems culminating in an asymmetric feature of human–machine interactions, this paper investigates the extent to which artificial agents may reasonably be attributed a share of moral responsibility. To elaborate on criteria that can justify a distribution of responsibility in certain human–machine interactions, the role of types of criteria (interaction-related criteria and criteria that can be deferred from socially constructed responsibility relationships) is examined. Thereby, the focus will lay on the evaluation of potential criteria referring to the fact that artificial agents surpass in some aspects the capacities of humans. This is contrasted with socially constructed responsibility relationships that do not take these criteria into account. In summary, situations are examined in which it seems plausible that moral responsibility can be distributed between artificial and human agents.

Keywords Moral responsibility · Social human–machine interaction · New type of social agents · Distributed responsibility · Moral agency

1 Introduction

Soon we will share much of our social life with various kinds of artificial systems. In addition to manifold forms of tool use, I assume that this will probably also lead to a new type of social interactions—social human–machine interactions. Such interactions are characterized by the fact that they are not reducible to tool use and, moreover, are very similar to human–human interactions. Thereby, unequal interaction partners, namely living and non-living agents, face each other.

Claiming that artificial systems may enter the realm of social interactions may sound rather radical for a Western audience since our understanding of sociality is restricted to living beings. But outside of Western cultural ideas, for example, in Shintoism and Animism, objects are characterized as animate that are considered inanimate from a Western perspective. Furthermore, the claim that some human–machine interactions resemble social human–human interactions rather than reminding us of tool use is supported, for example, by the fact that the assumption that human–machine interactions are comparable to human–human interactions has already found its way into empirical research [1]. In several studies, experimental protocols with artificial agents are used to gain insights into human socio-cognitive mechanisms [2]. If there were no similarities, such experiments could not contribute to the study of human features. Last but not least, there is no doubt that humans can form emotional attachments to inanimate objects and that our tendency to anthropomorphize is already leading to cases where we treat artificial agents as if they were social agents. In this paper, I assume that if artificial agents contribute to social interactions by utilizing socio-cognitive abilities and thereby add to a reciprocal

✉ Anna Strasser
annakatharinastrasser@gmail.com

[1] Faculty of Philosophy, Philosophy of Science and Religious Studies, LMU, Munich, Germany

exchange of social information, we are justified to consider them social interaction partners.

Suppose that non-living machines can qualify as a new type of social interaction partner in human–machine interactions. In that case, questions arise whether we are justified to ascribe moral responsibility to such artificial systems while they are interacting with humans. One might assume that the question of moral responsibility is already settled by the decision to regard an entity as a social interaction partner. However, while being a social agent is a necessary condition to qualify as a moral agent, it is not a sufficient condition. Not all social agents are also moral agents. Similarly, the attribution of moral agency does not exclude that an interaction partner can be released from responsibility. So it is conceivable that this new type of social interaction partner could be characterized by not being able to bear responsibility. Only if one assumes that responsibility can turn out to be ascribable, the subsequent question regarding the distribution of responsibility arises. In other words, presupposing that this new type of a social agent can also qualify as a moral agent that can be responsible, this paper aims to investigate potential criteria clarifying how responsibility can be distributed between such unequal interaction partners. Considering this special case of social human–machine interaction, it is by no means clear from the outset how much responsibility remains on the human side. To address this question, I explore the extent to which our practice of distributing responsibility in human–human interactions can provide a strategy regarding distributing responsibility in social human–machine interactions. In doing so, the widespread intuition that human interaction partners should in principle be ascribed more responsibility will be critically examined. That is, it will be investigated whether there are cases in which it is justified to assign a greater share of responsibility to artificial interaction partners, even though human moral agents are both creators of the artificial systems and initiators of these interactions.

Regarding the question of whether and if so, how moral responsibility might be distributed between human interaction partners, artificial systems, and their creators, one finds a variety of controversial disputes in philosophy. Besides extreme positions which deny the mere possibility of attributing responsibility to artificial systems [3, 4], other approaches discuss under which circumstances artificial systems may qualify as moral agents [5–10]. Acknowledging that there are crucial differences between living human beings and non-living artificial systems, which culminate in an asymmetric feature of any human–machine interaction, this paper examines the pros and cons that can be used to make a case for distributed responsibility regarding social human–machine interactions.

