

Trustworthy Artificial Intelligence Requirements in the Autonomous Driving Domain

David Fernández-Llorca , European Commission Joint Research Center and University of Alcalá

Emilia Gómez, European Commission Joint Research Center and Pompeu Fabra University

We identify the maturity level of the different requirements for artificial intelligence (AI) in autonomous driving and outline the main challenges to be addressed in the future to ensure that automotive AI systems are developed in a trustworthy way.

It is increasingly accepted that the adjective “trustworthy” linked to artificial intelligence (AI) refers to something that goes beyond the word’s literal meaning (which, according to the Oxford English Dictionary, is that something can be relied on to be good, honest, sincere, and so forth). Rather, trustworthy AI is a concept that encompasses multiple ethical principles, requirements, and criteria to guarantee that AI systems are designed following a human-centered approach and committed to social good.

Although the development of human-centric AI is a common trend worldwide, probably the most solid steps

so far have been taken in Europe. In April 2018, the European Commission published its AI strategy,¹ and in June 2018, the High-Level Expert Group on AI (AI HLEG) was set up to make recommendations for how to address mid- and long-term challenges and opportunities related to AI. In April 2019, the AI HLEG published “Ethics Guidelines for Trustworthy AI,”² setting out a set of seven key requirements that AI systems should fulfill to be deemed trustworthy. These requirements were further elaborated with the publication of “Assessment List for Trustworthy AI” in July 2020,³ defining several specific criteria for each requirement. Finally, in April 2021, the European Commission presented “Proposal for a Regulation Laying Down Harmonized Rules on AI” (at the time

Digital Object Identifier 10.1109/MC.2022.3212091
Date of current version: 8 February 2023

of writing, the regulation was being negotiated among legislators in the European Council and the European Parliament).⁴

This regulation adopts a risk-based approach, defining four levels of risk to safety and fundamental rights (unacceptable risk and prohibited practices, high risk, requiring transparency obligations, and minimal risk), so that AI systems used in

Autonomous vehicles are a challenging scenario, due to the high number of components involved, some of them incorporating AI, and the critical need for safety. Their complexity can be seen in the numerous global efforts to develop regulations (for example, United Nations Economic Commission for Europe WP.29) and standards [such as International Organization for Standardization (ISO) 21448, ISO

Automotive Engineers (SAE) levels⁵ 1 and 2 (driver) for assisted driving, SAE level 3 (a backup driver/user is in charge) for automated driving, and SAE levels 4 and 5 (passenger/unoccupied) for autonomous driving.] Autonomous vehicles have the potential to create new mobility services, develop new shared mobility schemes, and respond to the growing demand for the mobility of goods and people. Considering the importance of human factors in most accidents (for example, errors, distractions, and traffic violations), autonomous vehicles could significantly improve road safety. In addition, they could extend mobility to people who are unable to drive conventional vehicles themselves, either because of physical (for example, elderly people with visual impairment and disabilities) and legal (for instance, people without a driver's license and teenagers) issues. Moreover, autonomous driving, through its innovative nature, implies the acceleration of vehicle electrification and connectivity as well as the potential to free up urban public spaces that are currently used for parking.

Achieving high levels of automation in global scenarios (such as urban environments) without the need for a backup driver and allowing empty trips to pick up passengers (for example, shared mobility) and find parking places (for instance, private vehicles) requires solving highly complex problems in environments with an almost infinite variety of possible situations and interactions. Among the different problems that have to be addressed in autonomous driving,⁶ we highlight localization (Where am I? Where am I heading?), scene understanding (detecting agents and predicting their motion), local path planning (decision making and local ego vehicle

TRUSTWORTHY AI IS A CONCEPT THAT ENCOMPASSES MULTIPLE ETHICAL PRINCIPLES, REQUIREMENTS, AND CRITERIA TO GUARANTEE THAT AI SYSTEMS ARE DESIGNED FOLLOWING A HUMAN-CENTERED APPROACH.

high-risk contexts need to fulfill a set of requirements that are coherent with the ones defined by the AIHLEG. The requirements defined in the legal text need to be further detailed and standardized, with a main challenge being the definition of methodologies that are valid for different application domains, for example, from school admission to biometric identification. This approach represents a paradigm shift, and the complexity lies in finding a solution that is both specific enough to ensure compliance with safety and fundamental rights objectives and general enough to be applicable to different domains. In any case, it is reasonable to assume that in some sectors, it will be necessary to adapt the requirements to the specifics of the application context.

8800, ISO 4804, ISO 5083, and UL 4600] to ensure the safety and cybersecurity of these systems. But, as seen in the following, the trustworthy AI concept incorporates several requirements that go beyond safety. This article analyzes AI in this particular context, considering the seven requirements for trustworthy AI systems, their maturity, and related challenges in a comprehensive way.

FROM TRUSTWORTHY AI TO TRUSTWORTHY AUTONOMOUS VEHICLES

AI plays a fundamental role in autonomous vehicles, a domain with a great disruptive potential at the social, environmental, and economic level. [Our proposed approach for referring to vehicles with automated driving systems is to consider Society of

trajectories), control (lateral and longitudinal), and human interaction (people inside and outside vehicles).

