Bad Arguments for Responsibility Sharing

Pekka MÄKELÄ and Raul HAKLI

Abstract. We study whether humans and technological artifacts, such as robots, can form hybrid agents that would be fit to be held morally responsible for their actions. We give an argument against such a possibility and criticize several arguments that defend this possibility. We identify three argument forms that have been employed to argue for shared responsibility between humans and machines: (1) Argument from gradual properties, (2) Argument from responsibility gaps, and (3) Argument from extended agency. We analyse these arguments and aim to show that they are invalid.

Keywords. social robots, human–robot interaction, hybrid agents, responsibility, agency, extended agency, shared responsibility, collective responsibility

1. Introduction

It has been noted that the amount of automation is increasing in our societies. Algorithms and robots are operating and making decisions in areas that used to be controlled by humans alone, for instance, in stock trading, medical diagnosing, and car driving. The question has been raised what happens to responsibility when human beings give away control (see, e.g., [1] for recent discussion). Traditionally, control and responsibility have been closely related: One cannot be responsible for things over which one has no control. Knowledge has been seen as another necessary requirement of responsibility. Given that many of these algorithms are based on artificial intelligence and machine learning techniques, which make their operation unpredictable, humans cannot even know what kinds of actions will be performed and what their consequences might be. Hence, both traditional conditions of responsibility, control and knowledge, are out of reach for human beings, including the designers, builders, controllers, owners, and users of those machines. What makes things worse is that a collectivization of responsibility does not seem to help much: There is no collection of human beings that would possess the necessary knowledge and control of the algorithmic decisions.

Hence, some have suggested that the responsibility for the actions of the machines should be attributed to the machines themselves. However, it is not easy to see how that could be done. It is not even easy to understand what it could mean in practice to hold a robot responsible. The problem is that even if we were to admit that machines are agents and capable of possessing responsibility-relevant control and knowledge of their actions,
they do not seem to satisfy the general conditions of agents fit to be held responsible for their actions. Some often mentioned conditions are autonomy, self-awareness or consciousness, sensitivity to moral reasons, capability to experience emotions, etc. There is no consensus on what the exact conditions of moral agency are, but the majority view regarding current robots and AI agents is that they are not moral agents, and henceforth not fit to be held morally responsible.

However, it has been argued, that even if robots and AI agents are not agents capable of bearing full responsibility for their actions, they could still be held partially responsible. The idea would be that moral responsibility would go together with causal responsibility: Full moral responsibility for certain harmful consequences would be attributed to the collection of humans and machines whose contributions caused those consequences. Insofar as machines are part of this collective, they would be partially responsible for the consequences. They could be seen as relatively autonomous members of responsible groups [2]. At least prima facie this collective could be argued to satisfy general conditions of moral agency (because it contains people who are uncontroversially moral agents) and the particular conditions of responsibility for its actions (because it contains machines with the relevant control and knowledge of those actions), even though it would require sophisticated argumentative skills to construe such an argument without falling prey to the fallacy of composition.

In this paper, we critically scrutinize some arguments that have been presented to the effect that machines, even though not capable of carrying full moral responsibility, could participate in such human–machine hybrid agents and hence be partially responsible or share responsibility together with humans for the actions of the hybrid agent. Here we will not argue that robots cannot be full moral agents (but see [3] for our argument to that effect). Instead, we will take that as a premiss and defend the conditional claim that if robots cannot be full moral agents, then they cannot be partially responsible either.

A rather generic motivational background for writing this paper relates to our worry concerning the extended usage of important normative notions such as moral responsibility and moral agency. More generally, such normative notions that play an important constitutive role in our social practices which we find worthy of maintenance. With respect to such notions we tend to prefer a cautious, conservative if you like, approach. Our worry is that rushing into extended usage of such notions may have detrimental effects on normative practices that have a significant role to play in the lives of human communities. Our view is that before moving to extended usage, we need to first see how far the established meanings can go in terms of understanding and explaining new phenomena such as artificial intelligence and social robotics.

The situation with artificial agents constructed by technology can be compared with artificial agents constituted by collections of human beings. Think of the tentative risk of accepting corporate agents as members of moral community, as moral agents in their own right. If we hold corporate agents morally responsible in their own right, independently of the moral responsibility of the individual human agents in the corporation, it may happen that individual role-holders in the corporation find it tempting to hide behind a corporate veil—indeed feel incentivized to erect such a veil (via practices internal to the corporation). The result of the acceptance of new members to the moral community may turn out to be counterproductive with respect to the realization of the aims of the practice of holding agents responsible, namely, the aim of directing and changing the behaviour,
which in the case of corporate agents can only take place via the behaviour and actions of the individual human agents.

