IEAI Technical University of Munich
Munich Center for Technology in Society
Institute for Ethics in Artificial Intelligence

Research Brief – February 2021

AI and Autonomous Driving: Key ethical considerations

by Franziska Poszler and Maximilian Geißlinger

Researchers and companies around the world are working on advancing autonomous driving technology with a major goal of increasing road safety and comfort of motorized transport. Autonomous vehicles (AVs) are expected to have a global impact that will change society, the safety of roadways and transportation systems in the future. The field is moving quickly, leaving governance and ethical consideration to catch up. However, it is important to consider policy and ethical challenges and trade-offs, as well as potential solutions, already while AVs are being developed.

More than 25.000 people lost their lives on the roads of the European Union in 2018 (ERSO, 2018). Several studies claim that more than 90 % of these fatalities were caused by human error (Smith, 2013). In light of these facts, researchers and companies around the world are working on advancing autonomous driving technology with a major goal of increasing road safety and comfort of motorized transport. Autonomous vehicles (AVs) are expected to have a global impact that will change society, the safety of roadways and transportation systems in the future.

What are autonomous vehicles and how do they use AI?

The objective of AVs is to move in a goal-oriented way without the intervention of a human driver. In this case, sensors and actuators controlled by intelligent software perform the driving task. According to SAE (SAE International, 2018), five levels of autonomy in AVs are distinguished. While up to level 2 the human driver remains in charge, from level 3 the complete driving task is handed over to the software. At level 3, however, a human driver must be ready to take over the driving task from the system within a predefined time horizon (e.g., of 10 seconds). At level 4, this is no longer necessary under a set of conditions (such as good weather). In contrast, in level 5 the software can drive under all conditions. While level 2 of autonomy is already available in production vehicles today, developers worldwide are working towards level 3 and level 4. All major car manufacturers (such as Volkswagen or Toyota), but also software companies (such as Google's subsidiary Waymo or Apple) and start-ups (e.g., Zoox) currently engage in a technology competition for level 3 and level 4 systems. Most recently, Waymo drew attention with its developments on level 4 by foregoing the presence of a safety driver in their robotaxis in Phoenix (Waymo, 2020). Their vehicles are now only monitored remotely.

One day AVs will have to make decisions which would be morally difficult for humans

In order to further advance the development of autonomous vehicles from level 3, methods from the field of artificial intelligence (AI) are now being used more and more frequently. Particularly in the areas of detection and behavior prediction of other road users and the subsequent decision making of the AV, various AI methods already constitute the state of the art. Thus, the field is moving quickly and, as often is the case with rapid advancing in technology, governance and ethical consideration are left to catch up.

Potential impacts of AV on efficiency, accessibility and safety

In addition to positive impacts on traffic safety, significant efficiency gains are expected due to the elimination and coordination of driving tasks. Researchers and companies already demonstrate alternatives for spending time in the vehicle (Wadud & Huda, 2021). By automating transport routes and thus possibly shifting traffic away from rush hours, traffic could be relieved in the long term and time spent in traffic jams could be reduced. The use of so-called robotaxis or autonomous

buses in urban traffic could also potentially reduce the costs and increase accessibility of mobility in the long term. AVs enable people with limited mobility (e.g., due to age or disability) to increase their participation in traffic. By giving these groups of people the chance of increased mobility, they will gain greater social and societal participation. This effect is particularly important in times of demographic change, where there are increasing numbers of people with limited mobility.

AVs should conduct a responsible assessment and balancing of risks

The potential positive effects that autonomous vehicles are not without controversy, particularly in terms of safety. For example, road safety depends heavily on where and how autonomous vehicles are introduced on public roads. In addition, new sources of danger open up, for example through hacker attacks on the AV. The fears of potential users also include the issue of data privacy and surveillance. Thus, for the final introduction of AVs on public roads, the technological perspective is only one aspect.

It is assumed that one day AVs will have to make decisions which would be morally difficult for humans, and to which industry and research have not yet provided solutions. This is why policymakers as well as car manufacturers have to focus on the inclusion of ethical considerations into the software of AVs.

What are important ethical considerations?

The following represent some of the identified key issues that need to be addressed in AV development and corresponding recommendations to advance the development and implementation of AVs in a responsible manner, as well as some potential solutions to these problems. These insights are based on the findings of the AI4People-Automotive Committee (Lütge et al., 2021) as well as on the work of the ANDRE-project.[1]

(1) **Technical safety:**

Vision Zero states that eventually no one will and should be killed or seriously injured in road traffic (Ministry of Transport and Communications, 1997). In line with this goal, a prime rationale for introducing AVs onto streets is the expected potential of decreasing fatalities that usually would arise from human error (Bartneck et al., 2019). However, to do so, AVs' technical robustness and safety needs to be ensured. In this regard, relevant questions to be addressed are:

- What are 'safe' fallback plans for AVs?
- How can potential threats to AVs (e.g. cybersecurity threats) be prevented?
- How can we experiment with AVs and test AVs on the road without harming humans?

