Responsibility assignment won't solve moral issues of artificial intelligence

Abstract: Who is responsible for the events and consequences caused by using artificially intelligent tools? Is there a gap between what human agents can be responsible for and what is being done using artificial intelligence? Both questions presuppose that the term 'responsibility' is a good tool for analysing these moral issues surrounding artificial intelligence. This article will cast doubt on this presupposition and show how reference to responsibility obscures the complexity of moral situations and moral agency, which can be analysed with a more differentiated toolset of moral terminology. It suggests that the impression of there being responsibility gaps only occurs if we gloss over the complexity of the moral situation in which artificial intelligent tools are employed and if – counterfactually – we ascribe them pseudo-agential status.

Keywords: Responsibility; artificial intelligence; Ethical terminology

Introduction

Contemporary debates about the moral issues of generating and using artificially intelligent systems strongly focus on the question of who is responsible for these system's behaviour. A core suggestion is that there is gap between what moral agents can be responsible for (Matthias, 2004), or what they can suffer retribution for (Danaher, 2016) and the behaviour of AI systems. On the other hand, several authors either claim that there is no gap between our practice of responsibility ascription and the behaviour of AI (Köhler, Roughley, & Sauer, 2017; Tigard, 2021) or make suggestions how to modify our practice of responsibility ascription in order to bridge these gaps (Nyholm, 2018a). This article argues that we are looking in the wrong direction: responsibility is not a term suited to solve moral problems surrounding AI. Componential analysis of 'responsibility' as provided by many in the debate should be a reason to drop the unanalysed, general term and use the components instead, at least in ethical theorising.

Attempts to tackle questions regarding the use of artificially intelligent systems exist in a variety of contexts, ranging from autonomous driving (Nyholm, 2018b, 2018c) and the military use of robots (Sparrow, 2007) to epistemic responsibility for artificially intelligent diagnostic systems (B. Heinrichs & Eickhoff, 2020). The results are equally diverse. They range from a complete rejection of the use of artificially intelligent systems due to the impossibility of reliably attributing responsibility, to a rejection or significant revision of the concept of responsibility because it is not suitable for clarifying questions of robot ethics (Loh, 2019).

The diagnosis that technological developments challenge our legal and moral practise, especially that of ascribing responsibility, has predecessors in several previous technologies, starting with the steam engine. More recent examples include the car, and the issue has become most prominent in nuclear energy and genetic technologies. Gifford for example discusses how early automation in automated looms and railroads prompted "the replacement of the preexisting strict liability tort standard with the negligence regime" (Gifford, 2018, p. 1) and how the widespread introduction of automobiles made it necessary to generate "»financial responsibility laws« that required automobile owners to either purchase insurance or to provide proof that they had sufficient financial resources to pay claims" (Gifford, 2018, p. 41 f.). It has been pointed out in these and other contexts that there is a discrepancy between our established practice of attributing responsibility and the moral requirements of current technical developments. This discrepancy or gap is caused by at least two factors: It is caused firstly by the increasingly complex social and institutional contexts in which actions and their consequences can hardly be attributed to a single human actor (Horwitz, 1977). And it is caused secondly by the development of technical systems that increasingly determine the conditions, courses and consequences of actions (Sussman, 2009, p. 107 ff.). The extreme point of both factors can be found in the use of artificially intelligent systems. AI systems are obviously technical support systems for action that strongly influence said action's course. However, they are also – less obviously – integrated into complex social and institutional contexts, insofar as their production, training and even the use of AI systems involves multiple actors in different institutional dependencies.

Saying that AI is a technical support systems for human action narrows down the term's meaning for the current purpose. It is intended to refer to learning systems of narrow or at most moderate generality. Such systems can and often do show superhuman performance in a limited number of particular tasks, but typically do not generalize across tasks (Chollet, 2019). Such systems heavily depend on human intervention during design, training, and validation. They can, however, vary wildly in the need for human intervention in the way they are employed. Some systems such as language transformers heavily depend on human input, others such as automated picture recognition systems can work with automated sensor data acquisition and without human intervention for quite some time. While both remain support systems for human action, one is a close support systems, whereas the other one is a distant one. In contrast, systems with human level ability to generalize and potential superintelligences will probably have to be treated as genuine authors of actions in the full philosophical meaning of the term and thus cease to be a mere support systems to human action (Bostrom, 2014). Thus, they require a different treatment than the one I propose in this paper.

Why responsibility for artificial intelligence is controversial

Controversies about the responsibility for actions with the help of AI systems currently abound, a prime example being the discussion about the use of military robots. These applications are the focus of political attention in the campaign to outlaw killer robots and the core topic of the by now canonical example (Nyholm, 2018a; Tigard, 2021) of philosophical articles on the responsibility for the behaviour of Ais: Robert Sparrow's landmark article, 'Killer Robots'. In this text he asks "who we should hold responsible when an autonomous weapon system is involved in an atrocity of the sort that would normally be described as a war crime" (Sparrow, 2007, p. 62). Sparrow's killer robot thus is a case of distant support systems, i.e., a system comprising robotic parts which has little to no human intervention during its employment. He tries to show that *of all* possible candidates for responsibility, none are fit to take responsibility for actions carried out with the help of semi-autonomous military robots.

