Frontiers in Artificial Intelligence and Applications

SOCIABLE ROBOTS AND THE FUTURE OF SOCIAL RELATIONS

Proceedings of Robo-Philosophy 2014

Edited by
Johanna Seibt
Raul Hakli
Marco Nørskov

IOS Press

Sociable Robots and the Future of Social Relations
J. Seibt et al. (Eds.)
IOS Press, 2014

doi:10.3233/978-1-61499-480-0-265

Responsibility, Robots, and Humans: A Preliminary Reflection on the Phenomenology of Self-Driving Cars

Mark COECKELBERGH[1]
Centre for Computing and Social Responsibility
De Montfort University

Abstract. In this paper I discuss this question regarding responsibility for self-driving by reflecting on the phenomenology of car driving in general and the phenomenology of self-driving cars in particular. After setting up a relational view of responsibility and an approach which emphasizes the experience of responsibility, I point to what I take to be de-socializing effects of modern car driving, explore what it would mean to add self-driving cars to this picture, and draw implications for responsibility.

Keywords. responsibility, phenomenology, driverless cars, Levinas, motivation, moral technology

Introduction

As new autonomous technologies enter daily life, we have to face ethical questions and questions regarding responsibility. Consider for instance self-driving cars such as the one developed by Google. Today cars are already rather "robotic" considering the amount of built-in ICT, but self-driving systems effectively make it *autonomous* robots. This raises a number of ethical issues, such as problems with safety (the driver and others) and the security of the data (the car may be hacked for malicious purposes). But its autonomous character especially raises questions regarding *responsibility*. Is the human still responsible if (s)he is no longer "driver" but "operator"? Can we still take responsibility for driving at all in this situation? What is needed for us to exercise this responsibility? What is needed for responsible driving?

In this paper I discuss this question regarding responsibility for self-driving by reflecting on the phenomenology of car driving in general and the phenomenology of self-driving cars in particular. After setting up a relational view of responsibility and an approach which emphasizes the experience of responsibility, I point to what I take to be de-socializing effects of modern car driving, explore what it would mean to add self-driving cars to this picture, and draw implications for responsibility.

My reflection on this topic is preliminary since a more systematic and comprehensive framework is needed to fully describe and understand the phenomenology of driverless cars (and robotics and automation technology in general) and to deal with the related responsibility issues.

[1] Centre for Computing and Social Responsibility, De Montfort University, The Gateway, Leicester, LE1 9BH, UK; E-mail: mark.coeckelbergh@dmu.ac.uk.

1. Relational Responsibility and the Experience of Responsibility

The usual approach to this and similar responsibility problems, is to ask who or what is responsible (the robotic car or the human), which leads to the question if the robotic car is a moral agent. If it is a moral agent, the car can be held responsible. If it is not a moral agent, what exactly is its status? Is it still an agent, and if so, can it be moral to some degree? What would be necessary to make it into a "moral machine" [1]? To answer these questions one then needs to discuss which properties the entity has.

In *Growing Moral Relations* [2] I have questioned this properties-based approach and this way of asking the question regarding moral status. Instead I have defended a relational view of moral status which is "relational" in two senses: it focuses on relations between entities and which therefore supports a relational ontology/metaphysics of humans and non-humans (e.g. looks at how humans relate to robots rather than only what robots "are"), but it is also "relational" with regard to the subject-object relation. I have proposed a relational epistemology according to which what matters is our experience of entities, the way they appear to us, how we as experiencing subjects actively relate to them. For robots, this means that there is no robot-in-itself, the specific robot already appears to us in a specific way in a specific situation, relation, and context (practice, society, culture). I will return to this point below.

For responsibility, this twofold relational approach has the following implications: First, a relational view of responsibility means that responsibility is always about others, that we are in relation to (human and non-human) others and that therefore we are responsible *to* those others. Responsibility is always social responsibility. I guess most of us know this or could accept this point, albeit usually in philosophical discussions of responsibility the focus is on properties of the "agent" rather than on the other. (Note that a Levinasian ethics more radically chooses this other-oriented approach to responsibility; however, I will not discuss this here.)