To this end, I introduce a distinction between interaction-related criteria and criteria that can be derived from socially constructed responsibility relations. Both types of criteria are used as justifications to ascribe moral responsibility. The former refers to the manifestations of abilities that an agent needs to be a social interaction partner, assuming that the ability of both interaction partners to influence the outcome of an interaction is a crucial prerequisite for any social interaction. In addition to the mere agency, socio-cognitive abilities play an important role, as they enable an agent to anticipate, plan and control his or her behavior in social interactions. For example, to be a successful interaction partner, two kinds of anticipation abilities are required, namely, being able to anticipate actions of the interaction partners (mindreading) and consequences of the interaction. Regarding the distribution of responsibility, it is assumed that greater expertise (in the sense of further developed socio-cognitive abilities) in social interactions can serve as a reason to ascribe a higher share of responsibility. Beyond this, however, there are also criteria that we can derive from socially constructed responsibility relationships. Examples are responsibility relationships between adults (parents) and children, superiors and employees, or toward protégés. Some of these responsibility relationships are also reflected in jurisprudence. Such criteria seem to prescribe a certain weighting regarding the distribution of responsibility. In general, criteria derived from socially constructed responsibility relationships reflect the evaluations of interaction-related criteria. However, there are also cases in which socially constructed responsibility relationships argue for a different distribution of responsibility than case-by-case assessments based on interaction-related criteria. For example, a child might exhibit responsibility-related skills in a particular interaction that the adult does not possess at that very moment; in such cases, it seems reasonable that criteria derived from socially constructed responsibility relationships deserve a greater weight.

In this paper, it is shown that there are many cases in which evaluations of social human–machine interactions based on interactional criteria point in a different direction than evaluations based on criteria derived from our established, socially constructed responsibility relationships. I take this as a motivation to question whether in such cases established, socially constructed responsibility relations legitimately reduce the weight of interaction-related criteria. For it could be that precisely such cases provide arguments for bringing about a reassessment of socially constructed responsibility relations in relation to certain artificial agents.

2 Responsibility on the human side

Attributing moral responsibility to human beings is closely related to the question of necessary and sufficient conditions by which social agents qualify as moral agents. Our notion

of full-fledged moral agency requires demanding conditions such as consciousness, free will, reflective abilities, understanding, and the ability to process information along with a belief-desire-intention architecture. Moral responsibility must, of course, be distinguished from a purely causal responsibility. Even though we consider people to be morally responsible because they have conscious mental and emotional states, intentionality, intelligence, the ability to think, plan, judge, and act differently (free will), we may relativize the extent of responsibility if some of these abilities are impaired or not yet fully developed. This is where concepts such as *diminished culpability* come into play. In discussing the complex factors that affect the assignment of responsibility and its distribution, it becomes clear that moral agency is a necessary but not a sufficient condition. One can point here to debates about the role of conditions such as the presence of intent or lack of self-control, diminished impulse control, a general ability to have empathy [11].

Turning to distributed responsibility, human–human interactions present cases where individual agents are not held solely responsible because it is assumed that responsibility is shared with the interaction partner. Interestingly, the amount of responsibility between two interaction partners is not always equally distributed; for example, we argue for an unequal distribution of responsibility in adult–child interactions.

Although approaches regarding moral agency imply prerequisites for the capacity to be responsible, the notion of responsibility itself remains a highly complex, multi-faceted, and also disputed concept [12, 13]. Especially when it comes to questions of how much responsibility should be ascribed to an agent if more than one agent are involved in an action (e.g., many hands problem [14]). It is often unclear which criteria play a decisive role or what a clear-cut strategy regarding weighing potential opposing criteria could be. In addition, in the case of a reduced attribution of responsibility, it is not clear who is the addressee for 'left-over' responsibility. For example, one can refer to recent innovations in technology—autonomous machines, learning algorithms, and social robots—that challenge the attribution of responsibility based on instrumental theory and lead to debates about *responsibility gaps* [15]. In this paper, I focus on distributed responsibility in social interactions, thereby I support conceptualizing responsibility as distributed across a network of humans and machines [16]. It is also not clear from the outset whether responsibility can be divided up like a pizza, namely that one can infer the size of the share from the number of responsibility bearers [17].

Analyzing our practice of responsibility allocation in human–human interactions by distinguishing between interaction-related criteria and criteria arising from socially constructed responsibility relations, the interaction-related criteria provide information about the extent to which the interaction partner has an influence on the outcome of the interaction. For example, if an interaction partner cannot overlook the consequences of an interaction or is mistaken in his or her assumptions about the interaction partner's future behavior, he or she has diminished influence. The extent to which relevant abilities are developed here allows a conclusion to be drawn about the degree of influence. If the interaction partners have comparably developed abilities, it is obvious to argue for an equal sharing of responsibility. Conversely, a different expression of these abilities speaks for an unequal distribution of responsibility. The criteria that can be derived from socially constructed relationships of responsibility establish a certain distribution of responsibility from the outset. Regardless of the case-by-case assessment of individual interactions, parents, for example, are attributed a greater share of responsibility in adult–child interactions.

