Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: http://www.tandfonline.com/loi/uaai20

Responsibility and the Moral Phenomenology of Using Self-Driving Cars

Mark Coeckelbergh

To cite this article: Mark Coeckelbergh (2016) Responsibility and the Moral Phenomenology of Using Self-Driving Cars, Applied Artificial Intelligence, 30:8, 748-757, DOI: 10.1080/08839514.2016.1229759

To link to this article: http://dx.doi.org/10.1080/08839514.2016.1229759

APPLIED ARTIFICIAL INTELLIGENCE
2016, VOL. 30, NO. 8, 748–757
http://dx.doi.org/10.1080/08839514.2016.1229759

Taylor & Francis
Taylor & Francis Group

Responsibility and the Moral Phenomenology of Using Self-Driving Cars

Mark Coeckelbergh

Department of Philosophy, University of Vienna, Vienna, Austria

ABSTRACT

This paper explores how the phenomenology of using self-driving cars influences conditions for exercising and ascribing responsibility. First, a working account of responsibility is presented, which identifies two classic Aristotelian conditions for responsibility and adds a relational one, and which makes a distinction between responsibility *for* (what one does) and responsibility *to* (others). Then, this account is applied to a phenomenological analysis of what happens when we use a self-driving car and participate in traffic. It is argued that self-driving cars threaten the excercise and ascription of responsibility in several ways. These include the replacement of human agency by machine agency, but also the user's changing epistemic relation to the environment and others, which can be described in terms of (dis)engagement. It is concluded that the discussion about the ethics of self-driving cars and related problems of responsibility should be restricted neither to general responsibilities related to the use of self-driving cars and its objective risks, nor to questions regarding the behavior, intelligence, autonomy, and ethical "thinking" of the car in response to the objective features of the traffic situations (e.g. various scenarios). Rather, it should also reflect on the shifting experience of the user: how the new technology reshapes the subjectivity of the user and on the morel consequences this has.

Introduction

Self-driving cars can drive without human intervention. Since a few years, there is an acceleration in the development of such cars; many companies are experimenting with them, for example, not only Google but also major car companies such as Mercedes, Nissan, and Volvo. These developments promise more safety and less congestion on the roads, as it seems that less cars may be needed (for promises regarding self-driving cars, see for instance Thrun 2010). Yet, they also raise ethical issues. For instance, who is responsible when something goes wrong? The user of the car? The manufacturer? The car?

CONTACT Mark Coeckelbergh mark.coeckelbergh@univie.ac.at Department of Philosophy, University of Vienna, 1010 Vienna, Austria.

Recently, some work has been done on the ethics of self-driving cars (e.g., Goodall 2014; Lin 2013, 2015), including the question regarding responsibility (Hevelke and Nida-Rümelin 2015). Philosophers engaged in this discussion tend to focus on the responsibility of the user of the car (is she responsible for causing an accident if she did not drive?) and/or on the moral agency and moral reasoning of *the car* as artificial agent, which is expected to avoid causing harm to its passengers and other human traffic participants. For instance, one could try to make the car more "ethical," creating what Wallach and Allen would call a better "moral machine" (Wallach and Allen 2009). A typical scenario discussed is that in which a child runs into a street and the car cannot stop on time. There are also other crash avoidance problems, sometimes in the form of the so-called trolley problem, in which there is no win scenario and the car has to choose which person(s) to kill (e.g. Lin 2013). While this exercise may help designers to know more about how to make a safer autonomous system, and in addition may help philosophers to continue reflection discussion on *human* ethics, the scenarios created for the purpose of this exercise tend to take an external, third-person point of view—an overview of the road, the god's eye point of view of the duty ethicist or utilitarian ethicist, etc—and tend to neglect a first-person point of view, for instance, the point of view of the user of the self-driving car or other users of the road.

