Ethics in the design of automated vehicles: the AVEthics project

Ebru DOGAN[a,1], Raja CHATILA[b], Stéphane CHAUVIER[c], Katherine EVANS[ac],
Petria HADJIXENOPHONTOS[ab], and Jérôme PERRIN[d]

[a] *Institut VEDECOM, Versailles, France*
[b] *Sorbonne Universités, Université Pierre et Marie Curie,CNRS, Institute for Intelligent Systems and Robotics (ISIR)*
[c] *Sorbonne Universités, Université Paris-Sorbonne, Faculty of Philosophy*
[d] *Renault, Technocentre, Guyancourt, France*

Abstract. Automated vehicle (AV) as a social agent in a dynamic traffic environment mixed with other road users, will encounter risk situations compelling it to make decisions in complex dilemmas. This paper presents the AVEthics (Ethics policy for Automated Vehicles) project. AVEthics aims to provide a framework for an ethics policy for the artificial intelligence of an AV in order to regulate its interactions with other road users. First, we will specify the kind of (artificial) ethics that can be applied to AV, including its moral principles, values and weighing rules with respect to human ethics and ontology. Second, we will implement this artificial ethics by means of a serious game in order to test interactions in dilemma situations. Third, we will evaluate the acceptability of the ethics principles proposed for an AV applied to simulated use cases. The outcomes of the project are expected to improve the operational safety design of an AV and render it acceptable for the end-user.

Keywords. Ethics, artificial intelligence, robotics, automated vehicle, artificial moral agents

1. Introduction

Technological developments in sensors and wireless communication facilitate the development of sophisticated advanced driving assistance systems. Several subtasks of the driving task, such as lateral control and longitudinal control are now handled by the vehicle. The human driver is left more and more out of the control loop of the vehicle as the level of automation increases and the vehicle becomes an autonomous agent. A deployment of a fully automated vehicle (AV) in all contexts, however, is expected to take a few decades. Then an AV would become a social agent taking decisions to regulate its interactions with other road users and static objects in a mixed traffic environment. Some situations would involve complex decision making when life hazard is involved. Currently, an AV does not have a consensual minimal risk state, nor a crash optimization strategy. In fact, the decision-making architecture consists of a set of rules, mostly the Highway Code, applied by a programmer. Given the difficulty of predicting the behavior of dynamic objects in the traffic environment, there would be no way to completely avoid

[1] Corresponding Author.

a collision ("inevitable collision state") and the aim would be to minimize risks and damages [8]. Since the risk cannot be avoided, the decision turns into an ethical one: there will not be one "good" solution and the decision will involve a trade-off between interests of different parties in a given context [12]. An AV does not have a decisional autonomy, that is, "the capacity of reasoning on the perception and action in order to make non-trivial choices" [3, p.15]. Nor does it have a sense of ethics. Nonetheless, it would have to make real time decisions of risk distribution in ethical dilemmas involving high uncertainty. Although this issue relates to the general domain of "robot ethics" [1], the case of AV has specific features: i) open environment (public roads), ii) interaction with many different social agents, iii) involvement of many stakeholders of the road mobility system (car makers, insurance companies, public authorities, etc), and iv) entrust of the safety of the car occupants for the robot.

2. AVEthics Project

Figure 1 depicts a dilemma situation on a road that anyone can encounter. In such complex dynamic situations human drivers report that, even though they can explain their decision making processes once the situation is over, they do not reflect on the same terms while the situation is taking place [11]. Thus, human reaction in a dilemma situation is not pre-calculated; it is a split-second reaction [12], whereas an AV's will have prescribed decision algorithms to control the vehicle. In the situation depicted in Figure 1, no matter how the AV decides to manage a conflict, someone might be harmed.

