Can we forgive a robot?

Michael Nagenborg

Author's manuscript. Final (minor) revisions and edits are missing.

Introduction

If we hold a robot fully responsible for its actions, how should we deal with that robot if it did something wrong? For example, if a robot murders a human being, should it be punished just like a human being who commits the very same crime? Can we actually punish a robot? In "A Body to Kick, but Still No Soul to Damn" (2012), Peter Asaro argues that "various forms of corporal punishment presuppose additional psychological desires and fears – concerning pain, freedom of movement, mortality and so on" (Asaro 2012, p. 182). Along similar lines, Robert Sparrow (2007) has pointed out that to consider an act as punishment, we need to assume that the machine suffers from the punishment: "Indeed, the suffering involved in a machine being punished must be such that if we discover that it was in fact innocent […] then we feel that we have done it a serious wrong and owe it recompense" (Sparrow 2007, p. 72). Hence, for scholars like Asaro (2012) and Sparrow (2007), robots should not be held responsible for their actions since there would be no adequate response to unwarranted and undesired actions.

There are at least two ways to respond to the position that robots can't be punished since they do not suffer: (1) One could argue, like Daniel Dennett (1998), that it seems plausible that some robots may indeed have something like artificial sensations. (2) One could argue, like Gert-Jan Lokhorst and Jeroen van den Hoven (2012), that a bearer of responsibility may very well lack the capability to suffer and that punishment is only one way to respond to criminal offenders, namely: treatment.

While I do agree, that we need to think through how we can react to the wrong-doings of a robot *if* we are willing to hold the machine responsible for its actions, I will explore a different perspective. Instead of asking, if we can punish a robot, I will ask if we can forgive a robot.

The background of my inquiry is that forgiveness plays a crucial, yet often neglected role in human-human interactions. It seems reasonable to assume that forgiveness will play a similar role in a society where humans and robots coexist and, at least, some of these robots are held responsible for their actions. To avoid misunderstandings, I am not claiming that such robots do currently exist, and they may very well never exist. The following chapter is a speculative exercise to grasp what it could mean for human beings to live together with such machines, which is motivated by the claims of some scholars, that one day we may consider robots as (artificial) moral agents.

In a first step, I will introduce the complementary role of promises and forgiveness in Hannah Arendt's work.

Forgiving a person

In "The Human Condition" (1998, Original: 1958), Hannah Arendt highlights the complementary role of promises and forgiveness (Arendt, 1998, pp. 236–247). Arendt limits her reflection on forgiveness to the realm of politics due to the non-instrumental nature of *action* in contrast to *labor* and *work*. Thus, the outcome of (political) actions is not predictable. This is also due to the fact that actions always involve interactions with other persons, who – according to Arendt – are free to make their own decisions and, thus, unpredictable.

The ability to make and accept promises is a central mechanism for dealing with the basic unpredictability of human beings. Through these acts, we "islands of security" in the "ocean of uncertainty, which the future is by definition" (Arendt, 1998, p. 257). To put it the other way around: Since human beings are free to make their own choices, a human being's activities are not predetermined and, hence, unpredictable. Political actions require human beings to act jointly, which in turn requires various actors to become predictable. By making promises the individual limits her options for actions. While a human being might be capable of doing X or Y, a person can promise to do X rather than Y. Hence, her behavior becomes more predictable.

The possibility of forgiving a person for breaking a promise is central to Arendt since it is not always possible for people to keep promises. Persons may even break a promise for good and overriding reasons. If we could not forgive a person for breaking a promise, we only could react negatively. Hence, a broken promise would limit our own options for action – that is: our freedom – because another person's misconduct would force us to punish that person (at

least, if the breaking of a promise caused serious harm). The same applies if we make a promise:

"Without being forgiven, released from the consequences of what we have done, our capacity to act would, as it were, be confined to one single deed from which we could never recover; we would remain the victims of its consequences forever, not unlike the sorcerer's apprentice who lacked the magic formula to break the spell" (Arendt, 1998, p. 237).