3 Responsibility on the side of artificial agents

Specifying the conditions for moral agency plays a crucial role in the debates about responsibility, even though qualifying as a moral agent is not always a sufficient reason to justify the ascription of moral responsibility. However, the notion of full-fledged moral agency excludes artificial systems from the onset. If one assumes that meeting traditional criteria for full-fledged moral agency such as consciousness is a necessary precondition for responsibility then artificial systems cannot be (held) responsible [18, 19]. To assume morally responsible artificial agents, one needs an alternative to this demanding notion. This is the point where proposals suggesting a gradualist conception regarding the notion of moral agency come into play. A distinguishing feature of such gradualist conceptions is that they question the necessity of some specific conditions of full-fledged moral agency by presupposing that there are different ways how moral agency can be realized. Assuming multiple realizations, they argue, for example, that moral agency can even be ascribed to unconscious, non-living agents. For instance, taking different expressions of conditions such as autonomy and sensitivity to moral values into account, one can, according to Wallach & Allen [8, 9], distinguish operational moral agents from weak functional ones, and these in turn from full-fledged functional moral agents. Alternatively, one could follow Moor [6], who elaborated on several types of moral agents by distinguishing agents with ethical influence from implicit ethical agents and explicit ethical agents. The latter corresponds to full-fledged moral agency with consciousness, intentionality, and free will. According to Floridi & Sanders [5], artificial systems can be attributed

moral liability if they fulfill conditions regarding interactivity, autonomy, and adaptability.

The discussion of the principle question of whether artificial systems can at all be considered moral agents will not be addressed in detail in this paper. By investigating whether social human–machine interactions may lead to a form of distributed responsibility, I presuppose that some artificial agents can be qualified as a new kind of social interaction partner and that a gradualist notion of moral agency is applicable to some of those agents. Therefore, accepting a gradual notion of moral agency as applicable to certain artificial agents is a necessary presupposition for the following investigations. At this point, it is important to make a clear distinction between the different presuppositions assumed. If one assumes that there can be human–machine interactions that are not reducible to mere tool use and that such interactions can be meaningfully considered a new kind of social interaction, one must argue for expanding the notion of a social agent.[1] By this, one takes the position that certain artificial systems can qualify as social agents if they possess both a kind of agency (minimal agency) and a form of social competence (minimal socio-cognitive abilities) such that they can both contribute to an exchange of social information and have an influence on the outcome of a social interaction [20]. However, this claim that it is conceivable that certain artificial agents qualify as a new type of social agent does not yet imply that those agents automatically qualify as moral agents. For example, animals can be considered social agents without necessarily implying that they have moral agency. And even the attribution of moral agency does not conclusively clarify questions regarding the attribution of responsibility. For example, there may be social agents to whom we attribute moral agency, but to whom we nevertheless deny culpability under certain circumstances. The aim of this paper is to examine under which circumstances we are justified in ascribing how much responsibility to this new type of social agent, provided that they can also be qualified as a moral agent with minimal moral agency.

To approach this question, I investigate the interaction-related criteria, which build on the manifestation of certain abilities enabling social interactions, and the criteria derived from socially constructed responsibility relationships. The latter concerns the extent to which we as manufacturers of artificial systems and as initiators of social interactions are in principle entitled to greater (or full) responsibility in social human–machine interactions, similar to what is common in product liability regarding tools.

[1] Given that attitudes toward the status of social agents have proven changeable throughout human history, this is not an unfulfillable requirement. For instance, the changes in relation to the status of women, children, other ethnic groups, and nonhuman animals show how we have managed to expand the class of social agents.

With respect to interaction-related criteria, I will explore how the asymmetric distribution in human–machine interactions can meaningfully be compared. One could argue that humans fundamentally surpass artificial interaction partners with regard to the interaction-related criteria since artificial systems only fulfill minimally or even simply differently realized conditions through the introduction of a gradual conceptualization. Artificial agents are not alive, they have neither consciousness nor emotions, nor are they capable of suffering. These are all points that feed a justified skepticism towards the attribution of responsibility [21]. Following this line of thinking, a gradual conceptualization of moral agency complicates the evaluation of interaction-related criteria in human–machine interactions since some conditions that are necessities for humans are not required from artificial systems.

From my point of view, it is unlikely that artificial agents will be endowed with consciousness and emotions, and I think that on the basis of a gradualist conception of moral agency this need not be required. If one understands multiple realizations of moral agency as equivalent, one can point out that there are also abilities in which artificial agents surpass humans. For example, they can process and store a greater amount of data in a shorter time. This can have crucial consequences when examining the extent of their influence on the outcome of an interaction. Human reaction times can simply be too slow to intervene effectively.

Moreover, limits have already been reached beyond which humans no longer understand in detail how artificial systems work. Although humans are able to construct artificial systems, they are confronted with the so-called black box problem, i.e., they cannot understand the internal processes. For example, it is not obvious according to which criteria a trained neural network 'decides' to assign a category. Therefore, humans often cannot predict how artificial agents will behave. The question here is to what extent, for example, our limited ability to predict the behavior of artificial agents absolves us from taking on a greater share of the responsibility. In this paper, I assume that it is possible to compare the abilities of artificial and human agents.