Sparrow seems to assume that responsibility always belongs to an individual, and that it depends on the ability to predict or control the event in question. Because of the lack of control over the actions of an autonomous weapon system, Sparrow claims that neither the manufacturer, nor the programmer, nor the commanding officers are in a position to take responsibility for the system's behaviour if it amounts to a war crime. Because responsibility can only be assumed by those who can be punished, the machine *itself i*s not suitable for assuming responsibility. Because it neither suffers nor can its behaviour be corrected by punishment, let alone by praise or blame, it is simply not possible or appropriate to hold the machine responsible. There is what has later come to be called a gap between the events that require moral evaluation and the actions that allow no ascription of responsibility, a *responsibility gap*.

The term 'responsibility gap' has probably been coined in 2004 by Andreas Matthias, who argued that AI systems would create responsibility gaps that could not be bridged by the established notion of responsibility: "If we want to avoid the injustice of holding men responsible for actions of machines over which they *could not have* sufficient control, we must find a way to address the responsibility gap in moral practice and legislation." (Matthias, 2004, p. 183)

Sparrow makes use of the concept – if not of the term – of a responsibility gap. Holding anyone responsible for the action of an autonomous weapon system would, as introduced by Matthias, be unjust because of their lack of control over the systems behavior. Sparrows overall conclusion is straightforward: because it is not possible to reasonably hold anyone responsible for war crimes committed by an autonomous weapon system, but the possibility of holding specific individuals responsible for war crimes is a necessary condition for a just war, the use of autonomous weapons systems in war cannot be justified (Sparrow, 2007, p. 66). They should not be used and probably not be built in the first place.

A similar argumentative structure can be observed in a completely different type of document, namely in industry guidelines for AI engineers. The British Standards and the IEEE Guidelines both note that there is uncertainty in the prediction and control of some AI-based systems. Both argue that this uncertainty results in a gap between the responsibility of engineers (and users) and the real consequences of the machines behaviour. In this they concur with Sparrow. But unlike him – obviously – they do not call for abandoning the technology. Neither do they, however, simply accept the epistemic gap or call for abandoning our practice of ascribing responsibility whenever a person cannot predict and control a device. Rather, they call on engineers to design the systems in question in a way which allows for prediction or at least retrodiction of the system's behaviour (British Standard for Robots and robotic devices, p.5; IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, p. 90 ff.)

The guidance of these institutions serves not only the purpose of creating the best products for the user, but also to protect manufacturers and engineers from legal and moral blame and liability for undesired behaviour of their products. Philosophical authors as well as stakeholder organisations seem to doubt that our contemporary practice of attributing responsibility can solve the moral and legal problems that arise from the use of AI systems. Either unpredictable or uncontrollable AI systems must be dispensed with altogether, or they must be adapted to such an extent that they fit back into our practice of responsibility.

In the following I will argue that these authors are completely right in claiming that our contemporary practice of responsibility ascriptions does not solve the moral issue of how to respond to morally problematic actions which have been committed with the support of or by AI-systems. However, this is not because artificially intelligent systems change the structure of action or the requirements of moral evaluation in some fundamental way, but because the term 'responsibility' is not suited for detailed moral analysis of complex actions.

The claim that these moral problems of AI-assisted actions cannot be solved given our current practices of ascribing responsibility is due to the fact that 'responsibility' is a bundle term with too many internal tensions. The internal tensions of the term are adumbrated by the fact that it allows for a certain dismantling of distinctions, as can be seen in a prominent application to artificial intelligence: „There are many ways agents are held *responsible* for their actions within legal systems, corresponding to different available means of *punishment*. Human agents have historically been punished in a variety of ways: through infliction of pain, social ostracism or banishment, fines or other confiscation of property, or deprivation of liberty or life itself. Debates over whether it makes sense to hold (ro)bots *accountable* for their actions often center on whether any of these traditionally applied punishments makes sense for artificial agents." (Wallach & Allen, 2009, p. 208, my emphasis)

This short passage showcases how several distinct moral properties and relations – here liability, punishment, and accountability – are gathered under the term 'responsibility'. By trying to solve moral issues of artificial intelligence with this term, we try to solve several highly specialised tasks with one rather crude tool.

'Responsibility' - etymological structure and functional role

The concept 'responsibility' has originally been modelled in analogy to the interaction between a judge and a defendant. Its first philosophical occurrences in Reid[1] and Kant's Metaphysics of Morals refer to the person giving an account of her actions. This situation at first appears to be morally pretty straightforward. One person has to answer for her previous actions to another. This prototypically clear distribution of obligations and entitlements is what guides the use of the term. As such, 'responsibility' is a term used for the coarse-grained moral task of identifying who has to give reasons for his or her actions to whom (Brandom, 1998; Bert Heinrichs & Knell, 2021; Sellars, 1997).

Contrary to appearance, even this seemingly straightforward moral situation is characterised not by one but by a variety of moral relations between the persons involved. It involves mutual recognition, mutual attribution of moral status and states, and several concrete moral demands and obligations such as duties that one person has towards the other, claims that she can make on him, the justification that she owes him, the reparation that may be due, and so on. These moral relations can and need to be analysed with the complex conceptual tool box of ethical theory, involving terms such attributability, accountability, liability, culpability, individual and group agency, etc.

Using the term 'responsibility' is common in many contemporary moral discourses about AI as exemplified above in Sparrow's article, Wallach and Allen's book, the British standards, and the IEEE guidelines. The following analysis will try to show that this is using the term beyond its capability, i.e. beyond the coarse grained task of identifying a judge and a defendant.