We proceed by first sketching a positive argument for why technological artifacts can not be attributed even partial responsibility for their causal contributions to actions attributable to collections of humans and artifacts in Section 2. Then we scrutinize some recent arguments for such attributions of partial or shared responsibility and try to show that they are invalid in Section 3. Finally, we conclude in Section 4.

2. Positive argument against responsibility sharing

In our view when we talk about responsibility sharing we talk about joint moral responsibility which in turn is moral responsibility for joint actions. Roughly speaking, a joint action can be understood thus: Two or more individuals perform a joint action if each of them intentionally performs an individual action but does so with the (true) belief that in so doing they will jointly realise an end that each of them has [4]. On this view joint responsibility is ascribed to individuals. Each member of the group is individually morally responsible for the outcome of the joint action but each is individually responsible jointly and interdependently with the others.

In this view, each participant in the joint responsibility bears individual moral responsibility for his or her actions. As moral responsibility is an agentic notion, that is, it presupposes moral agency, it follows that each participant in moral responsibility must be a moral agent. In [3], we have argued that robots and other pieces of technology such as AI algorithms are not autonomous and hence not moral agents. Roughly stated it tries to establish that moral responsibility requires autonomy and that programming is the kind of manipulation that undermines the autonomy of robots and AI agents. If our previous argument based on Mele’s history-sensitive account of autonomy [5] is correct, it follows that robots and AI algorithms cannot be participants in joint responsibility either and that we cannot share responsibility with them. Obviously, the claim that moral agency requires autonomy is contestable, but the justification for such a requirement stems from our practice of holding agents responsible. This is a moral practice, and it would not be fair to hold agents that fall short of autonomy morally responsible.

Robots and AI algorithms cannot be properly seen as group members in a group that is morally responsible unless they are moral agents themselves. In our view, attributing collective responsibility to a group entails attributing the individuals full responsibility for their part-actions and joint responsibility for the collective action. Joint responsibility in turn means that each participant is conditionally responsible: One participant is responsible only if other participants are also responsible. Hence, responsibility sharing can only take place among full moral agents.

3. Unpromising ways to argue for responsibility sharing

We identify three argument forms defending the possibility of shared responsibility between humans and machines. In what follows, we analyse these arguments in detail aiming to show that they are invalid.

**Argument from gradual properties:** As a preliminary step to arguing that responsibility can be shared with robots, it has been argued that responsibility is something
that can be shared. One way to do that is by arguing that moral responsibility or moral agency is a gradual matter. According to what we call the argument from gradual properties, the properties and capabilities that are required for moral agency are gradual properties, hence moral agency itself is something gradual. Just as a child gradually becomes a moral agent through learning and development, a robot might be taught moral reasoning and moral decision-making, and hence gradually approach full moral agency. There are more and less sophisticated versions of this argument; one goes as follows [6]:

On the possibility of ascribing responsibility to artificial systems—autonomous driving systems [...] it is important to take into consideration that these three sets of capacities (communication, autonomy, and judgment), and the competences that come with them, vary, to say that someone is more or less able to act in a specific situation, more or less autonomous, more or less reasonable, and so on, it follows that responsibility itself must be attributed gradually according to the aforementioned prerequisites.

The conclusion does not follow: There can be properties that are not gradual, even though they are based on only such properties that are gradual. As a simple example, consider adulthood. In most countries there is a certain legal age, for example, 18 years, before which one is not considered an adult and after which they are. Adulthood such understood is not a gradual property even though it is defined on the basis of age which is a gradual property. Versions of the argument that are based on the analogy between raising children and teaching morality to robots can also be questioned by pointing out that children whose moral understanding is still developing need not be understood as partial moral agents. Our educational practices of teaching our children that they can be “blamed” or “punished” for their behaviour can be understood as treating them as if they were moral agents in order for them to eventually—but not necessarily gradually—become moral agents [7]. Moral agency can be understood as a categorical property obtained once certain threshold levels of gradual properties like moral understanding have been achieved. Even though an agent can be partially responsible for particular acts or outcomes, this presupposes fitness to be held responsible, which we take to be a categorical property: One either has it or not.