To elaborate how to evaluate an asymmetric distribution of abilities when one agent has full-fledged moral agency while the other has only minimal moral agency, I examine our practice of distributing responsibilities in child–adult interactions and ask then whether there is anything to be learned from this practice that can be transferred to clarifying the distribution of responsibilities in social human–machine interactions.

3.1 Child–adult interactions

Without doubt, abilities are asymmetrically distributed in child–adult interactions. Therefore, such interactions might

be a viable example for discussing the distribution of responsibility in social human–machine interactions. With respect to child–adult interactions, it seems natural that the asymmetry in abilities leads to an unequal distribution of responsibility. Consequently, we claim that adults bear more responsibility than children in child–adult interactions because they have further developed abilities. One could state that adults qualify as experts regarding many aspects of social interactions, whereas children are considered novices. Being an expert (in the sense of further developed social-cognitive abilities) in social interactions goes along with several features: for example, adults can contemplate and choose from a wider array of actions, are more often able to intervene, and can better anticipate both the future behavior of interaction partners and the consequences of actions. In other words, adults have more opportunities to intervene and have more control over the outcome because they can better assess circumstances and consequently know more about the consequences of actions. All these features count as reasons why adults are held more responsible. On the basis of interaction-related criteria that specify the development of the abilities for moral responsibility in interactions, an unequal distribution of responsibility between children and adults can be justified. The more pronounced the relevant abilities of an interaction partner are, the more responsibility is attributed to this interaction partner. However, there is another type of criteria that is crucial for our assessment regarding the distribution of responsibility. These are criteria that can be derived from socially constructed responsibility relationships. Concerning adult–child interactions, one can point out that adults, especially parents, are in principle attributed a greater share of responsibility. Parents are responsible for their children. Based on established responsibility relationships, such criteria may carry more weight than the case-by-case evaluation of interactional criteria when it comes to the distribution of responsibility in interactions between children and adults.

4 Distributed responsibility in human–machine interactions

Turning to social human–machine interactions, our first intuition might be to attribute less responsibility to artificial agents as we do to children in child–adult interactions because the agency of a child might be compared to that of the artificial agent [22]. However, by evaluating interaction-related criteria in more detail, it becomes questionable whether an analogous justification regarding the distribution of responsibility can be applied to social human–machine interactions. I argue that on the basis of interaction-related criteria, artificial agents cannot be assigned an equivalent role as children have in child–adult interactions. This is because we cannot assume that human agents will generally prove to surpass artificial agents regarding the relevant abilities for moral responsibility in social interactions. Considering cases where artificial systems outperform humans because they are able to process and store a greater amount of data in a shorter time is leading to more control regarding the outcome of the interaction, the role of artificial systems in this regard would be more akin to the role of adults in child–adult interactions. Moreover, drawing on less developed abilities of children in child–adult interactions, I argued that we attribute less responsibility to children because children are often not very good at predicting the interaction partner's behavior and cannot foresee the consequences of an interaction. Taking diminished anticipation abilities as a criterion that justifies why we consider children to be less responsible, one is confronted with cases of social human–machine interactions that speak for attributing less responsibility to human participants because humans often cannot anticipate artificial systems' behavior or the outcome of the interaction. Does this mean that artificial agents deserve more responsibility than human interaction partners?

One could object here that assuming a gradual conception regarding social and moral agency requires less demanding conditions of artificial agents and, consequently, this could be taken as a justification to attribute less responsibility to artificial agents. I argue, however, that the less demanding conditions of a gradual conception of moral agency cannot be the only decisive factor because we still find criteria regarding which artificial agents outperform human interaction partners. This means that one must finally weigh up whether the less demanding conditions regarding moral agency should weigh more heavily than other conditions in which artificial agents are superior to humans. At this point, I leave open the question of whether we might, in principle, attribute less responsibility to artificial agents because they do not meet the demanding conditions of full-fledged moral agency and examine criteria that speak in favor of attributing more responsibility to artificial agents, such as having more control over the outcome of an interaction.

A serious fundamental objection to attributing responsibility to artificial agents concerns criteria that we can derive from socially constructed responsibility relationships. Here it seems to be common sense that only humans as constructors or as users of artificial systems can be attributed moral responsibility for the outcome of interactions. From this, one could conclude that we as constructors of artificial systems and as initiators of human–machine interactions would be the only addressees for responsibility attributions. However, this is in conflict with the idea put forward here, namely that under certain conditions artificial systems can turn out to be a new type of social interaction partner, and that in this role, moreover, they can also possess minimal moral agency.

Consequently, I question whether it is appropriate to apply our socially constructed responsibility relationships to the special case of social human–machine interactions.