In the absence of the option to forgive a person, we would not only have to sanction the breaking of the promise. Having to sanction or punish another person may also very well undermine our opportunities for future interactions, since those who we made suffer will remember the suffering.[1] Forgiving, in contrast, allows for a new beginning.

I would like to broaden Arendt's considerations beyond the scope of (political) action, since the argument for the need of an alternative to punishment (in the broadest sense) seems to be applicable for all forms of human activity characterized by the giving and receiving of promises. For example, any act of delegating a task X to another person builds on the promise of that person to do X. Furthermore, promises can be considered an integral part of trust relationships, when we trust that another person or entity will do X to avert harm to us (see Nagenborg, 2010). Such a trust relationship is hard to explain without an (at least implicit) promise. From an Arendtian perspective, one can argue that these promises, too, serve to reduce the level of uncertainty by limiting the possibilities of the legitimate acts of the trustee: We trust someone to do X and not Y. However, since a person may not be able to keep her promises (and, thus, breach our trust), the possibility to forgive becomes essential in such cases, too.

The crux is that human beings are in principle unpredictable, because they could always act differently. Promises create security by voluntarily restricting the options for action by those who make a promise. Through this voluntary restriction, we enable joint actions and the delegation of tasks. However, if we could only respond to non-fulfillment of promises with sanctions (punishment in the broadest sense), we lose precisely what we wanted to gain in the act of promising: our freedom.

1 As I will argue later, it would be equally problematic if we could not punish and would always be forced to forgive. However, in the latter case there would be no suffering of the offender involved. Thus, it is reasonable to distinguish the problem of inflicting harm from the problem of having a limited set of options for reacting to an offense.

Forgiveability as a condition of responsibility

As we have seen, forgiveness is central to the coexistence of human beings. By analogy, I would like to argue, that forgiveness is of similar importance for the coexistence of humans and robots, if we consider robots to be responsible for their acts.

One may object, that we need to discuss first whether robots can be responsible for their actions at all before asking about the possibility to forgive a robot. After all: As long as there is no claim that robots are responsible, there is no need to discuss the option to relieve a robot from the consequences of its actions.

Scholars who argue in favor of considering robots as moral agents (that is: as bearer of responsibility) at times make strong claims. For example, Keith Abney states, that "one day robots could become moral agents, and, so, full moral persons" (Abney 2012, p. 50). Authors like Colin Allen and Wendell Wallach (2012) defend the weaker thesis, that robots may not have "full moral agency" (for now), but still can be regarded as "artificial moral agents" (with a special emphasis on "artificial"). As we will see later, such a weaker position is not unproblematic either.

One of the challenges in dealing with such claims is that they are always concerned with future developments, when it comes to full moral agency. Thus, we are forced to argue for or against the idea of a responsible robot on a very abstract and fundamental level. The high level of abstraction in the discussion on the conditions for ascribing personhood as well as the contemporary latent discomfort about anthropocentrism and speciesism prove to be of little help here, too. I actually do believe that we should be a little more relaxed about the question whether computers can think. As Alan Turing (1950) has noted in his seminal paper, we can never be 100% certain that other people actually are conscious. Yet, we are so polite to assume that they are. Why should not we do the same thing when dealing with machines? Nevertheless, I am reluctant to regard robots as moral agents, and I would like to pinpoint my discomfort with the question of whether we can forgive robots.

The question concerning the possibility to forgive a robot is relevant not only in view of moral agency. It already becomes relevant once we claim that we *trust* a robot (e.g., Coeckelbergh, 2012). As I have argued above, breaching trust opens up the possibility for the trusting person to forgive the trustee. If trustworthiness should mean more than reliability, how should we react, for example, to a robot's failure as a care-giver? Can we forgive a smart car for killing people? We certainly forgive some human drivers for their mistakes, even if

they caused serious harm. A human being confronted with a real-life trolley problem, would certainly have our sympathies since the person was confronted with a hard and unavoidable choice and, thus, we might be willing to forgive the driver, even if we don't approve the choice made.