**Argument from responsibility gaps**: In the literature concerning moral responsibility the concept of responsibility gaps can be defined as follows [8]: “The responsibility gap refers to the discrepancy between the ethical responsibilities managers must take upon themselves, on the one hand, and the realities of the human functioning that limit our human capacity to meet those responsibilities, on the other.” This is clearly compatible with the agentive conception of responsibility, because the gap is between the load of ethical responsibilities imposed on the moral agent by an institution or structure or practice and the capacities of the same agents to fulfill those responsibilities. In the literature it has been noted that there are situations in which there appears to be more responsibility to be shared than can be borne by the moral agents engaged in the situation. The question of responsibility gaps has been discussed mostly in the context of collectives and corporations, but also in the context of autonomous computer systems and robots [9,10,11].

The argument from responsibility gaps starts from the observation that robots cause morally bad things to happen and hence someone or some thing should be responsible for them. Due to unpredictability and other problems, oftentimes there is no human being who could be held morally responsible. Also since the robot itself cannot be held fully morally responsible, responsibility must be distributed to the collective or network
consisting of humans and possibly machines. The obvious point against this argument is that there are many other morally bad things that happen for which no-one, no thing, and no collective can be held morally responsible, for instance, people getting killed in earthquakes. Natural catastrophes aside, there are also unintended consequences of collective actions for which it can be claimed that there is no (individual or collective) agent that could be held morally responsible, for instance morally bad consequences of the market system. Hence the argument rests on a false premiss according to which responsible agents exist for all morally bad things.

One way to deal with such responsibility “surplus”, perhaps more used in the literature on computer systems than in literature concerning collective responsibility and corporations, is to extend the use of moral responsibility as a term to actors that do not qualify as moral agents. However, our incapacity to control an actor, system, or mechanism, does not turn the actor, system, or mechanism in question, into an entity capable of bearing moral responsibility nor does it turn the causal responsibility of such entities into moral responsibility. Calling the responsibility moral does not make it moral. Doing so is an argument from equivocation. Here is an example from Johnson and Powers [12]:

Our point in noting the complexities of tasks being divided among humans and computer systems [...] is to note that what is at issue is a reconfiguration of role responsibilities. As air traffic control becomes more automated, human controllers will have to acknowledge that the computer system shares the tasks formerly assigned only to them. [...] Holding systems “responsible” – in the sense Hart introduced in his analysis of role responsibility – seems to follow from the re-assignment of tasks. [...] In this context, it makes sense to speak of the role responsibility of the computer system because it has taken over morally-relevant tasks.

Their idea is taken up by others as well [13]:

We argue that to address the question of ascribing moral responsibility to intelligent systems we must adopt the functionalist view and see them as parts of larger socio-technological systems with distributed responsibilities, where responsibility of a moral agent is a matter of degree. From such a standpoint ascribing responsibility to an intelligent system has primarily a regulatory role. Delegating a task to a machine is also delegating responsibility for the safe and successful completion of that task to the machine [...]. A machine that takes care of certain tasks intelligently, learning from experience and making autonomous decisions gives us good reasons to talk about a machine as being “responsible” for a task in the same manner that we talk about a machine being “intelligent”. No doubt, technology is morally significant for humans, so the “responsibility” for a task with moral consequences could be seen as moral responsibility.

In our view, causally taking care of tasks of moral significance (safety, health, etc.) is to be distinguished from being morally responsible for tasks being accomplished.

In the case of responsibility gaps we rather need to analyse the systems and their organization and improve the match between the moral responsibilities and the capacities of moral agents. If we create systems over which we do not have the kind of control required by moral responsibility, we are responsible for such creation over which we do have control, and consequently indirectly responsible for states of affairs brought about by such systems when let to function in the wild. This is the human predicament.

Indeed, calling actors that do not qualify as bearers of moral responsibility “moral agents”, is a way to make matters fuzzier rather than clearer, and this strategy of equivocation may turn out to be counterproductive with respect to our aims of making the world
better for human beings by way of utilizing new technology. Responsibility attributions must be fair and justified—our practice of holding entities responsible is not a channel for our moral outrage or other moral emotions.

We agree that the possibility of responsibility gaps is indeed a problem, but its nature is practical instead of ethical and it requires practical and legal remedies rather than widening the extension of our moral concepts. Here we willingly identify ourselves as blunt Finnish fellow travellers of Sven Nyholm who writes:

[...] we should not be too quick to attribute agency of a very independent sort to robots like automated cars or military robots. Rather, we should stop to think more carefully about (a) whether any of these robots are really exercising agency in any real sense and if so, (b) what kind of agency it makes most sense to attribute to automated robots. These questions are crucial to ask if we are to avoid the sort of worries about responsibility-gaps that we find in the literature on autonomous machines [...] It is primarily if it would make sense to attribute a strongly independent form of autonomous agency to automated systems and robots that these kinds of worries of responsibility-gaps can plausibly be thought to arise. The mere presence of unpredictability and a lack of direct control are not by themselves enough to create responsibility-gaps.