This is the point where I want to emphasize that our established socially constructed responsibility relationships are based on the idea that all human–machine interactions could be treated as cases of tool use. In tool use, humans are the main addressees of responsibility, even though humans are not always held fully responsible. To avoid dangerous outcomes, our society requires proof of knowledge to allow people to use certain tools. For example, one needs a driver's license to use an ordinary car. Both external and internal reasons can justify a reduced attribution of responsibility. Suppose a person can neither foresee nor influence adverse circumstances that significantly influence the outcome of an action. In that case, this has a diminishing effect on the assessment of responsibility. For example, if unforeseeable environmental factors limit intervention possibilities, this leads to less responsibility regarding the human agent. In addition, there are situations in which certain human agents are in general ascribed diminished culpability. According to German case law, persons are incapable of culpability if, at the time of committing an action, they are unable to recognize the wrongfulness of the action or are unable to act on the basis of this recognition due to a pathological mental disorder, a profound disturbance of consciousness or a reduction in intelligence or another serious mental disorder (§ 20 StGB). Another case concerns technical devices failing despite proper handling. Here, the user is not held responsible; instead, the producer of the devices can be attributed responsibility. There is an extensive body of case law dealing with liability issues. Interestingly, there are various cases of diminished responsibility in which the question of to whom or what the remaining residual responsibility can be attributed is not finally clarified. This led to debates about responsibility gaps [15]. I argue that with respect to social human–machine interactions that cannot be reduced to mere tool use, the question arises to what extent diminished responsibility at the human side may justify attributing a higher share of responsibility to artificial interaction partners and thereby fill in potential responsibility gaps.

Since it may be questionable whether relying only on case-by-case assessments based on interaction-related criteria provides a sufficient justification to ascribe moral responsibility to artificial agents in social human–machine interactions, I suggest investigating whether currently applied socially constructed responsibility relationships, which speak for attributing responsibility exclusively to the human beings, are appropriate regarding social human–machine interactions. Describing child–adult interactions, I mentioned socially established pre-existing responsibility relationships adults, especially parents, have towards children. Regarding these responsibility relationships, one can argue that the adult's responsibility is already higher prior to any interaction—but at the same time, they reflect interaction-related criteria. In contrast, currently applied socially constructed responsibility relationships regarding human–machine interactions do not reflect interaction-related criteria of social human–machine interactions.

Questioning socially constructed responsibility relations is by no means to be meant to ignore reasons suggesting that humans are, nevertheless, responsible for being vigilant and suspicious of artificial agents before interacting with them [23, 24]. Due to an analysis of interaction-related criteria, one might be inclined to shift some responsibility away from us, but this should not hide the fact that there still can be a prior obligation to acquire expertise in dealing with artificial agents. This may even lead to considerations speaking in favor of establishing normative restrictions concerning the production of certain systems, such as killing drones [7, 25].

Nevertheless, it should be critically reflected that the current socially constructed responsibility relations treat all human–machine interactions as cases of tool use. Therefore, I argue that referring to these responsibility relations should not be taken as a free pass to absolve artificial agents from responsibility a priori. To illustrate what I am talking about when I raise the question of whether we should rethink our socially constructed responsibility relationships, let me give a fictional example. Imagine a normal driving lesson; here, two people interact together with a tool (the car). This is a joint action. In dicey situations, the driving instructor intervenes because the learner driver may be confronted with a situation that exceeds his/her abilities, while the driving instructor can manage the situation based on her expertise. An analysis of the interaction-related criteria will argue for attributing more responsibility to the driving instructor. The same result can be reached by referring to criteria that can be derived from socially constructed responsibility relationships (teacher/student). What concerns me is the extent to which our assessment would change if we were to replace the driving instructor with a driving assistance system (I admit this is quite fictional at this stage)—but how would we then judge about the distribution of responsibility? I suspect that we would arrive at a similar assessment on the basis of interaction-related criteria, but criteria that we can derive from socially constructed responsibility relationships could only consider human agents to be responsible here, or speak of a responsibility gap.

Instead, a detailed assessment of interaction-related criteria could inform a society on how to rethink established responsibility relationships in relation to this new type of social human–machine interactions. To carve out the relevance of interaction-related factors, I will discuss a theoretical case of a social human–machine interaction.

4.1 Autonomous vehicles and control

Social interactions are cases in which social interaction partners contribute together to an outcome of an action. Joint actions present a paradigmatical example of a social interaction. If social agents act jointly, it seems uncontroversial to claim that they share responsibility. An evaluation of the interaction-related criteria can deliver a justification of how responsibility should be distributed. For example, adults exhibit further developed abilities regarding interaction-related criteria than children. They can choose from a wider variety of actions, intervene more frequently and anticipate how interaction partners will behave. All this leads to a higher degree of control. Consequently, being in control is likely to be an important determinant of responsibility.

Analyzing interactions with autonomous vehicles, controllability can have varying manifestations. Not all human–machine interactions qualify as social human–machine interactions. In general, we can distinguish several levels of control. If a vehicle is under our total control (a so-called in-the-loop system), then this is a typical case of tool use that cannot be considered a social interaction because tools have no agency. That is, artificial systems here do not influence the outcome of the action together with the human interaction partner. Nevertheless, tools can be causally responsible, but we cannot attribute moral responsibility to them because they are no social interaction partners.