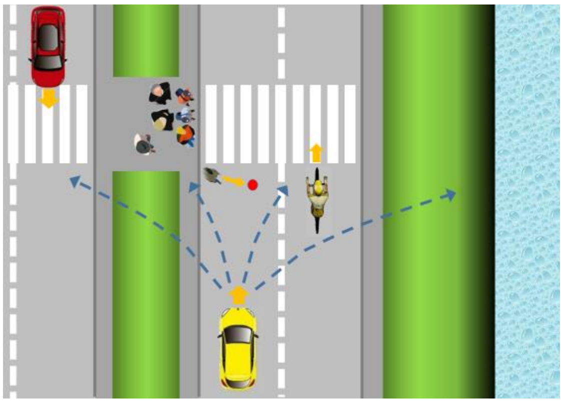

Figure 1. Sample use case

The current paper aims to present the AVEthics project (Ethics policy for Automated Vehicles). Its overarching goal is to provide a framework for an ethics policy for the artificial intelligence of an AV in order to regulate its interactions with other road users. This issue will be treated in three parts. First, we will specify the kind of (artificial) ethics that can be applied to AV, including its moral principles, values and weighing rules with respect to human ethics and ontology. Second, we will implement this artificial ethics numerically by means of a serious game in order to test interactions in dilemma situations. Third, we will evaluate the acceptability of the ethics principals proposed for an AV applied to simulated use cases.

3. Philosophy: from theory to casuistry

A common approach in robot ethics is to transfer *our* own way of reasoning on moral issues to artificial agents, creating a blueprint of our moral reasoning in robots, according to deontological or consequentialist theories (e.g. Kantian ethics, utilitarianism or virtue ethics). One of the problems of this approach is the arbitrariness of the choice of an ethical theory: why should we prefer a Kantian AV to a utilitarian AV? A more important problem is the lack of real operationalization of the capacities that enables humans to think morally.

The AVEthics project approaches the AV as a "*modular artificial moral agent" (MAMA,* [4, in press]*)*. Accordingly, an AV pursues its goals based on its artificial intelligence, and it is modular in the sense that it is not universal, but rather specialized to cover a limited set of goals. The artificial ethics endowed to a MAMA refers literally to the code that should guarantee that the MAMA's behavior should be sensitive to the rights, interests, and needs of all the entities that could be affected by its decisions. The challenge is the lack of consensus on *which capacities* to implement in the MAMA for it to successfully make ethical decisions.

One way to tackle this issue is to focus on the essential needs of artificial ethics, rather than trying to implement human morality into the robot. Human drivers' decisions are determined by prioritization of goals, and valences[2] of different action possibilities afforded by their perceptual environment, which are rarely calculated in advance [7]. Following this notion, the morality of a MAMA would mainly require sensitivity to a limited set of values and principles that would be morally salient in specific encountered situations depending on its functionalities (case-based approach), rather than general principles applicable to a great variety of situations (principle-based approach). However, the literature on the casuistic approach to ethical decisions (in robotics) is relatively limited [see 13, 9, and 1 for examples]. Moreover, it seems difficult to dissociate cases from principles. *How can an AV decide to crash into a dog instead of a cyclist, if the AV does not know the rule that human life has a higher value than the life of a dog?*

One way to favor a case-based over a principle-based approach is to dissociate deeds and valences. Human morality is rooted in the calibration of the content of the world that we experience [14], and our morality is shaped by situations and experiences we are confronted with. Hence, perception of the environment, valence entailed by the entities in the environment, and goal-directed behavior incorporating the notion of valence become common in human morality and artificial ethics.

In the philosophy part of the AVEthics project, we will argue that 1) an artificial ethics requires representation and categorization of the morally relevant entities in order to define its "ontology", which is a moral issue *per se*, 2) an awareness of different entities in the traffic environment could be implemented by assigning to each a numerical value that would be taken into account by the AV control algorithms, and 3) a special "self" value would be added, for an AV carrying humans may not share an ethics of self-sacrifice.

[2] Gibson and Crook (1938) use **"valence"** akin to hazard weight of a potential action in a given traffic situation.

4. Robotics: development and experimentation

An AV has limited decision-making capacity because its situation awareness is focused on the actual perceived situation. An AV is equipped with sensors, such as radars, lasers, and cameras, which provide it with 360°-vision. Perception, planning, and control algorithms of the vehicle enable it to move in a complex environment. Nonetheless, the information available to the vehicle by its sensors, how the vehicle interprets this information, and the way it uses the information for decision-making are all completely different from those of a human driver. Furthermore, an AV cannot cope with situations that were not anticipated and taken into account by the programmer (even if it includes a learning approach). Overall, an AV's decision-making is imperfect and uncertain.