Failing to carry out a delegated act may often lead to less dramatic results. Still, the delegation of task is a crucial aspect of human-robot collaborations (Nyholm, 2018). Thus, we need to ask how to react to a robot, which fails to carry out a delegated act. For now, the default option is to either stop using the robot or to redesign the robot. Hence, the default option reveals that we do not treat the robot as a moral agent. The responsibility of its actions fully lies with the user and/or the designers.

If we would consider a self-driving car to be responsible for killing a human being, we may very well be forced to punish the car – or be able to forgive it. While I am not concerned here with the conditions under which we can consider a robot to be responsible for its actions, I would like to argue in the following, that we should not ascribe responsibility to a robot unless we are willing to allow for the option to forgive it.

For lack of a better word definition, I will refer to the "possibility of forgiving X" as "forgiveability of X." If I can forgive someone (or something) that someone (or something) is *forgiveable*. Note that "forgiveability" is a property of the entity to be forgiven and as such is independent from the particular act to be forgiven. Based on this definition of forgiveability, I would like to argue that forgiveability is a necessary condition for ascribing (moral) responsibility to an entity. In other words, I would like to raise the bar in the discussion about robots as bearers of moral responsibility.

Forgiveability as a litmus test of the human-machine relationship

From a philosophical perspective, the question of forgiveability of a robot becomes interesting only if we exclude two trivial cases:

(1) The robot is de facto considered as a mere machine.

(2) The robot has all the relevant characteristics of a human being (or, person, if you prefer).

Regarding (1): If the robot is a strictly deterministic machine, there is no uncertainty regarding its behavior. Accordingly, it is not necessary to constrain its freedom by a promise. Strictly speaking, it would even be impossible to restrict its freedom, because such a machine

is not free to make any decisions. And while we may, for example, forgive our old car for letting us down again, we would only do so in a metaphorical sense.

Regarding (2): This case is, of course, trivial only in the moral, but not in the technical sense. But if there are certain qualities of the person that identifies a representative of our species as a person and, thus, as a bearer of responsibility, we should apply the same criteria to nonhuman individuals, too. Furthermore, the criteria should be applied regardless of the specific criteria to all kinds of nonhuman individuals. In other words, once we have established such criteria, we should be able to argue about the moral status of robots as we can argue about the personal status of monkeys.[2] Finally, it seems reasonable to interact with a robot which has all the relevant qualities in the same way as we do with fellow human beings. Hence, if a robot meets all the criteria and does something wrong, we may forgive that robot as we would forgive a human being. This is not to say, that forgiving a person is a trivial act. But there would be no need for additional considerations about the forgiveability of a robot that is just like a human in all relevant regards. Thus, I consider a robot which meets all the relevant criteria for personhood as a trivial case in our discussion.

Accordingly, I will focus on the case of a robot that is more than a mere machine but does not have all the necessary qualities of a person. What should we expect from such a machine? A preliminary suggestion: First, the robot should be able to make decisions and its behavior is not determined solely by external influences. Second, the robot should have reasons for making its choice.

The first capability mirrors the unpredictably of human behavior. Just like a human being might have to choose between doing X and doing Y, the robot should be able to make a decision about doing X or Y. In order to raise the bar in terms of 'freedom of choice' not too high: If freedom of choice manifests itself in uncertainty about future action (the system could always behave differently), then it may be sufficient if the behavior of the system is unpredictable for a human observer. We could, of course, predict the behavior of a strictly deterministic system *if* we had all the relevant information about the environment, the exact construction and the current state of the system. However, this is often not the case because both the systems and their environment are too complex. The non-trivial case of a robot that is more than a mere machine can be understood as an opaque technology in the sense of Sherry

2 In view of the current state of technological development the question regarding the moral status of monkeys seems more relevant to me. Clement (2013), e.g., provides a good overview of the question of moral agency of nonhuman animals.