Prototype and variations of attribution of responsibility

In the current literature, 'responsibility' is differentiated into different meanings of the term and into different relations it refers to. According to the first type of differentiation going back to Hart's seminal work, being responsible for something can for example mean that someone has a duty to

[1] Reid in his d's Essays on the Active Powers of the Mind uses the term 'accountability, but his argument is taken up by Mill who uses 'responsibility' as a synonym (McKeon, 1957, p. 7).

do something, or that he or she is liable for something (Hart, 1968; Haydon, 1978). The second differentiation distinguishes the relata of the relation of responsibility, i.e. who can be responsible for what to whom, according to which norm and for whose sake (Lenk & Maring, 1991; Ropohl, 1994). These two differentiations are drawn together in newer analysis (van de Poel, 2011; Vincent, 2011). The term 'responsibility' thus refers to a broad cluster of practices that answer the very general question "In what way (1) is who (2) responsible for what (3) to whom (4) according to what norm (5) and for whose sake (6)?". Each interrogative term in this question can have different answers depending on the context, but some answers form the core of our way of speaking, others are variations or even deviations.[2] In the following, I will provide some detailed support for the thesis that the prototypical answer to all six interrogative words is modelled on the simplified dialogical situation and refers to a specific component of the bundle term 'responsibility'. I'll show that, in contrast, analysing moral issues of artificial intelligence requires a different model than the dialogical situation and the employment of a multitude of moral relations, not just one. To show how this is the case I will focus on the answers to the first three interrogative terms.

There is another tradition of analysing the concept of responsibility, which dominates the writings of Köhler and colleagues (Köhler et al., 2017) as well as Tigard (Tigard, 2021). Tigard in particular, following David Shoemaker, distinguishes the concept of responsibility into accountability, attributability, and answerability, rejecting a gap in the respective practices of responsibility: "Like our accountability practices toward fellow humans, we can hold AI to account by imposing sanctions, correcting undesirable behavioral patterns acquired, and generally seeing that the target of our responses works to improve for the future – a bottom-up process of reinforcement learning." (Tigard, 2021, p. 604)

Both are versions of pluralism concerning moral responsibility, and they overlap in some of the components of responsibility. However, there are mismatches in nomenclature – e. g. what Tigard calls 'answerability' is here subsumed under 'accountability' – and in depth of analysis. While Tigard uses a tripart distinction of the types of responsibility into attributability, accountability and answerability, the present differentiation goes beyond an analysis into types and identifies slightly more types, too.

[2] A pragmatic prototype conception of concepts will be used here, i.e. I will understand concepts as linguistic tools and their meaning as determined by a prototypical use and variations (Lakoff, 1987; Rosch, 1975; Wittgenstein, 1997)

(1) The different answers to “Responsible in what way?”

“Responsible in what way?” can have several different descriptive and normative answers. In the following, I will only discuss the normative versions. It should, however, be mentioned that the prototypical descriptive answer refers to being a cause: rain is responsible for the road being wet. This is a common way of speaking and is presupposed in many (not all) normative judgements of responsibility.

The normative dimensions of the term ‘responsibility’ are given by three retrospective dimensions accountability, praise- and blameworthiness, liability on which I'll focus in the following and two prospective ones, obligation and virtue (van de Poel 2011; Vincent 2011). Most of these pick out a separate dimension of the dialogical model of ‘responsibility’. If one person approaches another on grounds of her actions, she can ask for justification, praise, or blame the person or assign some kind of liability.[3]

The prototypical answer to ‘responsible in what way’ is, in my opinion, 'accountable'. Here is why: Being accountable is a necessary precondition of the further moral relations subsumed under ‘responsibility’. To be responsible in the sense of being accountable means to be able and to be expected to give reasons for, i.e., to justify past actions. Someone who is accountable in this sense can be asked about their actions in direct interaction and *answer, i. e. give an account* why they said or did something (cf. Bovens, 2007). Being asked for and giving reasons for one’s behaviour is the entrance ticket to every form of rule-governed interaction, a sufficient (and maybe necessary) condition for taking part in discourse. Being able to give justifications and getting challenged to do so is, moreover, the core characteristic of rational agents, i.e. of beings who can act for reasons (Brandom, 1998; Sellars, 1997). This is why responsibility in the sense of accountability and ability to justify oneself/one’s actions is sometimes also equated with the ability to act, sometimes even with personhood. Moral discourse necessarily requires the ability for justification, for giving and taking reasons (Bert Heinrichs & Knell, 2021), the ability to receive or give praise and blame or to assign or accept liability fully depend thereon.

[3] Another prominent approach to further analyse the ‘in what way’ bundle of ‚responsibility‘, the one used by Tigard, is introduced in (Watson, 1996). Watson distinguishes between accountability and attributability, with accountability being used more or less as above, attributability as the ability to be subjects of moral appraisal (p. 240). Shoemaker (Shoemaker, 2011) in turn focusses on the distinction between attributability, answerability and accountability. According to him attributability is the ability to get attributed a predicate which expresses one's practical attitudes, answerability is the ability to give the reasons with which one justified one's conduct and accountability is the “capacity to recognize and appreciate the demands defining the various relationships as reason-giving” (p. 631). For the present purpose, I will leave attributability to one side. This is not to say it is not an important dimension of our moral practice and the reactive attitudes we have towards others. However, it is more of a precondition of holding someone responsible in any meaning of the term rather than one of the components of responsibility. Thus, it is not an answer to the question in what way someone can be held responsible, rather it is an answer to the question who can be held responsible and thus will be discussed below.

There is a danger of confusing accountability and explainability. Being accountable implies being the kind of person who can provide reasons for one's behaviour, not explaining its causal antecedents. When asked "Why did you do that?" the explanation "the combination of internal states $x_{1..n}$, input $y_{1..n}$ and the mechanisms $z_{1..n}$ operating on the former caused this behaviour as an output" is not a valid reply. It might at best be the start of one, but lacks any justificatory relevance. It fails to place the action in question in a net of mutually accepted norms of conduct.