However, autonomous vehicles are not under our complete control; they display grades of autonomy and are eventually even able to adapt and learn. This can be reflected by further distinguishing between on-the-loop systems and out-of-the-loop systems [25]. Out-of-the-loop systems describe machines in which humans have no intervention options besides initiating the action. These are also cases in which we cannot speak of a social interaction because the human agent does not contribute much to the action. Of course, in cases of serious misconduct, humans would still be held responsible for initiating this interaction.

Whereas, regarding on-the-loop systems, which have some autonomy, both the artificial and human agents can intervene and have an influence on the outcome of the interaction. These cases can be described as a new type of social interaction. And then, it is feasible to ask to what extent each interaction partner is responsible for the outcome of an interaction. If autonomous vehicles are on-the-loop systems, I suggest discussing how responsibility should be distributed between the two interaction partners and whether artificial agents could, in some sense, contribute to filling in the responsibility gap. Otherwise, our society is "facing a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription" ([15] p. 175). This is because with respect to autonomous, learning artificial systems based on neural networks, genetic algorithms, and agent architectures, neither the producers nor the users could be held fully morally responsible or liable since they are unable to predict the future behavior of the artificial systems.

However, it is difficult to arrive at a clear-cut evaluation because interaction-related criteria and criteria that can be deferred from socially constructed responsibility relations point in different directions regarding how to distribute responsibility. Assuming that both human and artificial agents contribute to the outcome of the interaction, it seems reasonable to distribute responsibility along interaction-related criteria. This is, for example, to argue that limited anticipation abilities on the human side regarding the prediction of the artificial interaction partner's actions can speak for an unequal distribution. Furthermore, reduced cognitive abilities regarding processing and storing data resulting in diminished possibilities to contribute and intervene can also be used to argue that human agents deserve a smaller share of responsibility. In contrast, criteria deferred from current socially constructed responsibility relationships regarding human–machine interactions only speak for ascribing responsibility to human agents. Before turning to the question of whether there are reasons to reconsider socially constructed responsibility relationships, I will discuss another example.

4.2 Interacting with expert systems

Looking at interactions with expert systems, we can observe that people making use of expert systems especially tend in the case of failures to attribute some kind of responsibility to the systems. For instance, we often excuse failures in wayfinding by saying that it was not our fault but the fault of Google Maps. However, such examples are no cases of a social human–machine interaction. Rather, this is a special case of tool use. Therefore, investigating interactions with expert systems, which turn out to be mere tool use, can do little to clarify the question of to what extent artificial interaction partners can be held jointly responsible in social human–machine interactions. In the case of tool use, the human being makes the final decision whether to accept the offered advice, and the artificial system has no obvious option to intervene. All we can attribute to artificial systems is causal responsibility.

However, this example can serve to highlight important differences between tool use and social human–machine interactions. Regarding tool use, a reduced attribution of moral responsibility with respect to the human agent cannot justify assigning the rest of the moral responsibility to the tool. Tools can only be causally responsible. Addressees of moral responsibility regarding tool use are exclusively human beings, namely the producers of the tools and the users. Interestingly, we seem to apply different kinds of evaluations depending on whether such an interaction was

successful or not. In cases of success, we usually do not emphasize the causal part our tools played in the success. Instead, we tend to take the whole credit and responsibility for the outcome of our actions. However, analyzing failures, we can distinguish between two cases. Some failures result from the fact that the human ignores the advice of the artificial system. In other cases, following the advice can turn out to be the cause of the failure. In the latter case, we tend to blame the producer of the artificial system. In the former case, the entire responsibility lies on the shoulders of the human user because the artificial system had not even the chance to have an influence on the outcome.

Returning to the initial question of whether social artificial and human interaction partners can be held jointly responsible, one can highlight that cases of distributed moral responsibility in social interactions presuppose that both interaction partners have to a certain extent the possibility to intervene. A potential subclass of being able to intervene might be constituted if the artificial interaction partner would be able to nudge the human one [26]. Via changing the way decisions present themself to the human agent, the artificial agent would have an influence on the human decisions. Future research regarding expert systems might lead to cases that cannot be reduced to tool use.

Suppose cases in which the human attributes some authority to expert systems. In such cases, the human decides to take the expert system's advice seriously before interacting. Thus, the expert systems' suggested actions play a role with respect to the outcome of the interaction. Here we are in an intermediate area between social human–machine interactions and tool use. However, even if expert systems are granted an influence here, this influence is in the last instance dependent on the previously made decision of the human agent. Therefore, I conclude that interactions with expert systems are not a good example to elaborate on social human–machine interactions.