For the problem of robotic cars, this means that the question regarding responsibility is also and especially about how to be responsible *to* others – other drivers or operators – in traffic, rather than about the properties and agency of the robotic car.

Second, however, I wish to make a deeper, epistemological point about the relationality of responsibility, which mirrors the epistemological point I made about moral status: knowledge and experience of responsibility makes sense only within such a relation and my experience of responsibility *depends on how the other appears to me*. This relationality concerns again the subject-object relation and is a more controversial point in philosophy, even after so many years of Kant and phenomenology. It shifts the emphasis from abstract reasoning about morality and responsibility to concrete situations and questions concerning the experience of responsibility in these situations and contexts.

For the problem of robotic cars, this means that we need to attend to how drivers/operators experience responsibility in these new driving and traffic situations, rather than to so-called "objective" properties of the robotic cars.

In addition, this approach shifts the emphasis from justification to motivation: the main question is no longer "What is the right thing to do, and can robots think about what is the right thing to do?" but rather: "How can humans as social beings be motivated to do the right thing, and how do machines shape that motivation?"

Let me now further apply these two approaches – what we may call a "remote" abstract one and a "close" relational one – to the responsibility problem of the self-driving car.

2. The Phenomenology of Self-Driving Cars

As said, the standard approach to responsibility problems with autonomous robots is to focus on the status of the machine. The "remote" approach is to ask about the status of for example the Google car: does it have moral agency? Can it make moral decisions in difficult traffic situations? Can we build-in moral rules to *make* it more moral? The moral status and the capacity for responsibility are assessed from the outside. And if neither the car nor the human can be held responsible, can we solve the problem by means of law and insurance? For instance, today car insurance already partly de-moralizes accidents. Responsibility questions are replaced by remote procedures and payment.

A "closer", more relational approach to responsibility, by contrast, focuses on the experience of the drivers, including the appearance of the cars and their drivers/operators. Let me explore the phenomenology of car driving in general and driverless cars in particular.

First, even today non-autonomous cars are already de-facing (to use a Levinasian term after all) in so far as they hide the face of the drivers. Modern cars render their drivers anonymous and – literally - faceless. This has implications for responsibility. I don't know you therefore I do not feel responsible. I only see a car with a "driver", if at all. I do not experience *to whom* I am responsible. Responsibility remains abstract; something the law tells me. Therefore, it is not responsibility at all, in the rich social, relational sense articulated above. Contemporary cars and traffic are effectively de-socialized, de-relationalized, and hence de-moralized. Traffic is not experienced as a "social" setting and participants in traffic are not experienced as part of a community, therefore difficult to take up responsibility. Social relations are hidden; this encourages irresponsible behaviour. (Note that empirical research in social psychology seems to support this analysis when it suggests that anonymity stimulates aggressive driving [3].)

Second, this loss of responsibility experience and responsibility motivation seems to be even higher in the case of self-driving cars, which become entirely "driverless" cars: as a participant in this kind of traffic I do not feel any responsibility whatsoever, because what I encounter on the street is no longer a human being but a robot (which transports a human being, but that is not very visible), an autonomous machine. The automation of the car thus contributes to the already on-going de-socialization and de-moralization of traffic and car driving.

This re-description of the problem thus illustrates how shifting the focus from abstract arguments about moral agency and justification of (ethical) behaviour to the practical study of human moral subjectivity, of human experience with technology within practices, gives us rather different questions and different answers. The approach asks how responsibility is experienced by practitioners and what motivates people to act responsibly within a practice, given the use of particular artefacts, particular habits, rituals, etc. Indeed, the approach reveals not the properties of an object but a practice and an entire *culture*. How a particular artefact is viewed is not something that can or should be "objectively" defined, but something that should be

understood *within* the practice and *within* the society and culture. For instance, here I describe (the experience of) the self-driving car as part of a practice which is all about anonymity and about things, moving objects – as part of a culture which is focused on moving things rather than people. Traffic is part of anonymous modern urban life and is experienced as a-social.

Note that due to the appearance of contemporary cars, traffic may even be described as *anti*-social. It does not take many examples to illustrate that today many cars look *aggressive* and often this is actively promoted by advertisements.