This paper brings in the latter, first-person perspectives, and discusses issues concerning responsibility and self-driving cars by (1) focusing on how the new technology changes the phenomenology of using a self-driving car (and of perceiving such a car, for instance by a pedestrian), indeed reshapes the subjectivity of the user and (other) traffic participant, and (2) exploring the moral implications of this changing experience, in particular, the implications for responsibility.

The structure of the paper is as follows: First, a working account of responsibility is presented, which identifies two classic Aristotelian conditions for responsibility and adds a relational one. A distinction is made between responsibility *for* and responsibility *to*. Then, this account is applied to a phenomenological analysis of what happens when we use a self-driving car and participate in traffic. It is argued that self-driving cars threaten the excercise and ascription of responsibility in several ways. These include the replacement of human agency by machine agency, but also the user's changing epistemic relation to the environment and others, which can be described in terms of (dis)engagement. The conclusion further reflects on what this means for the discussion about the ethics of self-driving cars and today's cars.

Responsibility: A brief working account of responsibility conditions, including epistemic and relational aspects

The concept of responsibility is a perennial issue in the history of philosophy; for the present analysis, I set up a brief working account that I will use as a

tool to explore ethical issues with self-driving cars. What I need for my analysis is an account of the conditions of responsibility: what is needed for the exercise and ascription of responsibility to work? Under what conditions is it possible to exercise and ascribe responsibility?

A classical account of this has already been provided by Aristotle (1984) in the *Nicomachean Ethics*. He argued that exercising responsibility—at least in the sense of *being responsible for what one is doing*—requires that at least the following conditions are met. First, the agent needs to be in control of what she is doing. If we lack control, we are not responsible. For instance, I lack control over the weather, therefore I am not responsible for it. By contrast, and to get closer to the issue at hand: I do have control over my car. Therefore, I am responsible for my driving. Second, the agent needs to know what she is doing. If I sleep, for instance, I am not responsible for what I do. If, on the other hand, I am wide awake and I see that my car is going in the direction of a pedestrian who will most likely be harmed or even killed if I do not change direction and/or hit the brakes, then I am responsible for whatever I do. More generally, knowing what you are doing does not only include knowing how to operate a machine, for instance, but also knowing the entire situation, that is, it includes knowing the environment and the situation in which the (here: technological) action takes place, for instance, the road and its environment and the traffic situation. This then enables the person to respond to the situation.

Usually when a question concerning responsibility is being discussed, all this is judged by ethicists and legal scholars from a third-person point of view: an overview, perhaps a god's eye point of view. What counts, it seems, is the "objective" situation, and the focus is on "the individual." However, responsibility can also be conceptualized in a different, more social and relational way. It can be understood as answerability (Duff 2007), for instance, and not only in a criminal context but in general: responsibility is always *relational*, it is responsibility *to* someone (Gardner 2003; see also Coeckelbergh 2010). Let me here use and further develop this conception of responsibility in a way that does justice to a more relational and phenomenological moral epistemology.

I already interpreted the epistemic condition in the classic account in a more relational way, since one could say that it is part of the epistemic condition that there is a relation to, and engagement with, the environment. However, the environment does not only include "things," "artefacts," "architecture," and "nature"; it also includes *others*. These others, I submit, are crucial in any account of ethics. The point of responsibility, one could say, is that we are responsible *to* others. Therefore, here is what at first seems like an alternative view. Instead of focusing on the agent, we might want to focus on the person *to whom* the agent is responsible. Here, the condition is that there is a relation—or, better: one or more (morally relevant) relations—between me and the person, and that I know this relations, that I am aware of this relations. The relation can