Potential solutions:

To bypass technical failures and outages, AVs need to pass an official test that assures the system's accuracy, reliability and fallback options. Standards such as the IEEE P7009 (Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems) (IEEE, 2019) or the SAE Driving Safety Performance Assessment Metrics (SAE International, 2018) could serve as a baseline to develop appropriate tests. Furthermore, cybersecurity threats are particularly "new" and important to AVs compared to regular vehicles. Therefore, in addition to conventional safety tests, cybersecurity management systems should be developed relying on existing guidelines such as SAE J3061 (SAE International, 2016). Concerning the rollout of AVs, a stepwise approach is recommended meaning that

[1] This is a project of the IEAI, for more information see: https://ieai.mcts.tum.de/research/andre-autonomous-driving-ethics/

simulations and hardware-in-the-loop testing should be conducted before experimenting on open roads (European Commission, 2020a).

(2) **Responsible balancing of risks:**

Realistically, AVs do not need to make decisions between the outright sacrificing of some individuals to protect others. Instead, they need to implicitly decide about who is exposed to greater risks (Bonnefon et al., 2019). For example, adjusting the lateral position of AVs on a lane can influence the risk posed to other traffic participants (e.g., granted distance to cyclists). Therefore, at every time in mundane traffic scenarios, AVs should conduct a responsible assessment and balancing of risks. This balancing should never be based on personal characteristics of individuals such as gender or age (Lütge, 2017), but rather should take into consideration more objective features. In this regard, the relevant questions to be addressed are:

- What are the objective factors that AVs can rely on in their decision-making and risk allocation process?
- How can this be technically implemented in AVs?

By rapidly processing huge amounts of data, AI can replace complex transport systems problems

Potential solutions:

More objective factors are, for example, factors that influence the collision probability and/or the estimated harm, such as the speed of the traffic participants or the impact angle under which the collision would occur (Geißlinger et al., 2021). These risk assessments can then be integrated into the trajectory planning of AVs in the form of an optimization problem and validity checks (e.g., for maximum acceptable risk). A corresponding mathematical formulation of risk in the context of AVs is developed within the ANDRE-project.

(3) **Human agency:**

AVs have enormous potential to influence human agency, either in a positive manner by, for example, offering solutions to mobility-impaired individuals, or in a negative manner by, for instance, restricting self-determined, independent decisions and interventions by drivers. To ensure effective human agency and clarity over personal responsibility during the operation of AVs, relevant questions to be addressed are:

- To what extent and in which situations should humans be able to override an AV?
- Through what exact processes can we enhance human agency in AVs?

Potential solutions:

The admissibility of human override should be conditioned on the level of automation (up to level 3: at any time; level 4: corresponding to safety mechanisms of an AV, perhaps using a time lag; level 5: not required), as well as on the state and behavior of the driver (e.g., impaired ability). Furthermore, the exact processes that are needed to enhance human agency are threefold and include monitoring drivers (e.g., help drivers remain awake through driver availability recognitions systems), training drivers (e.g., on the limitations and capabilities of AVs) and providing external human-machine interfaces (e.g., LED strips to convey perception information) (Lütge et al., 2021).

(4) **Privacy & data governance:**

AVs will need to collect and process a vast amount of data to ensure proper and safe functioning (Future of Privacy Forum, 2017). Despite the AVs dependence on such data, personal privacy still should be respected by, for example, transparently communicating how and what kind of data is collected and governed or by explicitly requesting affirmative consent from the driver. In this regard, relevant questions to be addressed are:

- What types of data inside and outside the AV need to and should be collected?
- Under which circumstances and in which format can valuable data be shared with third parties?

Potential solutions:

First, products or services that collect and share data such as AVs should comply with pertinent data protection standards and regulations including the GDPR, the ePrivacy directive for information access on the terminal equipment of a user (EDPB, 2020; European Commission, 2020b). In addition, manufactures of AVs should follow a strict privacy and data governance policies (Future of Privacy Forum, 2017) that prescribe transparent communication to drivers about data collection and usage, demand affirmative and explicit consent before sensitive data is collected, and allow only limited and anonymous sharing of vehicle data with third parties (including governments) (Lütge et al., 2021).

(5) **Responsibility, liability & accountability:**

In case of an accident where an AV is involved, the vehicle itself cannot be held morally accountable for the outcomes (Gogoll & Müller, 2017). Responsibility will rather be distributed between a various amount of involved parties such as manufacturers, component suppliers, technology companies, infrastructure providers or car holders and drivers. To identify the true cause of an accident and subsequently the responsible party during an investigation, explicit measures of transparency need to be implemented beforehand. In this regard, relevant questions to be addressed are:

- In what way do we need to change regulations on product liability for AVs?
- To what extent should AVs comply with traffic laws?
- What are explicit measures of transparency that allow retrospective investigation of the true cause of an accident where an AV was involved?