5 Do socially constructed responsibility relationships outweigh interaction-related criteria?

With respect to social human–machine interactions, I suggested reconsidering established socially constructed responsibility relationships on the basis of case-by-case evaluations utilizing interaction-related criteria. Suppose one applies interaction-related criteria in evaluating the moral responsibility of designers and that of the human interaction partner deciding to engage in a social interaction with an artificial system. In that case, one could argue for weakening the share of responsibility on the human side due to a limited anticipatory capacity and due to the fact that the artificial agents are able to process and store a greater amount of data in a shorter time. Above considerations showed that attempts to justify attributing moral responsibility to artificial systems in social human–machine interactions are leveled by criteria deferred from current socially constructed responsibility relations, which exclude the possibility of moral responsibility regarding artificial agents. In favor of defending the claim that certain artificial agents can be held morally responsible, one has to question whether existing socially constructed responsibility relations are subject to a design flaw by unjustifiably reducing social human–machine interactions to tool use. Suppose socially constructed responsibility relations are revised on the basis of the analysis concerning interaction-related criteria. In that case, we will face particular social human–machine interactions in which artificial agents are attributed more responsibility than their human interaction partners.

5.1 Objections

But even if one would introduce new socially constructed responsibility relations considering that social human–machine interactions cannot be reduced to tool use, the project of ascribing moral responsibility to artificial agents remains counterintuitive for other reasons. One potential objection raises the question of what it should mean when such agents are considered morally responsible, and at the same time, this has no further consequences. In the human sphere, the attribution of responsibility is usually accompanied by the justification of sanctions and punishments. Regarding artificial agents, we have no idea what could be understood as a sanction or punishment at all because artificial agents cannot suffer from any punishments [27]. That is, even if we ascribe some form of moral agency to artificial agents and grant them the possibility to influence the outcome of joint actions as social interaction partners, this is accompanied by an uneasy feeling resulting from a categorical difference between humans and this new kind of social interaction partners. Taking into account the lack of possible sanctioning measures, the general project of attributing responsibility to artificial agents seems to falter. Neither a gradual notion of moral agency nor the attribution of minimal socio-cognitive capacities, such as a minimal capacity to act jointly, clarify what consequences should follow from the attribution of responsibility.

At this point, one might ask whether one should insist on a connection between the attribution of moral responsibility and subsequent consequences at all. However, this fundamentally calls into question the meaningfulness of attributing moral responsibility. Similar to the debate about responsibility gaps, there is no culprit that we could punish. A temporary solution to this problem might be to demand that social artificial agents should be provided with some sort of liability insurance that could at least monetarily cover

the damage if they were responsible for causing negative consequences.

Future research would need to address another fundamental problem, which consists in the fact that we seem to ascribe the status of being a new type of a social agent to artificial agents only during a social interaction. Outside the interaction, the momentarily social agent seems to turn back into a tool. At the latest, when we switch off artificial systems, we deprive them of the status of a social agent. This contradicts our intuition that social agents also have a social status outside of interactions and should also have the ability to understand the wider context of a single interaction.

6 Conclusion

Returning to the initial question of whether we are justified to assume that human beings may in some social human–machine interactions deserve less responsibility than their artificial interaction partners, the above considerations showed that an analysis of interaction-related criteria can support a positive answer. Referring to interaction-related criteria regarding anticipation skills and the ability to process and store a greater amount of data in a shorter time, one can argue that human interaction partners with reduced anticipation skills and a less developed ability to process and store data deserve a smaller share of responsibility in social human–machine interactions. Applied to future social human–machine interactions, this would mean that the more artificial systems surpass humans, the more morally responsible they can become in social interactions.

However, it has also been shown that current socially constructed responsibility relations do not support this because they treat artificial systems as tools and not as social interaction partners. Only if our socially constructed responsibility relations would distinguish between artificial agents that qualify as social interaction partners and those who can be reduced to tools, one would be in a position to defend distributed responsibility in social human–machine interactions. This shows that the project of describing moral agency as a gradual phenomenon would have to entail more far-reaching consequences than it currently has. But even if the idea of artificial agents as interaction partners in social interactions were to gain acceptance, one faces further fundamental problems. This concerns the close connection between responsibility and possible sanctions and the question of what status artificial systems have outside of social interactions. Future research should therefore investigate the extent to which artificial systems could face up to their responsibilities if, for example, they were equipped with liability insurances. Furthermore, detailed research is needed on how to assess artificial systems outside of social interactions. The special thing about people's responsibility is, after all, that they do not only prove to be responsible during a social interaction but that the responsibility goes beyond that. The continuous existence as social agents plays an essential role in our practice of dealing with attributions of moral responsibility. How to judge artificial agents that are currently not in any social interaction remains an open question at this point. Considering those objections shows that there is still a lot of work to be done before we are able to take into account a new type of interaction partner with all its ethical consequences.

Funding Open Access funding enabled and organized by Projekt DEAL. The author did not receive support from any organization for the submitted work.