**Argument from extended agency:** According to the argument from extended agency, the locus of agency should not be understood as an individual human being, but as the individual human being extended with various technological tools whenever these tools are necessary for performing certain actions. The basic idea is familiar from discussions on extended mind, but it has been put to a new use in the context of responsibility: If the agent of an action is such an extended agent then the responsibility for the action should be attributed to the extended agent as well. Here is a straightforward formulation [14]:

To bring this into the practical realm, consider the well known phrase “guns don’t kill people; people kill people.” That is an absurd thing to say because it falsely implies that the possession of a gun is not pertinent to killing, or to the subject that possesses it. Of course guns do not go around shooting people all by themselves, but everybody knows that a person with a gun is a far greater threat to kill someone than a person without one. Selinger and Engström (2007, pp. 575–576), following Ihde (1990, 2002), explain that human beings are changed when they use certain technologies: a man-with-a-gun is a different being or subject than the same man without a gun. Similarly, it is commonly said that some people are transformed when they get behind the wheel of a car. Or consider the saying, “if all you have is a hammer, everything looks like a nail.” That is, the possibilities for action depend not just on human beings, but on the available means of action as determined by the relationship of humans with technology and/or other components of extended agencies. The entity that acts—the subject—is the extended agency (person/gun, person/hammer), not just the human individual. [...] If moral responsibility for an act lies with the subject that undertakes it, and if the subject includes nonuman as well as human beings, then so may moral responsibility.

It can be admitted that the tools we have affect us as subjects and so does the environment in which we are, but it is not clear whether they affect us in a way that turns us into participants in joint agency; at least the conditions of the above definition of joint action are not satisfied. The effect of the tools and environment can rather be understood as modifying our affordances and our psychological state. In fact, we would claim that this is the standard way of understanding action involving tool use. Both in philosophical action theory and in everyday language use we would rather say that a person killed
someone with a gun, rather than a person and a gun killed someone or that a person killed someone together with a gun. Of course, both the standard view in philosophy of action and our ordinary understanding may be faulty; however, the point is that there is no valid argument from extended agency to shared responsibility as there are other possible ways of understanding responsibility attributions in cases of extended agency that do not involve attributing partial responsibility to technological items.

4. Conclusion

In this paper we have discussed the possibility of responsibility sharing between human and technological artifacts. We have provided an argument for thinking that such responsibility sharing is not possible. We have also identified several instances of unsuccessful argument types for responsibility sharing appearing in the literature. In our view, and in the view of quite a few others, moral responsibility is an agentive notion: It presupposes moral agency. Previously, in [3], we have argued on the basis of Mele’s history-sensitive account of autonomy [5] that robots are not moral agents for their lack of autonomy, and here we have argued that also responsibility sharing can take place only between moral agents.

While discussing the arguments that we find faulty, we have expressed two kinds of worries concerning tendencies to extend the use of our moral vocabulary, one which is methodological or analytical, and the other which is normative or moral-political. The first worry concerns the key concepts such as agency, responsibility, and autonomy. When it comes to agency, it is a continuum notion. There are very different kinds of agents. For instance, there are canine agents that are capable of intentional action but which fall short of being moral agents. Then there are moral agents, typically exemplified by adult human agents.

As well as the sense of agency, also senses of responsibility vary and in our view these different senses should be kept separate. There is causal, normative, role, prospective, retrospective, individual, and collective responsibility. Also autonomy can be understood in various ways: There is the AI sense of autonomy and the moral sense of autonomy as self-rule or self-government [5]. When extending the use of these senses, we run the risk of equivocation. The point is not linguistic policework but the worry of losing the analytical clarity in the study of these highly relevant and extremely complicated phenomena.

The moral or political worry concerning the extended use of the normative key concepts is that such an extension may have unforeseen consequences to our social practices and communities in which these key concepts play an important role. We suggest that we should move cautiously here. It is the moral responsibility of the researchers of these issues. This paper is a part of a larger project in which we search for arguments that make the extension of such notions as moral agency and moral responsibility necessary. We are still looking.

References