Potential solutions:

As mentioned earlier, due to the increasing involvement of various parties during the development and operation of AVs, regulations on (product) liability need to be reviewed and adapted (European Commission, 2018). For example, one could argue that liability should be determined by the driver's level of autonomy and solo action (Lütge et al., 2021). To test such different regulatory approaches in a controlled manner, regulators could introduce Law Labs (Joaquin Acosta, 2018), similar to regulatory sandboxes. Lastly, applicable measures of transparency could be to prescribe storing records and data of the underlying system logic (e.g., used training data sets) (European Commission, 2020b) and implementing logging mechanisms and black boxes into AVs (e.g., event data recorder) (Lütge, 2017).

To ensure that AVs are programmed and function in a non-discriminatory manner, the systems need to be trained and tested for unfair bias

(6) **Non-discrimination & inclusiveness:**

Past studies have shown that implicit biases and discrimination may unintentionally be incorporated into algorithms (e.g., Goddard et al., 2015). For example, some AI object detection systems are less likely to detect pedestrians with darker skin color compared to those with lighter skin (Wilson, Hoffman & Morgenstern, 2019), which may influence the occurrence and distribution of fatalities between individuals of different ethnicity. To ensure that AVs are programmed and function in a non-discriminatory manner, the systems need to be trained and tested for unfair bias. In addition, AVs should exhibit a non-discriminatory design, meaning that they are equally usable for and accessible to all individuals (Lütge et al., 2021). In this regard, relevant questions to be answered are:

- How can companies ensure and test that biases are not incorporated into the systems of their AVs and that certain fairness standards are met?
- What exact features need to be included in the design of AVs to allow accessibility and inclusiveness to all individuals?

Potential solutions:

To eliminate biases during the creation of algorithms, companies should test their vehicle's AI system for unfair performance differences across personal characteristics such as skin tone, gender and age (Lütge et al., 2021). In doing so, companies can rely on existing standards such as IEEE P7003 that provides protocols to developers and highlights key criteria for selecting validation data sets (IEEE, 2019). To ensure the possibility of wide-scale adoption and inclusiveness, companies

need to demonstrate plans and actions that show how their AVs can be customized to differing abilities and needs (e.g., possibility to include ramp for entering via a wheelchair) (Lütge et al., 2021).

(7) **Societal & environmental wellbeing:**

In line with the United Nation's Sustainable Development Goals (United Nations, 2015), AVs have great potential to bring forward societal and environmental benefits such as increased mobility, better traffic flow, less congestion and decreased carbon emission. On the other hand, as AVs will make driving more convenient and easy for individuals, it is also likely that per vehicle-mile traveled will increase, potentially leading to greater total pollution and congestion (Geary & Danks, 2019). Therefore, if not managed properly or without according policies in place, inefficiencies and counterproductive effects may arise. In this regard, relevant questions to be answered are:

- How can AVs be deployed to increase societal and environmental benefits?
- How can autonomous vehicles be safely integrated into mixed traffic with human drivers?
- How should the appropriate infrastructure be developed accordingly?

Potential solutions:

To achieve net benefits, the rationale behind introducing AVs should be to enhance mobility (e.g., though increased options offered in public transport) without promoting an increase in overall road traffic that could arise, for example, from private drivers. To moderate demand and incentivize more socially and environmentally optimal travel choices, for instance, the implementation of congestion pricing schemes or road tolls has been proposed (Simoni et al., 2019).

Furthermore, since AVs will be gradually rolled out onto streets, the co-existence of conventional vehicles and AVs will be inevitable. Therefore, it is necessary to adapt the physical and digital infrastructure simultaneously to allow mixed vehicle traffic flows (Lütge et al., 2021).

Several programs, such as the Inframix project, work on designing and testing physical and digital elements (e.g., novel visuals signs or electronic signals) that may be relevant for the road infrastructure of mixed vehicle flows (Inframix, 2020). Such research efforts will be key to prepare for the introduction of AVs without jeopardizing safety and efficiency of the road network. The use of AVs as shared mobility and in connection with electromobility also has great potential to positively influence environment in the long term.

Final Thoughts

In this research brief, we highlighted some pressing questions that relate to important ethical considerations in the field of autonomous driving. Certainly, incompatibilities and tradeoffs between these ethical considerations can emerge. For example, AVs may meet the principle of inclusiveness by offering greater mobility for all individuals but, at the same time, AVs may decrease environmental wellbeing if the amount of overall travel and congestion rises as a result of better accessibility and convenience. AI can play a major role in mitigating some of these tradeoffs. For example, by rapidly processing huge amounts of data, AI can replace complex transport systems problems (e.g., traffic congestion or overcrowding) with smart traffic (Voda & Radu, 2018).