As mentioned, the moral relation of *praise- or blameworthiness* closely depends on the that of *accountability.* It does so because praise and blame have – at least among other things – the function of correcting behaviour. If a being cannot respond to praise or blame with changes in behaviour, it is not a candidate for them (apart from providing an outlet for the blamer's anger). Modifying one's own behaviour on the basis of praise and blame in turn requires the ability to act for reasons. This is different from the mere ability to modify one's behaviour on the basis of punishment and reward. The former is possible for discursive beings only, the latter for discursive as well as non-discursive beings. Responsibility in the sense of praiseworthiness or blameworthiness makes a being an active member of the moral community in the sense that it can apply moral reasons to its actions.

Related is the understanding of 'responsibility' as *liability*, i.e., that individuals can be expected to compensate or apologise for consequences of their behaviour (Capes, 2019). The extent of an agent's liability typically depends on particular norms and their goal. It will for example make a difference whether norms are designed for the aim of retribution or for the aim of betterment. In the former case the dialogical situation might be a suitable model, in the latter case a more complex social arrangement needs to stand in as a model. In any case, however, taking an agent to be liable presupposes a certain set of mental capacities, without which the application of the norm becomes pointless (Vincent, 2011, p. 25). This presupposition of certain mental capacities is what Sparrow referred to when he claimed that autonomous weapon systems are not suited to take responsibility because they do not suffer and thus cannot be punished. Even liability is closely connected to the dialogical model for the term 'responsibility'. We usually talk about agents being liable only if they can in some minimal way grasp the norms for the breach of which they are liable. That implies that these norms have at least to be possible reasons for action for this agent. They must be able to relate their own behaviour to these norms when asked to justify their conduct. It does not suffice if the agents are merely conditioned according to these norms.

In the debate about *how* someone – anyone – could be responsible for the behaviour of artificially intelligent agents, two answers dominate: first, a deviant form of accountability for the system itself. As already mentioned, industry standards recommend that artificially intelligent systems be designed in such a way that their actions can be traced at any time. The same idea lies at the heart of the explainable AI movement. However, the system does not give reasons for actions, but is designed to make the causes for its behaviour transparent – or worse: allows for another – often equally opaque – system to make the causes for the original behaviour transparent (cf. Tigard, 2021, p. 602; Zeiler, Krishnan, Taylor, & Fergus, 2010). As noted above, the explanation of causes is not a justification with reasons. At best, the people who created or used the system are enabled to consult its logs in *their* reasons for continuing to operate, modify or shut down the system (Rudin & Radin, 2019). Thus, this deviant case runs into a dilemma: Either the information about the causes of behaviour is not an answer to the normative "Why did you do this?" question. Then making the system explainable tracks causation, which is at best a nonnormative version of 'responsibility', not a case of accountability. Alternatively the information is taken to be an answer to a normative question, but then it is a question aimed at the engineers, providers, and users of the system: "Why was the system trained on these data?", "Why were these outputs during training considered correct or acceptable?", "Why was the system employed in this environment and for this purpose?" etc. Then this is not accountability of the system but of the people creating, selling, or employing it.

Second, there is regular discussion about who is liable for the events in which artificially intelligent systems are involved and their consequences. One proposal is a form of group liability. The idea is to generate a joint liability by producers, distributors, users etc. of AI systems, by generating a money fund from which damages caused by the system – and possibly fines incurred – are to be compensated (Beck, 2016). This is a borderline case of responsibility as liability for two reasons. First, it is an extremely narrow and purely legal scope of liability. It lacks any regard for genuinely moral reactions and for behavioural change. Second, and this point is closely related to the next interrogative term, it involves a deviant answer to the question "Who is liable for actions involving artificially intelligent systems?". It suggests a form of group responsibility, but even that would be a deviant case, simply because the group in question does not fit the standard conditions for any type of group agency and responsibility. There is barely a joint context of action, no self-identification of an acting body or anything similar (Crone, 2020). The only connection between the members of this liable group is some involvement with the product.

From the description it should have become clear that applying the term 'responsibility' in both cases obscures more of the moral relations than it reveals. This is not to say that the solutions to specific moral questions of dealing with AI assisted actions are bad. They are not. But they are oversimplified or misdescribed by the term 'responsibility'. Neither is the pursuit of explainable AI simply a case of accountability or of any other normative version of responsibility, nor is the complex legal constellation of joint liabilities for consequences of AI employment adequately described by the undifferentiated 'responsibility'. Both questions, that concerning accountability and that concerning liability, involve not just one way of being responsible, but several; and they both refer to a situation which simply cannot be modelled on a dialogical situation but always involves more parties. That leads us to the second interrogative term, the 'who' of responsibility.

(2) The different answers to "Who is responsible?"

The *subject of responsibility* which the interrogative clause 'who is responsible?' inquires about is the responsible person from whom a reaction or response is required. The prototypical subject of responsibility is a typical interlocutor in rule-governed discourses. These are the individuals we usually turn to when we ask why someone did or said something, blame them or even demand compensation or punishment.

This characteristic of being an interlocutor is mirrored in what has come to be called attributability, i.e. the ability to be attributed moral properties independently of external ascription (Tigard, 2021; Watson, 1996). Another human being can correctly be attributed practical reasons, reactive attitudes, or virtues and vices. The moral properties of algorithms on the other hand are not attributes of the models themselves, but of their being designed in a certain way and being placed in a specific use context for a particular purpose. It is to the combination of (human) design, employment, and purpose that moral properties can be attributed, not to the algorithm itself.