Turkle (2005): a machine that invites us to explain its behavior in psychological rather than in mechanistic terms.[3]

However, the difficulty in predicting the behavior of the system may not yet entitle us to qualify the robot as a bearer of moral responsibility. A robot that leaves its decisions to pure chance would undoubtedly be a source of uncertainty but would not show any degree of autonomy (in the sense of self-governance).[4] Yet, the same applies for human beings. A human person who would *only* act on random (e.g., by rolling a dice) would not be considered as a moral agent either.[5] At the very least, we would expect from an autonomous person that she can give reasons for her behavior (Nagenborg, 2005, p. 69). Hence my proposal, for the second requirement for robots which are more than mere machines: To have comprehensible and justified reasons for actions.[6]

Since it has been assumed that the robot is more than a mere machine yet different from a person, such a machine presents a challenge to the idea of forgiveability. As I have stated earlier, we may accept that a person had reasons to break a promise, which he or she considered to be good reasons at the moment of decision making. One of the reasons, why we may forgive a human being making such a decision, is that we can understand the reasons for breaking the promise. We may say something like: 'If I had been in your position, I probably would have acted like you for similar reasons. That does not make your behavior any better, but it is understandable to me.' If a robot is different from a human being, these considerations no longer apply. It is precisely the common human nature that enables us to forgive human beings who act on similar reasons like us.[7] In contrast, we may simply not

3 Of course, in the case of an opaque technology we can always switch back to the modernist perspective and try to understand the system's behavior in mechanistic terms. The point is, we usually do not do that.

4 Such a robot could not even be called 'autonomous' in a technical sense, since deterministic machines can't generate random. Computers always rely on the measurement of hard-to-predict events in their environment, when random numbers are needed. Random decisions of the system are therefore always heteronomous since the source of the random numbers is located outside the system.

5 While it can be perfectly fine in certain context to take a random decision in specific contexts (e.g., by having a lottery), I am here thinking of a person making *all* decisions on a random base. Think DC Comics' Two-Face taken to the extreme.

6 I acknowledge the temptation to introduce criteria for autonomy or personhood, which can only be met by humans (at least, for now). However, a little caution seems to be required, because it is not at all clear that we as humans can always meet such criteria.

7 There are, of course, those cases, where our understanding of human behavior fails. The radical evil is characterized precisely by the fact that we can no longer understand how a human being can do such a thing. Franziska Dübgen (2016) therefore reminds, following

understand *why* a robot took a certain decision, because there is no guarantee that good reasons for such a robot will be the same as good reasons for a human person. And what about cases, where a human being simply failed to uphold her promise? Humans – at times – fail. Failing is part of our common nature. What about robots? It seems hard to understand, what it means for a robot to fail upholding a promise in a similar way that human beings fail.

If, however, we should find one day that people sincerely begin to forgive robots, it would be a strong indication of a fundamental shift in the relationship between humans and robots – and to that extent the act of forgiveness seems to be a good litmus test of whether or not we are considering a robot as more than a mere machine. Such a machine would be admirable not only from a technical point of view, but it also arouses our philosophical interest because it is hard to imagine – at least for me – what would characterize such a machine and what it would be to co-exist with such machines.

Furthermore, following Hannah Arendt (1998), it would be worthwhile to inquire deeper into such acts of forgiving if they should ever occur. For Arendt, the central element in the act of forgiving is not the deed, but the person to be forgiven; and forgiving the person is a way to recognize that person: "Forgiving and the relationship it establishes is always an eminently personal … affair in which *what* was done is forgiven for the sake of *who* did it" (Arendt 1998, p. 241). To me, the difficulty of understanding how to forgive a robot is rooted in our inability to grasp the essence of a robot which is more than a machine and, yet, not a person. But if we were to begin to forgive machines, then perhaps in the act of forgiving the robot, we would start recognizing them for what they are.