simply be that the other is human, or there can be additional and particular relations: perhaps the other is part of my family or of the same community, for instance. The particular relations will influence the particular responsibilities I have. For instance, I may not only have the responsibility not to harm someone, but also have the responsibility to provide care for that person because she is part of my family. Traffic participants have responsibilities in virtue of their relation to others as humans (and some nonhumans, perhaps even all nonhuman living beings—this depends on the ethics we want to endorse), which at the very least should not be harmed, and in virtue of their relations to others as traffic participants, which gives them particular responsibilities. For instance, as a driver, I am responsible for the safety of other people in my car and for the safety of pedestrians who may also participate in the traffic, and I will take (different) actions to ensure the safety of both kinds of others I am related to. In addition, in all these cases, I can only exercise that responsibility if there indeed *is* a (morally relevant) relation between me and that other and if I *know* that relation. For instance, if someone were to dress up in a way that disguises her humanity or nature as a living being and lie down on a dark and foggy road, then a driver who does not even perceive her cannot be expected to be responsible for hitting that person. It is impossible for the driver to exercise her responsibility, and we cannot ascribe responsibility to her. Similarly, if someone is perceived as a "monster" or a thing (e.g. "dirt"), there is no perceived relation to a human being or a (friendly) living being, so someone who sees the "monster" or the "piece of dirt" will not be able to exercise responsibility toward it. The perception of a morally relevant relation is lacking. Thus, for a relation to give rise to responsible behavior, the other and the relation to the other need to be perceived, the other needs to appear to me as other, in particular as a morally relevant other. This may be described as highlighting more "subjective" aspects as opposed to "objective" aspects, and more "social" aspects as opposed to "individual" aspects, although I prefer to move beyond such dualistic ways of thinking.

At the very least, we can say that both aspects should be part of any adequate account of responsibility. Indeed, that the latter view does not necessarily exclude the form; instead, I propose that we combine it with the classic one, which gives us a more comprehensive and a more complete picture of responsibility: We are responsible *for* what we do *to* others who might be affected by what we do. With regard to ascribing and exercising responsibility, it is important to see if all these conditions are met, that is, all conditions for both kinds of responsibility: the agent must have control over the action and know what she is doing, *and* the conditions should be such that there are others, or an other, that are perceived, appear as morally relevant, and as standing in relation to her. These conditions ensure that she not only should be but also *can* act morally responsibly. In traffic, for instance, it means that a car driver can only be responsible for her driving if

(1) she is in control of the car, that is, she is driving and not someone else, and she knows what she is doing with respect to operating the car and knowing the environment—this enables her to respond to it, and thus to drive "responsibly" in this sense. However, (2) she also needs to perceive, and experience, that she is related to others to whom she should act responsibly. She needs to experience these others as others, rather than things or elements of the environment (although she may also have responsibilities to that non-living environment). In order to respond to other traffic participants in a morally responsible way, these participants also have to appear *as* humans or morally relevant non-humans, and as traffic participants. Therefore, with regard to them, driving responsibly means responding to them: both in the traffic situation and potentially after something has happened and after she has done something, when she may be asked to give account of what happened and what she did. In addition, driving responsibly in this sense means to keep in mind this responsibility-*to* (and not only responsibility-*for*), to keep in mind that the question of responsibility, and the question of ethics, is not only about "who does it" and "who has done it" but at least also about who is, or might be, affected by that agent's actions, about what is being done, has been done, and could be done *to whom*. This future aspect is important. Driving a car and traffic, like all human activities and practices, involves risk. Something might happen because of my actions, because of *what I do*. And something might happen *to* others. An appropriate moral imagination, then (that is, a moral imagination that takes into account these different senses of responsibility), entails imagining what my actions could do to others. It is about acting (e.g., driving) in a way that takes into account the morally relevant features of the situation (e.g., the traffic situation), but this crucially includes others. However, this exercise of moral imagination and moral responsibility only works if the conditions are in place, if I am in control and know what I am doing, and also see others as others to whom I should act responsibly.