The prototype of typical interlocutors obviously does not exhaust the range of possible subjects of moral relations. The range of possible subjects of responsibility – in any meaning of the term – is typically described by a set of characteristics which need to be correctly attributable to the system in question, such as rationality, receptivity to reasons, receptivity to correction through praise and blame or punishment, etc. Since many of these properties come in degrees, it is plausible that being able to stand in the moral relation in question is either itself a gradual or a threshold phenomenon. Nor must all variations of responsibility be alike in this regard, for example, legal practice seems to know cases of full responsibility (liability, ability to give informed consent) given a certain threshold of competency, while at least educational practice sees responsibility (accountability) as a gradual phenomenon that adolescents acquire successively. Depending on which characteristics at which threshold are constitutive for the ability to stand in a given moral relation, it is a matter of dispute whether, for example, some animals, human infants, people with severe mental disabilities or artificially intelligent systems can be part of the relation in question. On the other hand, moral relations can tie together more than just two individuals. For example, one agent can be morally obligated towards a number of different individuals within the same relation, e.g., the obligation to respect one's parents, both of them. There is also the special case of group responsibility. Admittedly it is an as yet undecided question in philosophy whether moral rights or obligations of groups, which are firmly established in the legal context, have a moral equivalent that cannot be traced back to the rights and obligations of individual group members (cf. Crone, 2020).

In the debate about attributing responsibility for the behaviour of artificially intelligent systems, two marginal answers to "Who is responsible?" dominate. The first of these has been discussed in science fiction scenarios since at least Samuel Butler's novel Erewhon (Butler, 1872): the artificially intelligent system itself. As seen above in Sparrow's example, some philosophers mention this option, only to dismiss it immediately. This seems to be the reaction of most authors in this field: artificially intelligent systems are rarely seriously discussed as bearers of responsibility. Others take is slightly more serious and metaphorically talk of the duty of a programme to solve certain tasks, or to adhere to certain limits. This usage is even found in the prominent philosophical contribution by Wallach and Allen, who speak of designing systems in such a way that they do not transgress concrete norms of action. But in the end they admit that artificially intelligent systems are at best deviant cases of responsibility (Wallach & Allen, 2009). Some very few pick a special aspect of the term 'responsibility', namely the descriptive use as in 'causally responsible' and on this basis ascribe responsibility to artificially intelligent systems (Floridi, 2013).

The second major strand of attributing responsibility for AI systems constructs a form of group or systems responsibility for the consequences of actions carried out with the help of artificially intelligent systems. As mentioned above, Beck for example constructs a rough analogy to the concept of a legal person and proposes to generate a special legal status for artificially intelligent systems. It is intended to consist of the partial responsibilities of all parties involved in the creation and use of the artificially intelligent system: "This legal person for robots would only be the bundling *of* all the legal responsibilities of the different parties (users, sellers, producers, etc.). This bundling is actually the main reason why a new classification for these machines is necessary." (Beck 2016, p. 479). In philosophy, a similar suggestion has been made by Sven Nyholm, who argues that not much of a responsibility gap remains if we understand responsibility as a relation property of the network or system of agents and tools involved in a certain behaviour or event, what he calls a human–machine collaborations (Nyholm, 2018a). But as Nyholm himself admits, his suggestion involves a significant re-conceptualisation of 'responsibility' and 'agency'. In the terms introduced here, it drives the use of the term further apart from its etymological origin instead of trying to salvage the dialogical constellation.

Nyholm introduces several illuminating differentiations in the terms of agency and of responsibility. In particular he differentiates agency into *individual* versus *collaborative*, into *domain specific* versus *domain independent* as well as into *basic* versus *principled* and into *supervised versus deferential* and *responsible* (Nyholm, 2018a, p. 1207 f.). Contemporary and near future machine learning systems which Nyholm exemplifies with automated cars and military robots have "supervised and deferential agency […] of a collaborative type" (Nyholm, 2018a, p. 1211). Given that the agency of artificially intelligent systems is collaborative, defers to and is supervised by a human in the collaboration, Nyholm can assign responsibility to the pair or group of collaborators but place the locus of responsibility with the human part. Admittedly, Nyholm can thereby show that there is not much of a responsibility gap, but he does so by showing that the humans designing, training, selling, using artificially intelligent systems are the responsible parties: "The most difficult questions here instead concern what humans are most responsible for any potential bad outcomes caused by their robot collaborators." (Nyholm, 2018a, p. 1214)

What is more, Nyholm succeeds in bridging the gaps in moral relations because he further differentiates the concept of responsibility. The humans designing, training, selling, using artificially intelligent systems are not just plainly responsible, but rather we should try "to determine who is responsible for what aspects of the actions the car performs" (Nyholm, 2018a, p. 1214). Thus, Nyholms project to vindicate the use of the concept of 'responsibility' for artificially intelligent systems only succeeds by a) differentiating types of responsibility and b) assigning it to the humans involved with the systems. The underlying idea of Nyholm's solution to dealing with moral issues raised by the use of AI-systems is quite convincing, namely, to consider them either collaborative tools or scaffolds of human agency (J.-H. Heinrichs, 2020; Hernández-Orallo & Vold, 2019). But it is, contrary to appearance, not a solution which relies on the bundle concept of 'responsibility'.

In general, analyzing moral issues of artificial intelligence with the term 'responsibility' which is modelled on a dialogical situation tends to obstruct the view of the complex contexts in which, as Nyholm demonstrates, these issues occur. The moral issues of artificially intelligent systems typically involve more than just two persons, and most of the time there is not neat grouping of the affected and involved persons into collective agents, much less into two parties.