Forgiveness as virtue

Another difficulty in thinking about what it means to forgive a robot lies in the fact that forgiving is a sovereign act. In this respect, no criteria can be specified for forgiving a person or a robot. As I explained earlier, the purpose of forgiveness for Arendt is to provide an alternative to punishment, so that we are not forced by the wrongdoing of another person to react in a certain way. Being able to but not having to forgive, allows and requires a free

Jacques Derrida (2001), that there are "those atrocities that cannot be forgiven, which at the same time demand forgiveness" (Dübgen, 2016, p. 19, my translation). See also Arendt's remarks on "radical evil," where she observes: Men "are unable to forgive what they cannot punish and that they are unable to punish what has turned out to be unforgivable" (Arendt 1998, p. 241).

choice. But this freedom would be lost if, under certain circumstances, we were forced to forgive wrongdoing. Offenders may hope for forgiveness but not demand it.[8]

But even if we interpret forgiveness as a sovereign act, there might be good reasons to forgive someone or something. For example, we may no longer want to make our behavior dependent on the past. Forgiving someone can be a way to break free and make room for a new beginning – with or without the forgiven.

Furthermore, we should bear in mind that punishment is a morally ambivalent act. In essence, punishment means making individuals suffer. To intentionally inflict harm on another being, however, remains morally problematic, even if the act is justified.[9] In the absence of forgiveness, we would be forced to put ourselves constantly in a morally ambivalent situation. In that sense, forgiveness seems to be a virtue in dealing with robots, not only for the sake of the robot, but also for our own sake.

Finally, if certain robots are responsible but not forgiveable, it is aggravating that these robots will become second-class moral agents. After all, we will still be able to forgive people. Thus, we would live in a world where there are two types of moral agents: those who are forgiveable and those who must be punished. We would create an unbridgeable gap between the two groups, not because we exclude such robots from the outset as a moral responsibility bearer. On the contrary, we create the gap by attributing responsibility to certain robots while being unwilling to forgive them. Hence, we should be very careful in considering alternative approaches to moral agency that fall short of ascribing full moral agency to robots. I would rather suggest abstaining from considering anyone or anything to be responsible unless we are willing to forgive that individual.[10]

Conclusion

The attribution of moral responsibility without forgiveability is a risky enterprise. If we cannot forgive and always have to punish wrongdoing, we bring ourselves into a morally

8 In view of freedom, *having to forgive* is as problematic as *having to punish*. In both cases, our behavior would be predetermined by the behavior of the offender. It is the ability to either forgive or punish that enables and requires a free choice.
9 The problem is often not evident to us because we have delegated punishment to the state.
10 It might be reasonable, to disguish between two different types of moral agency and two corresponding responsibilities: human moral agency and robo-moral agency which comes with robo-responsibility. Still, I would like to uphold my claim that we should not ascribe robo-responsibility to a robot if we are not willing to ascribe to it also robo-forgiveability.

ambivalent situation. Even if the joint action of humans and robots, at first sight, may make possible new options for action – and thus extends our freedom of action – the necessarily asymmetrical relationship threatens to undermine our freedom again and again.

'Forgiveness' is certainly a complex phenomenon, and I do not claim to have done justice to all its aspects. But I hope to have demonstrated that 'forgiving' provides us with an excellent lens to think through human-technology relations. I focused in this chapter on robots, because of the single-sided interest of other scholars on the question of punishment. This made me wonder, if it shouldn't be possible to give robots a break and not to focus solely on punishment. Given the role of promises in some forms of trust relationships, however, it might be feasible to raise similar questions about other technologies as well.