Now we have a working account of responsibility, which I have combined with a relational ethics and with what we may call a moral phenomenology of traffic (and more generally: action), which starts from a first-person point of view and emphasizes that when it comes to the conditions for the exercise and ascription of moral responsibility, these conditions crucially depend on the perception and appearance of the situation, environment, and others. Whatever may be said from a third-person point of view (or even god's eye point of view, if this is possible at all), the concern about responsibility—with regard to traffic and other human actions and practices—should not only involve the exclamation that people *should* act but also should inquire into the conditions that make possible responsible action and responsible practice. It should also ask the question *how* people *can* act responsibly. In addition, to articulate these conditions and understand what they require of us, a

phenomenological-hermeneutic approach helps since it reminds us that our epistemic relation to the environment and others is not so straightforward: whatever reality may be, real (!) humans always have access to reality in a mediated way. If the exercise and ascription of responsibility are only possible when we (1) know what we are doing and (2) know what we are doing to others, then it is important to study what we *can* know. Maybe what we are doing is not always clear to us. And there is no guarantee that others appear to us as others. For instance, driving is a skill, and skills involve tacit knowledge. We know what we are doing, but maybe we cannot explain it to ourselves and others. Is this a problem for responsibility? And as I already suggested, others do not always appear as others. Maybe the technology of driving makes it more difficult to recognize others as others, for instance if we drive at a high speed or—as I will argue below—when these others are hidden in a car. The technology mediates our relation to our environment and others.

A term that also can be added to the analysis is engagement: the conditions for responsibility seem to require that we engage with our environment and others, with the human and non-human environment. Ethics, it seems, is about such engagement. A failure to act responsibly is a failure to engage. However, this engagement should be possible. There may be conditions under which it becomes more difficult to engage. This is especially clear in the case of the self-driving car, to which I shall now turn.

A moral phenomenology of self-driving cars

Even the use of non-automated cars raises problems concerning the conditions for responsibility as identified previously. For instance, one may ask if someone knows what she is doing if she does not only have her biological body at her disposal but a machine, power of which far exceeds the powers of humans and (other) animals. Our epistemic capacities may be adapted to a time when our body was not yet extended with this kind of technology. How can we have morally adequate awareness of our environment, for instance, at a speed of 100 miles/hour? And, are we really able to properly and safely navigate contemporary complex urban environments? Moreover, as I noted driving is a skill and we may learn well how to operate the machine, but when suddenly the traffic situation requires a response from us, do we really know what we are doing (or did) at that moment, let alone what we *should* do, given that our know-how is implicit and that there is no time for moral reasoning? Of course, we may know *in general* that there are risks associated with using a car (Hevelke and Nida-Rümelin 2015, 628), and this may give us *some* responsibility, but that does not mean that, when we are driving and experiencing our driving, we know the specific traffic situation and its morally relevant features. We might know about "objective" risks associated with using cars. However, we do not necessarily know specific risks *as we are*

driving. The Aristotelian account is in trouble, and only moral philosophers thinking about self-driving cars may have time for moral reasoning; drivers lack that time.

Furthermore, it is doubtful if others always appear as others in traffic. Car traffic, in particular, seems to raise the problem that when it comes to relations with other car drivers, others no longer appear as others in so far as they are literally shielded off from view. Other drivers disappear and what remains is a car, which is perceived as having agency on its own. This phenomenology has moral consequences: it becomes at least more difficult to feel responsible *to* other drivers since they hardly appear as others. It is possible to feel responsible to another human being, and perhaps to an animal. Luckily because of animism, cars also have an "animal" face, they appear as animals or even persons (think about "cute" or "aggressive" cars). This may give some reason for hope. However, so far, as the car appears as a machine, it is impossible to feel responsible-to. (Unless the machine were to appear as a quasi-other; I will return to this point.)