Availability of data and material Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest There are no conflicts of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Hortensius, R., Cross, E.S.: From automata to animate beings: the scope and limits of attributing socialness to artificial agents. Ann. N. Y. Acad. Sci. **1426**, 93–110 (2018)
2. Wykowska, A., Chaminade, T., Cheng, G.: Embodied artificial agents for understanding human social cognition. Phil. Trans. R. Soc. London ser. B. Biol. Sci. **371**, 20150375 (2016)
3. Nida-Rümelin, J., Weidenfeld, N.: Digitaler Humanismus: Eine Ethik für das Zeitalter der Künstlichen Intelligenz. Piper Verlag, Munich (2018)
4. Bryson, J.: Robots should be slaves. In: Wilks, Y. (ed.) Close engagements with artificial companions: key social, psychological,

ethical and design issues, pp. 63–74. John Benjamins Publishing, Amsterdam (2010)
5. Floridi, L., Sanders, J.W.: On the morality of artificial agents. Mind. Mach. (2004). https://doi.org/10.1023/B:MIND.0000035461.63578.9d
6. Moor, J.H.: The nature, importance, and difficulty of machine ethics. Intell. Syst. IEEE **4**, 18–21 (2006)
7. Misselhorn, C.: Grundfragen der Maschinenethik. Reclam, Ditzingen (2018)
8. Wallach, W., Allen, C.: Moral machines. Contradiction in terms or abdication of human responsibility? In: Lin, P., Abney, K., Bekey, G. (eds.) Robot ethics. The ethical and social implications of robotics, pp. 55–68. MIT-Press, Cambridge (2012)
9. Wallach, W., Allen, C.: Moral machines: teaching robots right from wrong. Oxford University Press, Oxford (2009)
10. Verbeek, P.P.: Materializing morality: design ethics and technological mediation. Sci. Tech. Human Val. (2006). https://doi.org/10.1177/0162243905285847
11. Vincent, N.A.: On the relevance of neuroscience to criminal responsibility. Crim. Law Philos. (2010). https://doi.org/10.1007/s11572-009-9087-4
12. Shoemaker, D.: Attributability, answerability, and accountability: toward a wider theory of moral responsibility. Ethics **121**(3), 602–632 (2011)
13. Scanlon, T.: Moral dimensions: permissibility, meaning, blame. Harvard University Press, Cambridge (2008)
14. Van de Poel, I., Nihlén Fahlquist, J., Doorn, N., Zwart, S., Royakkers, L.: The problem of many hands: climate change as an example. Sci Eng Ethics (2012). https://doi.org/10.1007/s11948-011-9276-0
15. Matthias, A.: The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol (2004). https://doi.org/10.1007/s10676-004-3422-1
16. Gunkel, D.J.: Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf Technol (2020). https://doi.org/10.1007/s10676-017-9428-2
17. Coverdale, H.B., Wringe, B.: Introduction: nonparadigmatic punishments. J Appl Philos (2021). https://doi.org/10.1111/japp.12499
18. Coeckelbergh, M.: Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics (2020). https://doi.org/10.1007/s11948-019-00146-8
19. Himma, K.E.: Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics Inf Technol (2009). https://doi.org/10.1007/s10676-008-9167-5
20. Strasser, A.: From tools to social agents. Rivista Italiana di Filosofia del Linguaggio (RIFL) (2020). https://doi.org/10.4396/AISB201907
21. Véliz, C.: Moral zombies: why algorithms are not moral agents. AI & Soc. (2021). https://doi.org/10.1007/s00146-021-01189-x
22. Nyholm, S.: Attributing agency to automated systems: reflections on human-robot collaborations and responsibility-loci. Sci. Eng. Ethics (2018). https://doi.org/10.1007/s11948-017-9943-x
23. Deroy, O.: Rechtfertigende Wachsamkeit gegenüber KI. In: Strasser, A., Sohst, W., Stapelfeldt, R., Stepec, K. (eds.) Künstliche Intelligenz—Die große Verheißung. Series: MoMo Berlin Philosophische KonTexte 8, pp. 471–488. Xenomoi Verlag, Berlin (2021)
24. Hauswald, R.: Digitale orakel? Wie künstliche Intelligenz unser System epistemischer Arbeitsteilung verändert. In: Strasser, A., Sohst, W., Stapelfeldt, R., Stepec, K. (eds.) Künstliche Intelligenz—Die große Verheißung. Series: MoMo Berlin Philosophische KonTexte 8, pp. 359–378. Xenomoi Verlag, Berlin (2021)
25. Loh, J.: Roboterethik. Eine Einführung. Suhrkamp, Frankfurt (2019)
26. Alfano, M., Robichaud, P.: Nudges and other moral technologies in the context of power: assigning and accepting responsibility. In: Boonin, D. (ed.) The palgrave handbook of philosophy and public policy. Palgrave Macmillan, Cham (2018)
27. Sparrow, R.: Killer robots. J. Appl. Philos. (2007). https://doi.org/10.1111/j.1468-5930.2007.00346.x