However, what becomes evident from this argument is that vast amounts of data will be necessary for bypassing inefficiencies. This draws attention to another important tradeoff, namely that AVs may meet the principles of technical safety, responsible balancing of risks and accountability, but this may come at a cost of needing increased access to and disclosure of personal data (such as the vehicle's position). In the future, industry, policymakers, researchers in the automotive sector will need to focus on the above-identified issues, develop an agreement on compromises and prioritization among these ethical considerations, as well as advance relevant solutions.

References

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2019). Ethik in KI und Robotik. Carl Hanser Verlag GmbH Co KG.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2019). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. Proceedings of the IEEE, 107(3), 502-504.

ERSO, "Annual Accident Report 2018, European Road Safety Observatory", pp. 1–86, 2018.Retrieved from https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/asr2018.pdf

European Commission (2018). Liability for emerging digital technologies. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018SC0137&from=en

European Commission (2020a). Ethics of connected and automated vehicles. Retrieved from https://ec.europa.eu/info/sites/info/files/research_and_innovation/ethics_of_connected_and_automated_vehicles_report.pdf

European Data Protection Board (EDPB) (2020). Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications. Retrieved from https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_202001_connectedvehicles.pdf

Future of Privacy Forum (2017). Data and the connected car. Retrieved from https://fpf.org/wp-content/uploads/2017/06/2017_0627-FPF-Connected-Car-Infographic-Version-1.0.pdf

Geary, T., & Danks, D. (2019). Balancing the benefits of autonomous vehicles. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 181-186). New York, NY: Association for Computing Machinery.

Geißlinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous driving ethics: From Trolley problem to ethics of risk. Working paper.

Goddard, T., Kahn, K. B., & Adkins, A. (2015). Racial bias in driver yielding behavior at crosswalks. Transportation research part F: traffic psychology and behaviour, 33, 1-6.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. Science and engineering ethics, 23(3), 681-700.

IEEE (2019). Ethically aligned design – A vision for prioritizing human well-being with autonomous and intelligent systems. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

Inframix (2020). Expected impact – a step by step introduction of automation. Retrieved from https://www.inframix.eu/expected-impact/

Joaquin Acosta, A. (2018). Autonomous vehicles: 3 practical tools to help regulators develop better laws and policies. Berkman Klein Center for Internet and Society at Harvard University. Retrieved from https://cyber.harvard.edu/sites/default/files/2018-07/2018-07_AVs04_1.pdf

Lütge, C. (2017). The German ethics code for automated and connected driving. Philosophy & Technology, 30(4), 547-558.

Lütge, C., Poszler, F., Acosta, A. J., Danks, D., Gottehrer, G., Mihet-Popa, L., & Naseer, A. (2021). AI4People: Ethical Guidelines for the Automotive Sector–Fundamental Requirements and Practical Recommendations. International Journal of Technoethics, 12(1), 101-125.

Robinson, J. (2014). Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells Us about Right and Wrong.

SAE International (2018). J3016 – Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Retrieved from https://saemobilus.sae.org/content/J3016_201806

SAE International – Vehicle Cybersecurity Systems Engineering Committee (2016). SAE J3061: Cybersecurity guidebook for cyber-physical vehicle systems. Retrieved from https://www.sae.org/standards/content/j3061_201601/

Simoni, M. D., Kockelman, K. M., Gurumurthy, K. M., & Bischoff, J. (2019). Congestion pricing in a world of self-driving vehicles: An analysis of different strategies in alternative future scenarios. Transportation Research Part C: Emerging Technologies, 98, 167-185.

Smith, B. 2013. "Human Error as a Cause of Vehicle Crashes" Retrieved from http://cyberlaw.stanford.edu/blog/2013/12/human-error-cause-vehicle-crashes

Thomson, J. J. (1984). The trolley problem. Yale LJ, 94, 1395.

United Nations (2015). Transforming our world: The 2030 agenda for sustainable development. Retrieved from https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

Voda, A. I., & Radu, L. D. (2018). Artificial intelligence and the future of smart cities. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 9(2), 110-127.

Waymo (2020). Waymo is opening its fully driverless service to the general public in Phoenix Retrieved from https://blog.waymo.com/2020/10/waymo-is-opening-its-fully-driverless.html

Wadud, Z. & Huda, F. (2021) Fully automated vehicles: the use of travel time and its association with intention to use. Proceedings of the Institution of Civil Engineers - Transport 0 0:0, 1-15

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. arXiv preprint arXiv:1902.11097.