In the case of self-driving cars, these problems get worse. First, if we try to apply the Aristotelian conditions, the machine may be in control but it is doubtful if it really "knows" what it is doing, since this seems to suppose consciousness. However, even if this problem would be debunked, it is clear that while the user may be aware that using a self-driving car is dangerous *in general*, clearly *when driving* and in specific *situations*, the human the user of the machine is (1) not in control of the machine and (2) does not really know what the machine is doing, let alone that one has knowledge of the environment and the situation and its morally relevant features (this was my environmental, more relational interpretation of the condition). Phenomenologically speaking, the user stops being a driver and becomes a passenger, and passengers are not responsible for what the driver does. The driver is responsible, but the driver has been replaced by a machine. The passenger may know certain statistics or other "objective" risks associated with using the car, but lacks knowledge *as a driver*. She has no driving skills, has lost driving skills. In addition, the passenger does not usually, and certainly not necessarily, engage with her environment, let alone that she might perceive the morally relevant features of it. This means that, assuming that all agency is entirely transferred to the machine, as users of the self-driving car, we can neither exercise responsibility nor can moral responsibility be ascribed to us. The Aristotelian conditions for responsibility are not met.

Moreover, self-driving cars fare even worse when it comes to the conditions for relational responsibility. The problem is not only that a machine, having no consciousness, cannot feel responsibility-to, cannot really recognize a morally relevant relation and cannot recognize others *as others*, but also that humans will perceive the car and its actions as "machine" actions, that is, they will not at all recognize that car and its machine driver as

"other." This means that, if in the case of contemporary cars, they already feel less responsibility-to, and in the case of self-driving cars, the condition for relational responsibility is entirely lacking. Unless the car is perceived as other, human drivers who encounter the machine-car will be unable to relate to it in a morally relevant way, and social-relational autonomy cannot get off the ground.

However, one could argue that there are at least two reasons why this sketch of the situation may be too pessimistic. Consider the following qualifications I made in the previous argument.

First, I had to add the condition that agency is fully transferred to the car. If this is not the case, then human beings might be able to do some of the moral-epistemic work required for exercising responsibility. However, this may not be that easy: if the car is semiautomated, humans may lose some of the skills and/or pay less attention than they did before, and less engagement endangers the conditions for responsibility: we may fail to see the morally relevant features of the situation and the environment, and we may fail to see others as others, at times when we are the passenger but also when we are the driver, since we are no longer used to driving all the time. Therefore, some of the ground for exercising and ascribing responsibility may be saved; yet much might already be lost.

Second, humans may perceive the car as a *quasi-other*. This may happen to the extent that we (still) have animistic and anthropomorphic tendencies: if we perceive the car as a living being or human, then the car may appear as morally relevant and as other. This can then encourage more moral engagement and responsibility-to in other (human) drivers than one would expect. Humans may feel responsible to the machine, even if it "really" is a "machine" and not an other. Note, however, that this exercise of engagement and responsibility is not mutual; the non-human driver cannot really recognize the human driver as other, and will at best be programmed and learn to act *as if* it treats humans as others (if even this can be programmed at all or learned by a machine).

In addition, even if the machine is seen as a quasi-other, it will be difficult for any human user of the road to assess the "intention" of the machine. This may already be difficult in the case of encountering humans, sometimes, but since humans and machines may "think" in different ways, it is unclear whether humans can properly know the new situations that will emerge due to different, non-human machine behavior and reasoning not transparent to the human user. The "intention" of the machine will then be difficult to determine. In this sense, when driving, they will no longer fully know what they are doing (and might do) because they will not fully know what the machines are doing (and might do). One could reply that humans will adapt to the machine "minds" and new situations; but whether that will happen is very uncertain.

To conclude, it seems that the self-driving machine can fulfill one condition for responsibility (it can have control over the action, it can assume agency), but not the other, epistemic and social-relational ones. In addition, the humans interacting with the machines will have difficulties to know the new behaviors, situations, and the new environment. They may have difficulties to know the "intention" of the car. And they will have difficulties to exercise responsibility-*to* if they encounter machines—unless, perhaps, if these machines would be perceived as quasi-others.