The decision-making architecture of an autonomous system consists of three levels with decreasing decisional autonomy [6]: the higher "strategic" level manages goals, beliefs, and plans; the intermediate "tactical" level supervises the system; the lower "operational" level carries out the actions, e.g. longitudinal and lateral control for an AV. One of the challenges is the traceability of the decisions of an artificial intelligence agent. Given the possibility of liability and insurance problems, AV stakeholders would be keen on traceability. Implementation of the decision is another challenge: decisions based on the AV "ethical" principles should be representable in the control algorithms of the vehicle as tangible components such as speed, brake, time headway (time between the ego vehicle and a lead vehicle), and steering wheel movements. Hence, we need to test the feasibility of the ethical decisions of an AV.

In the previous section, we advocated categorization of the perceived entities and assignment of valences to these entities in an AV's ethical decision-making. For this end, the AV should be, first, able to quantify the reliability of the information acquired by its sensors. Only then can it rely on this information in order to distinguish among the entities in its environment and to categorize them. The categorization will determine the valence assignment and the action plan of the AV depending on the ethical theory being tested. Assuming that the sensor data is of good quality and reliable, this process has two sources of uncertainty. *The first uncertainty is related to the categorization* of entities, which carries a probabilistic confidence value. *The second uncertainty is related to the action implementation*: the course of the action planed by the AV based on an ethical decision is also probabilistic. Hence, decision-making should account for the uncertainties in categorization and action. *How can we handle these two uncertainties?*

In the robotics part of the AVEthics project, we will 1) test different approaches, such as fuzzy, belief-based or Bayesian, in order to tackle uncertain categorization, 2) investigate the best course of action for an AV, considering uncertain perception and non-deterministic actions, and 3) study optimal decisions that could be taken jointly by the AV and other agents in its surroundings (vehicles, pedestrians, and infrastructure). We will also develop a test tool, a serious game interface that can be connected to a driving simulator, so that we can apply the model of artificial ethics to the use cases and test this with human drivers.

5. Psychology: public acceptability

To assume that an AV would be acceptable because it would increase safety is not necessarily valid. Human ethical decision making is often seen as a mix of emotions and reason. The end-user might consider the overall collective safety gain to be insufficient to merit taking certain individual risks, even if they would be rare cases. Indeed, research in social psychology indicates that socio-cognitive constructs such as values, contextual features, and trust have a notable effect on acceptability [15].

People who do not have sufficient knowledge on complex, new, and most of the time, controversial technologies, such as AVs, rely on their trust in the main stakeholders [16]. Competence-based trust (i.e. trust in a stakeholder's experience and expertise) is rather straightforward: positive information about a stakeholder's expertise is associated with higher trust and acceptability (and vice versa). Integrity-based trust (i.e. trust in a stakeholder's honesty, openness, and concern), on the other hand, is more complicated: when people perceive a stakeholder as biased and dishonest, they *go counter to the organizational position*. More precisely, if the organization is a proponent of a new technology, people are negative about the same technology [18]. In fact, when the issue is of high moral importance for the individual[3], the objective information about the competence loses its persuasive power [5]. One can even denounce the legitimacy of the decisions of a stakeholder when morality becomes salient [17]. *The relationship between trust and acceptability is thus sensitive to the moral importance of the issue*.

Another concept related to trust and acceptability is values. People are more likely to trust involved parties if they share values similar to their own [16]. Information on value similarity also influences integrity-based trust, but not competence-based trust [5]. Two main clusters of values have been identified: self-transcendence values, which are concerned with collective outcomes, and self-enhancement values, which are concerned with individual interest. These two may be in conflict in controversial issues. For instance, scenarios which involve taking risks with people on-board in an AV would be in line with societal outcomes, but contradictory to individual outcomes. Perlaviciute & Steg (2014) propose that people's tendency to adopt deontological or consequentialist reasoning might depend on people's values.

What are people's preferences in situations of ethical dilemma situations? Recent survey research revealed that drivers had positive evaluations about a utilitarian AV that is programmed to minimize the casualty in unavoidable accidents, including the self-sacrifice scenarios [2]. Utilitarian thinking is observed in public policy evaluations as well. People's ethical preferences for road safety policies changed as a function of the value of the age and the responsibility/vulnerability of the victim: protection of young (*vs* elderly) road users and pedestrians (*vs* drivers) is favored [10]. However, findings in the neuroscience of moral decision making hint at the complexity of this process.