(3) The different answers to "What is someone responsible for?"

The *object of responsibility* is the issue addressed in the demand for response, what the subject is responsible for. In everyday language, very different things can take the place of the object in 'responsible for ...', for example physical objects, actions, states of affairs. A person can be responsible for his car, for the care of his children or for the safety of a workplace. However, many authors think that these different kinds of things a person can be responsible for can be reduced to only one or two kinds, namely actions and possibly the success of actions. The undisputed prototype of the object of responsibility is a previous action.

The person responsible for his or her car is, according to reductive positions, actually responsible for certain actions concerning her car: for regular maintenance, for driving it in a certain, safe way, for parking it only in marked parking spaces and so on. It is an open question whether people can be responsible for the success of actions or only for the action itself: Does a person who is responsible for the safety of a workspace have to remove obstacles, provide guidance, etc., and has fulfilled his responsibility when he has performed all these tasks? Or is she responsible for ensuring that the workspace is safe beyond these actions and has neglected her responsibility if someone is harmed in this workplace, even though all specific measures have been carried out to make the workplace safe? The answer to this question depends on the exact meaning of 'responsibility' as introduced above. For example, a person may be legally (and even morally *(Capes, 2019)*) *liable for* events that have occurred, even though they have performed all the actions they were *obligated to* perform.

The discourse of responsibility for behaviours of artificially intelligent systems is fundamentally at odds with this conceptual prototype for two reasons. First, because it moves events into the focus, which definitely are not actions in any but a very lose sense, namely the data processing and output of computer systems. An action is behaviour carried out on the basis of practical reasons, which in turn are composed of some pro-attitude like a wish or desire and some instrumental belief. Even if it is convenient to describe the behaviour of AI-systems as actions, this is merely borrowing the psychological terminology. It might from an intentional stance make sense to ascribe beliefs to AI-systems. However, ascribing them pro-attitudes is at best a derivative use of the term, if not just a figure of speech.

This derivative use has become deeply entrenched in our description especially of reinforcement learning and tree search algorithms. We talk about them maximising a reward function or a preference function respectively. The term has been imported from behavioural sciences, be it animal behaviour studies or behavioural economics where organisms' behaviours are explained in the same terms, namely as reward and preference functions. We should, however, differentiate between notion of maximising a reward or preference function and that of being susceptible to rewards (Silver, Singh, Precup, & Sutton, 2021) or having preferences, though these are easily confused. While it might be possible to describe animal behaviour within a given domain as guided by the maximisation of reward functions for the purpose of scientific explanation, this abstraction leaves out quite a bit of relevant psychological information. The same is true for describing behaviour, especially human behaviour, as guided by a preference function. This might be a useful abstraction, but it captures only a minor part of what it means to have a preference. The mere fact that AI-systems can be designed making use of these explanatory models of psychological and behavioural science doesn't mean that these systems have the same type of state as the organisms so described. It is one of the scientifically exciting aspect of AI-models in science that they do not need to imitate the structure of the phenomenon which they explain (Napoletani, Panza, & Struppa, 2016). This holds equally true for models of the human mind.

It is exactly for this reason that ascriptions of cognitive and conative states or of agency to AI-systems are carried out with great care to what *exactly* is being ascribed. Take the examples of Nyholm: "Domain-specific principled agency: pursuing *goals* on the basis of *representations* in a way that is regulated and constrained by certain rules or principles, within certain limited domains." (Nyholm, 2018a, p. 1208, my emphasis) or Floridi: "*Agent* $=_{\text{def.}}$ a system, situated within and a part of an environment, which initiates a transformation, produces an effect, or exerts power on it over time." (Floridi, 2013, p. 140, my emphasis). Only few authors would insist that a human agent's pro-attitudes and beliefs together with the resulting actions are of the same kind as an AI-system's goals or objectives and representations together with the resulting behavior (Wooldridge, 2003).

The second reason why the discourse of responsibility for behaviours of artificially intelligent systems is fundamentally at odds with the conceptual prototype is because it rarely refers to the real actions involved in the moral issues raised by artificially intelligent systems, namely the actions of individual human agents. This is probably the most disconcerting diagnosis of the current debate. The question whether a software engineer is accountable for the way he programmed and trained an AI system, a manager liable for ordering or selling it, or an officer blameworthy for the order to use an autonomous weapon system is rarely asked, although these are the basic moral phenomena we should inquire about. In contrast, the question people in fact ask is whether any of these individuals can be responsible *tout court* for an artificially intelligent system or the events brought about by such a system. Several authors wonder whether it is the software engineer who is responsible for an artificially intelligent system or whether the commanding officer is responsible for the behaviour of an autonomous weapon system or to what degree – as a contrast to: in what way – each of these are responsible.

What does the concept of responsibility do for the moral analysis of the use of artificially intelligent systems?

Two things should have been particularly noticeable in the exposition of the concept of responsibility and its particular use in discourses on artificially intelligent systems: firstly, that the concept of responsibility contains an internal tension, between the simplified dialogical prototype and the very differentiated practice of attributing responsibility. Secondly, that the practice of using the term, i.e., of attributing responsibility, draws on the entire set of instruments for answering normative questions, at least in ethics and law. It tries to reconstruct the different moral relations of the dense network of obligations, duties, expectations, liabilities etc. and model them on the dialogical situation between typical moral subjects. Identifying different components or versions of 'responsibility', such as accountability-responsibility or responsibility as duty, is first and foremost a task of isolating distinguishable moral relations between different constellations of agents. Analysing a complex moral situation such as that of actions involving AI-tools according to these distinguishable moral relations is the logical next step. Unsurprisingly, this works better for some novel moral relations and worse for others. However, these analyses are at best misnamed as investigations of responsibility for the actions of AI, when in fact they show that the undifferentiated term 'responsibility' itself does no analytical work in this regard.