While I have argued elsewhere, that we should restrain from talking about 'robots' in general (Capurro & Nagenborg, 2009), I decided to leave aside details and specific context in this chapter. The decision was partly due to the speculative nature of this chapter but is also caused by the difficulties to understand, what a robot that is more than a mere machine and, yet, not a person may look like. A philosophical discussion about agency, responsibility, and forgiveability of robots seems pointless to me, if we merely assume that robots are not (or too) different from human beings (or any other kind of person).

Finally, I have shown that it is not so easy to understand what that means, that we forgive a robot. This is undoubtedly due first of all to the complexity of the 'forgiveness' phenomenon, but also due my explicit assumption that it should be a robot that is more than a machine but does not share all the relevant characteristics of humans. To reiterate, the discussion about the attribution of moral responsibility to a machine seems philosophically useless if we do not assume that this machine is different from humans. However, it is not easy to explore this difference and make it tangible. That is why we should be distrustful of the supposedly weaker demand that robots be moral agents of their own kind (artificial moral agents, in the sense of Allen and Wallach, for example). The moral status of these creatures seems to be unclear at the moment. It is not just a question of which requirements such a machine must meet, but also of the question of what it could mean for us to co-exist with such creatures. The act of forgiveness offers an excellent focus to pursue this question.

Acknowledgements

An earlier version of this chapter has been published in German:

Cordula Brandt, Jessica Heesen, Birgit Kröber, Uta Müller und Thomas Potthast (Hrsg.): *Ethik in den Kulturen – Kulturen in der Ethik*. Tübingen: Narr Franke Attempto 2017, S. 291-300. ISBN 978-3-7720-8611-3.

I would like to thank Cordula Brandt for all her feedback on the German version. The extended English version has been presented at the Dutch-Japanese Workshop on Philosophy of Technology in Sendai in summer 2018. I would like to thank all the participants for their encouraging and useful feedback. Finally, I would like to thank Melis Baş for taking her time to discuss Hannah Arendt's work with me.

References

Abney, K. (2012). Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 35–52). Cambridge, MA, London: MIT Press.

Allen, C., & Wallach, W. (2012). Moral Machines: Contraction in Terms or Abdiction of Human Responsibility? In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 55–68). Cambridge, MA, London: MIT Press.

Arendt, H. (1998). The human condition. 2nd edition. Chicago and London: Chicago University Press. (Originally published: Chicago: University of Chicago Press, 1958.)

Asaro, P. (2012). A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. In: P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 169–186). Cambridge, MA, London: MIT Press.

Clement, G. (2013). Animals and Moral Agency: The Recent Debate and Its Implications. *Journal of Animal Ethics*, 3(1), 1–14.

Coeckelbergh, M. (2012). Can we trust robots?. *Ethics and information technology*, 14(1), 53–60.

Dennett, D. C. (1998). *Brainchildren: Essays on Designing Minds*. Cambridge, MA: MIT Press.

Derrida, J. (2001). *On Cosmopolitanism and Forgiveness*. London, New York: Routledge.

Dübler, F. (2015). Grenzen der Versöhnung? *Polylog*, 34, 13–25.

Lokhorst, G. J., & Van Den Hoven, J. (2012). Responsibility for military robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 145–156). Cambridge, MA, London: MIT Press.

Nagenborg, M. (2005). *Privatheit unter den Rahmenbedingungen der IuK-Technologien*. Wiesbaden: VS Verlag.

Nagenborg, M. (2010). Vertrauen und Datenschutz. In M. Maring (Ed.), *Vertrauen* (pp. 153–167). Karlsruhe: KIT Scientific Publishing.

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, 24(4), 1201–1219.

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, 24(1), 62–77.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 49, 433–460.

Turkle, S. (2005). Computer Games as Evocative Objects. In J. Raessens, & J. Goldstein (Eds.), *Handbook of Computer Game Studies* (pp. 267–279). Cambridge, MA, London: MIT Press.