Again this is a matter of phenomenology: it is a matter of how we experience these features. It is something that happens in the subject-object relation, and thus given appearance and "behaviour" of the car in a situation and context. Perhaps the car "is" not aggressive, but we may feel it is when we look at it and when we see "its" behaviour. Perhaps the car "is" not an animal (according to a scientific point of view), but we may *experience* it as such. We may feel the *car* has a "face" – e.g. a cute face or an angry face. Perhaps it also has a "body". From the point of view of the driver, the car may become a kind of exoskeleton or (other) extension of the human body. Maybe the car becomes part of the body schema of the driver. The exterior of the car is then the "skin" of the driver. From the point of view of the spectator (including the driver-spectator) the car may equally appear as the body of the other driver. But it may also appear as an animal or a machine. Cars may have different "personalities". Cartoons for children which animate cars and make them look like humans with a personality do not have to do much "work"; we already animate the car *before* it even enters the world of fantasy.

Indeed our experience of technology is often shaped by animism and anthropomorphism. We attribute life to what is "objectively" a "mere" machine. We interpret artefacts in human terms. This also happens to our experience of cars.

For evaluating the moral implications of self-driving cars, this means we should see the discussion of the status and properties of these cars as representing only *one* possible experience – one heavily influenced by objectivist science; there are many more possible experiences and interpretations. There are also more ways of driving and different traffic cultures, some of which may well be less de-socialized than ours. It is important not to disconnect the discussion about self-driving cars and responsibility from broader questions about driving, traffic, and culture.

3. Conclusion

In this paper I have argued for an alternative approach to responsibility problems raised by self-driving (or driverless) cars. I re-framed the responsibility problem in relational, social terms, and connected this problem to motivation (rather than justification) and to appearance and experience (rather than "objective" features and status). The latter way of framing the problem is not only better in terms of ontology and epistemology, but may well be better in solving practical responsibility problems since it is prevention-oriented rather than post-accident (who's to blame?): if it is indeed the case that current traffic and current cars encourage a-social or anti-social behaviour, then the challenge is to recreate traffic as a practice that enables and encourage responsible behaviour. Re-designing cars may be part of that project. But whereas the standard approach would work on prevention by making the car more "moral" in the sense of it having (better) moral capacities, the alternative approach I propose faces the challenge to create cars

that contribute to a re-socialization and re-moralization of traffic practices, traffic spaces, and traffic cultures. Automation and artificial intelligence may or may not be part of that, and if they are then we have to make sure the technologies contribute to re-socializing traffic.

This is a strong normative position. Indeed, taking into account the phenomenology of practices does not lead to "anything goes" moral relativism, but instead enables us to ask new questions which are linked to a rather clear normative direction that emerged *in* and not *outside* the phenomenological analysis – here it was already present in my analysis of the driver experience – and which is open to discussion and revision (one can present a different phenomenological description and interpretation, and voice a different normative concern). The new questions I ask here are the following: How can we re-socialize and perhaps re-face traffic? What should be the appearance of future cars, given that current appearances tend to a-socialize if not anti-socialize?

For robotics and *robophilosophy*, answering these questions means that we should not focus on the "mind" of the robot (e.g. the self-driving car) but on how humans experience the robot and how the new machine may change the practice (e.g. the practice of driving). It also suggests that we make an interdisciplinary effort that includes for example studies of animistic experiences and learns from (other) anthropological work, and that pays attention to the relation between technology and world views and religion (e.g. animism and nature religion again, but also post-Christian culture). Modern concepts, for instance, may not be sufficient to fully understand how humans experience and use robotic technologies, and conceptual work may not be enough. Philosophers, social scientists, and roboticists must work together to modify and create new words and new things in a way that supports rather than undermines social responsibility.

References

[1] W. Wallach & C. Allen, Moral machines: Teaching robots right from wrong, Oxford University Press, Oxford/New York, 2008.
[2] M. Coeckelbergh, Growing Moral Relations: Critique of Moral Status Ascription, Palgrave Macmillan, Basingstoke/New York, 2012.
[3] P. Ellison-Potter, P. Bell, & J. Deffenbacher, The effects of trait driving anger, anonymity, and aggressive stimuli on aggressive driving behavior. Journal of Applied Social Psychology 31(2), 431-443, 2006.