Let me put it into a rough analogy: the original endeavour (moral analysis) is characterised by a complex set of subtasks (identifying different moral relations) for which we need different specialised (terminological) tools. These tools are amongst others: 'right', 'duty', 'justification', 'liability' 'recompensation', etc. It turned out that we can package a relevant number of these tools together in the multitool 'responsibility' – much like a swiss army knife. And much as in real multitools, we do not assemble a full version of the specialised tool, but one that fits into the casing. Here, the casing is the dialogical construction of the term 'responsibility'. Now we analyse that multitool and generate differentiations such as 'liability responsibility'. But these primarily refer to the multitool with a certain sub-tool extended. Liability responsibility is nothing over and above liability and responsibility as obligation is nothing but obligation. When our original endeavour takes a critical and partially unfamiliar turn (in the analysis of moral issues of artificial intelligence), we try to stick to our multitool instead of opening the full toolbox, although we encounter gaps in the multitool's applicability: responsibility gaps.

At no point in the above analysis of the attribution of responsibility did it turn out that the term does more than the differentiated set of instruments. The moral and legal questions regarding the production and use of artificially intelligent tools by persons in their institutional embeddings can be solved with the more differentiated conceptual tools of the relevant disciplines, it cannot with the multitool 'responsibility'.

Deborah Johnson has made a similar observation, when she approached our practice of ascribing responsibility to engineers: "Although there is broad consensus that engineers have social responsibilities, what is owed in the name of social responsibility is not well understood." (Johnson, 2017, p. 85). After analyzing the codes of several engineering societies which predominantly talk in terms of obligations and duties, she suggests settling for an analysis of engineers' responsibility as accountability, which in turn consists of obligations according to shared norms. Instead of sticking to the blanket identification of the involved parties as 'engineers' and 'the public' she disentangles who exactly is involved in the social practice of accountability. If this is the path set out for our analysis of responsibility for artificially intelligent systems and their use, we would do better so skip the detour and use the more specialized terminology from the outset. To return to Sparrow's example: Instead of talking about the responsibility for the events caused by an autonomous weapons system, we should ask for the diverse moral rights and obligations of everybody involved. Here are some: Does a commanding officer have the duty to prevent any action carried out by his troops and with his equipment, which would amount to a war crime? Is she blameworthy if her troops or their equipment are involved in a violation of the codes of just war? Should she feel regret or blame herself? Can she be excused, or her guilt be mitigated if she fulfilled all her obligations of oversight? Can victims or their relatives make claims against her in that case? Is a company liable for damages caused by their goods if there is no fault on the side of the user, in this case the military unit? Is the company obligated to inform their user about the possibility of unforeseen events and the risks thereof? Is a company which produces equipment that is likely to be involved in events that count as war crimes morally blameworthy? Is a military or political leader morally blameworthy for equipping a military with tools which can even under good conditions cause events which would have to be described as war crimes? I have little doubt that all these questions can be answered – without gaps – given some specifics of the cases under scrutiny.

Under which circumstances would the analysis with the full toolbox of ethical terminology result in gaps? I think this question is diagnostic for the debate about responsibility for artificially intelligent systems. The answer seems to be that gaps would occur, if there were a subject of some moral relation, which for some reason cannot stand in this relation, e.g., a subject of obligation which cannot have obligations. The thought behind this is, that artificial intelligence can be understood as an agent, but it cannot be understood as a subject of moral relations that come with being an agent. I think the problem is with the former and not with the latter part of this thought. Only if we think of artificial intelligence systems as pseudo-agent, do responsibility gaps occur. The disconcerting fact, however, is that at the moment an artificially intelligent system becomes an agent, there are no responsibility gaps anymore. An agent is a person who can be involved in all the moral relations packaged under the term 'responsibility'. Pseudo- and non-agents, however, simply do not stand in any type of moral relation, they can merely be the subject matter of moral relations between agents. The latter are typically not analyzed with the more general term 'responsibility', but with the more specialized components. There is no reason to do otherwise if artificially intelligent tools are being used by these agents.

Conflict of Interest Declaration:

The author declares that there is no conflict of interest.

Literature

Beck, S. (2016). The problem of ascribing legal responsibility in the case of robotics. *AI & SOCIETY, 31*(4), 473-481. doi:10.1007/s00146-015-0624-5

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework1. *European Law Journal, 13*(4), 447-468. doi:https://doi.org/10.1111/j.1468-0386.2007.00378.x

Brandom, R. (1998). Action, Norms, and Practical Reasoning. *Nous, 32*(Supplement 12), 127-139.

Butler, S. (1872). *Erewhon*. London: Trubner & Co.

Capes, J. A. (2019). Strict Moral Liability. *Social Philosophy and Policy, 36*(1), 52-71. doi:10.1017/S0265052519000220

Chollet, F. (2019). On the measure of intelligence. *arXiv:1911.01547*.

Crone, K. (2020). Foundations of a we-perspective. *Synthese*. doi:10.1007/s11229-020-02834-6

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299-309. doi:10.1007/s10676-016-9403-3

Floridi, L. (2013). *The ethics of information* (First edition. ed.). Oxford: Oxford University Press.