In the psychology part of the AVEthics project, we will 1) test a model of acceptability integrating people's trust in the competence and integrity of the stakeholders and the value similarity with the stakeholders, and 2) investigate public acceptability of different ethical principles for an AV decision making by using the game interface mentioned above, as well as end user surveys. We will also collect stakeholders' acceptability judgments.

6. Conclusion

The ethics of automated vehicles is becoming a major issue from legal, social, and vehicle control perspectives. We acknowledge that the AV will have to make decisions that might eventually harm an agent and that these decisions should not contradict the

[3] We presume that a harmful decision of an AV is of high moral importance for the end user of this technology

interests of the end users or the principal stakeholders. An ethics policy for automated vehicles is a vast subject, and AVEthics is only the beginning of a long path. The expected outcomes of AVEthics are i) an initial proposition of ethical principles for an AV, ii) a software system and interface to apply these principles to different use cases, and iii) end user's acceptability judgments of the proposed ethical principles and the following action plans. This should contribute to improvement of the operational safety design of an AV and render it acceptable for end-users and stakeholders of the mobility system.

References

[1] Arkin, R.C. *Governing lethal behavior in autonomous robots*. Chapman & Hall, NY, 2009

[2] Bonnefon, J.F., Shariff, A., & Rahwan, I.. *Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars?* http://arxiv.org/pdf/1510.03346v1.pdf. Accessed on 28/10/2015.

[3] CERNA.*"Ethique de la recherche en robotique"*, rapport n°1, Novembre 2014. http://cerna-ethics-allistene.org/digitalAssets/38/38704_Avis_robotique_livret.pdf

[4] Chauvier, S. L'éthique artificielle. In press for *L'Encyclopédie philosophique* (Ed. M. Kristanek)

[5] Earle, T.C. & Siegrist, M. On the relation between trust and fairness in environmental risk management. *Risk Analysis,* 28 (2008), 1395-1413.

[6] Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. Ethical choice in unforeseen circumstances. In *Towards Autonomous Robotic Systems*, A. Natraj, S. Cameron, C. Melhuish, M. Witkwoski (eds.), 14th Annual Conference, Oxford, UK, August 28-30, (2013) 433-445.

[7] Gibson, J.J & Crook, L.E. A theoretical field-analysis of automobile-driving. *The American Journal of Psychology,* 51 (1938), 453-471.

[8] Goodall, N. Ethical Decision Making During Automated Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board,* 2424 (2014), 58–65.

[9] Guarini, M. Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, 21 (2006), 22–28.

[10] Johansson-Stenmann, O. & Martinson, P. Are some lives more valuable? An ethical preferences approach. *Journal of Health Economics*, 27 (2008), 739-752.

[11] Klein, G., Orasanu, J., Calderwood, R., & Zsmbok, C.E. *Decision making in actions: Models and methods*. Ablex, Norwood, 1993.

[12] Lin, P. Why ethics matter for autonomous cars? In *Autonomes fahren* (M. Mauer, J.C. Gerdes, B. Lenz, & H. Winner Eds.). Springer, Berlin, 2015.

[13] McLaren, B. Computational models of ethical reasoning: Challenges, initial steps and future directions. *IEEE Intelligent Systems,* 21 (2006), 29–37.

[14] Parfit, D. *On What Matters.* Oxford University Press, Oxford, 2011.

[15] Perleviciute, G. & Steg, L. Contextual and psychological factors shaping evaluations and acceptability of energy alternatives: integrated review and research agenda. *Renewable and Sustainable Energy Reviews,* 35 (2014), 361-381.

[16] Siegrist, M. & Cvetkovich, G. Perception of hazards: the role of social trust and knowledge. *Risk Analysis,* 20 (2000), 713–719.

[17] Skitka, L.J., Bauman, C.W., & Lythe, B.L. Limits of legitimacy: Moral and religious convictions as constraints on deference to authority. *Journal of Personality and Social Psychology,* 97 (2009), 567-578.

[18] Terwel, B.W., Harinck, F., Ellemers, N., & Daamen, D.D.L. Competence-based and integrity-based trust as predictors of acceptance of carbon dioxide capture storage (CCS). *Risk Analysis,* 29 (2009), 1129-1140.