Conclusion

In summary, discussions concerning the ethics of self-driving cars should be restricted neither to general responsibilities related to the use of self-driving cars and its so-called objective risks, nor to questions regarding the behavior, intelligence, autonomy, and (ethical?) "thinking" of the car in response to the objective features of abstract traffic situations (e.g., various scenarios). Instead, these discussions should also reflect on the shifting experiences of the user: how the new technology reshapes the subjectivity of the user and on the moral consequences this has. In this paper, I have highlighted the epistemic and social-relational problems that render it more difficult to fulfill the conditions for responsibility *for* and responsibility *to* in the case of self-driving cars.

This discussion has also suggested a different way of understanding responsibilities in non-automated traffic. Already today, the machines we use for driving and the environments we use them in, render it increasingly difficult to fulfill epistemic and social-relational conditions for exercising and ascribing responsibility. For instance, already today, we can ask the question if we experience other drivers as others, or if, *when driving*, we experience either machines or quasi-others. Here too, a broader and rational account of responsibility and a phenomenological, first-person point of view approach can help to bring out these problems and, hopefully, find a solution to them.

More research is needed to explore what, exactly, happens to human experience when, if these and similar developments continue, we are confronted with self-driving cars and other automated machines. Important for these discussions and indeed for the design of these systems (if they have to be developed at all) is to realize that we will confront them as humans, that is, as beings with specific epistemic and social possibilities and experiential peculiarities. If we want to encourage responsible use of machines, we should take this into account—in this discussion and in similar discussions in robotics and artificial intelligence.

Finally, paying attention to the moral experience of users of automated machines and its relational aspects also opens the discussion up to discussing the *cultural* dimensions of automation. The perception, experience, and use

of automation may be different in different cultural contexts. This is likely to have consequences for the conditions for exercising and ascribing responsibility. Moreover, the very concept of responsibility is usually defined in a culturally independent way; but various cultures may have slightly different ways of understanding responsibility. A more relational or "mixed" non-relational/relational account as suggested here may have a better chance of doing justice to such different ways of understanding responsibility. Indeed, perhaps the very facts that attention to responsibility-to is often missing in mainstream modern Western thinking about responsibility and that the ethics and development of technologies are discussed without taking into account cultural differences, should themselves be problematized.

References

Aristotle. 1984. Nicomachean ethics. In *The complete works of Aristotle*, ed. J. Barnes, Vol. II, 1729–867. Princeton, New Jersey, US: Princeton University Press.

Coeckelbergh, M. 2010. Criminals or patients? Towards a tragic conception of moral and legal responsibility. *Criminal Law and Philosophy* 4:233–44. doi:10.1007/s11572-010-9093-6.

Duff, R. A. 2007. *Answerability for crime: Responsibility and liability in the criminal law*. Oxford, UK: Hart Publishing.

Gardner, J. 2003. The mark of responsibility. *Oxford Journal of Legal Studies* 23 (2):157–71. doi:10.1093/ojls/23.2.157.

Goodall, N. J. 2014. Machine ethics and automated vehicles. In *Road vehicle automation*, eds. G. Meyer, and S. Beiker, 93–102. Cham, Switzerland: Springer.

Hevelke, A. and J. Nida-Rumelin. 2015. Responsibility for Crashes of Autonomous Vehiles: *An Ethical Analysis*. *Science and Engineering Ethics* 21:619–630.

Lin, P. 2013. The ethics of saving lives with autonomous cars are far murkier than you think. Wired. http://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars (accessed October 9, 2015).

Lin, P. 2015. Why ethics matters for autonomous cars. In *Autonomes Fahren*, eds. M. Maurer, et al., 69–85. Berlin, Germany: Springer.

Thrun, S. 2010. Rethinking the automobile. (TEDx talk). https://www.youtube.com/watch?v=r_T-X4N7hVQ (accessed January 18, 2016).

Wallach, W., and C. Allen. 2009. *Moral machines: Teaching robots right from wrong*. New York, New York, US: Oxford University Press.