Gifford, D. G. (2018). Technological Triggers to Tort Revolutions: Steam Locomotives, Autonomous Vehicles, and Accident Compensation. *Journal of Tort Law, 11*(1), 71-143. doi:doi:10.1515/jtl-2017-0029

Hart, H. L. A. (1968). *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Oxford University Press.

Haydon, G. (1978). On being responsible. *Philosophical Quarterly, 28*(110), 46-57.

Heinrichs, B., & Eickhoff, S. B. (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Hum Brain Mapp, 41*(6), 1435-1444. doi:10.1002/hbm.24886

Heinrichs, B., & Knell, S. (2021). Aliens in the Space of Reasons? On the Interaction Between Humans and Artificial Intelligent Agents. *Philosophy & Technology*. doi:10.1007/s13347-021-00475-2

Heinrichs, J.-H. (2020). Artificial Intelligence in Extended Minds: Intrapersonal Diffusion of Responsibility and Legal Multiple Personality. In B. Beck & M. Kühler (Eds.), *Technology, Anthropology, and Dimensions of Responsibility* (pp. 159-176). Stuttgart: J.B. Metzler.

Hernández-Orallo, J., & Vold, K. (2019). *AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI*. Paper presented at the Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA. https://doi.org/10.1145/3306618.3314238

Horwitz, M. J. (1977). *The Transformation of American Law, 1780–1860*: Harvard University Press.

Johnson, D. G. (2017). Rethinking the Social Responsibilities of Engineers as a Form of Accountability. In D. P. Michelfelder, B. Newberry, & Q. Zhu (Eds.), *Philosophy and Engineering: Exploring Boundaries, Expanding Connections* (pp. 85-98). Cham: Springer International Publishing.

Köhler, S., Roughley, N., & Sauer, H. (2017). Technologically blurred accountability? Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Debiel (Eds.), *Moral Agency and the Politics of Responsibility* (pp. 51-67). London: Routledge.

Lakoff, G. (1987). *Women, fire, and dangerous things. What categories reveal about the mind.* Chicago: University of Chicago Press.

Lenk, H., & Maring, M. (1991). Deskriptive und normative Zuschreibungen von Verantwortung. In H. Lenk (Ed.), *Zwischen Wissenschaft und Ethik*. Frankfurt am Main: Suhrkamp.

Loh, J. (2019). Responsibility and Robot Ethics: A Critical Overview. *Philosophies, 4*(4), 58. Retrieved from https://www.mdpi.com/2409-9287/4/4/58

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175-183. doi:10.1007/s10676-004-3422-1

McKeon, R. (1957). The Development and the Significance of the Concept of Responsibility. *Revue Internationale De Philosophie, 11*(39 (1)), 3-32. Retrieved from http://www.jstor.org/stable/23940271

Napoletani, D., Panza, M., & Struppa, D. C. (2016). Is Big Data Enough? A Reflection on the Changing Role of Mathematics in Applications. In P. Mircea (Ed.), *The Best Writing on Mathematics 2015* (pp. 293-304): Princeton University Press.

Nyholm, S. (2018a). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics, 24*(4), 1201-1219. doi:10.1007/s11948-017-9943-x

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass, 13*(7), e12507. doi:https://doi.org/10.1111/phc3.12507

Nyholm, S. (2018c). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass, 13*(7), e12506. doi:https://doi.org/10.1111/phc3.12506

Ropohl, G. (1994). Das Risiko im Prinzip Verantwortung. *Ethik Und Sozialwissenschaften, 5*(1), 109-120.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7*(4), 532-547. doi:https://doi.org/10.1016/0010-0285(75)90021-3

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review, 1*(2). doi:10.1162/99608f92.5a8a3a3d

Sellars, W. S. (1997). *Empiricism and the Philosophy of Mind.* Cambridge, Mass. / London: Harvard University Press.

Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics, 121*(3), 602-632. doi:10.1086/659003

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence, 299*, 103535. doi:https://doi.org/10.1016/j.artint.2021.103535

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy, 24*(1), 62-77. doi:10.1111/j.1468-5930.2007.00346.x

Sussman, H. (2009). *Victorian Technology: Invention, Innovation, and the Rise of the Machine.* Santa Barbara, Cal, Denver, Col, Oxford: Praeger Publishers.

Tigard, D. W. (2021). There Is No Techno-Responsibility Gap. *Philosophy & Technology, 34*(3), 589-607. doi:10.1007/s13347-020-00414-7 van de Poel, I. (2011). The Relation Between Forward-Looking and Backward-Looking Responsibility. In N. A. Vincent, I. van de Poel, & J. van den Hoven (Eds.), *Moral Responsibility: Beyond Free Will and Determinism* (pp. 37-52). Dordrecht: Springer Netherlands.

Vincent, N. A. (2011). A Structured Taxonomy of Responsibility Concepts. In N. A. Vincent, I. van de Poel, & J. van den Hoven (Eds.), *Moral Responsibility: Beyond Free Will and Determinism* (pp. 15-35). Dordrecht: Springer Netherlands.

Wallach, W., & Allen, C. (2009). *Moral machines. Teaching robots right from wrong*. Oxford ; New York: Oxford University Press.

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics, 24*(2), 227-248. Retrieved from http://www.jstor.org/stable/43154245

Wittgenstein, L. (1997). Philosophische Untersuchungen. In *Schriften* (Vol. 1). Frankfurt am Main: Suhrkamp.

Wooldridge, M. (2003). *Reasoning about Rational Agents*: MIT Press.

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010, 13-18 June 2010). *Deconvolutional networks*. Paper presented at the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.