Although AI systems are not the only ones required to address these different tasks/layers—AI has become more present in some of these other systems than in others during recent years—AI is positioning itself as a core technology in almost all instances, being indispensable in the most complex cases where conventional algorithmic solutions are not sufficient. Moreover, it is important to note that we assume a definition of AI in line with Europe’s Artificial Intelligence Act,⁴ which not only includes machine learning but also considers, for instance, knowledge-based approaches. In addition, on the one hand, there can be multiple AI systems to address different parts of the same problem and even systems tackling the same problem with combined outputs for better performance and robustness (for example, sensor fusion). On the other hand, some systems depend on the outputs of others (for instance, path planning depends on localization and scene understanding), so there is a considerable degree of interdependence.

All this raises questions regarding the degree of abstraction needed to implement the requirements for trustworthy AI. Considering the autonomy that each system may have, an approach at the level of each AI system might seem the most reasonable option at first glance. However, its effective implementation would be intractable due to the number of systems, the dependencies, and the number of requirements. We must also take into account the interrelationships and dependencies existing not only among different AI systems (see Figure 1) but also among the requirements themselves (for example, human agency and oversight are highly correlated with transparency, and fairness and accuracy should be addressed jointly). All this suggests that a holistic approach is most appropriate. In other words, there is a need to map the general requirements of trustworthy AI into a set of specific requirements for trustworthy autonomous vehicles. That is precisely the goal of this article.

The following is a brief discussion of the most important elements to consider to adapt the seven key requirements for trustworthy AI² to the autonomous driving context. The

main conclusions are summarized in Table 1. For a better understanding of the criteria used, we refer the reader to the original documents^{2,3} published by the AI HLEG and an extended report on trustworthy autonomous vehicles.⁷

TRUSTWORTHY FOR WHOM?

As discussed in the following sections, the various requirements for ensuring that AI systems are trustworthy are focused within a context of use and interaction with humans (that is, human-centric). The requirements are set in relation to the people who are affected by the technology. In many cases, we can consider that a technology brings risk to safety and fundamental rights to certain stakeholders. For example, in AI-based medical devices, we can have the patient in mind as the affected stakeholder, while in AI systems used for university admission, we may consider students the citizens to be affected. We can then place these stakeholders in the center when designing AI systems. In the case of autonomous driving, however, and similar to other contexts, a heterogeneous multistakeholder approach is a must.

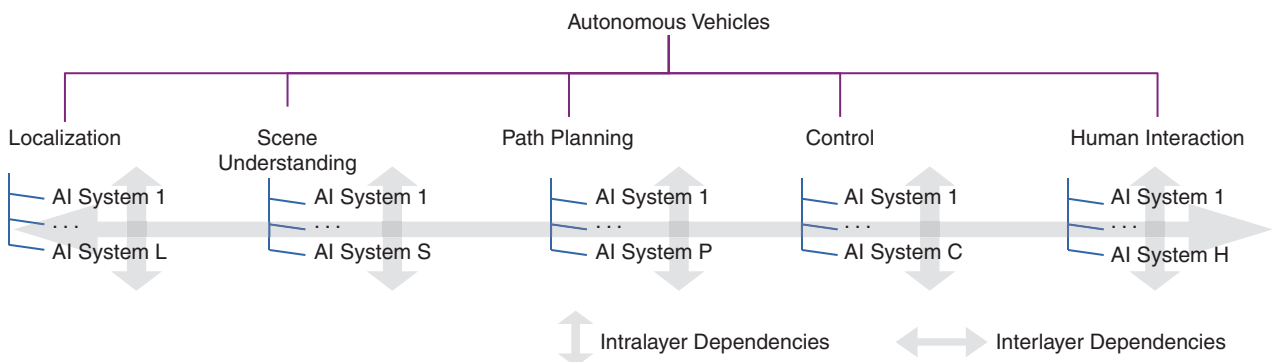


FIGURE 1. The main tasks/layers of autonomous vehicles. Multiple AI systems per layer and inter- and intralayer dependencies are evident.

Figure 2 illustrates the different users and related perspectives that need to be integrated into the human-centered view, corresponding to two main locations with respect to vehicles. On the one hand, there are different users inside a vehicle: (assisted) drivers (for SAE levels 1 and 2), backup/assistant drivers and users in charge (for SAE level 3), and mere passengers with no responsibility for driving tasks (for SAE levels 4 and 5). On the other hand, there are users outside the vehicle; that is, there are other road users interacting with the vehicle (even if they are mere bystanders). These

TABLE 1. The requirements for trustworthy AI systems in the autonomous driving domain.

Requirement	Subrequirements	Particular aspects	Maturity	Future challenges
Human agency and oversight	Human agency and autonomy Human oversight	Multimodal/multiuser perspective New skills	Low	Human agency calibration Agency-oriented HMIs/external HMIs Vehicle status communication
Technical robustness and safety	Resilience to attacks and security General safety Accuracy and reliability (fallback and reproducibility)	Scenario variability Minimal risk condition Multidimensional accuracy Intention to use	High	Heterogeneous constantly updated security approach Safety assessment New testing approaches Fallback strategies Minimal risk condition
Privacy and data governance	Privacy Data governance	Privacy versus safety Agent behavior modeling	Low	Data anonymization-preserving attributes Privacy by design Multiuser consent
Transparency	Traceability Explainability Communication	High complexity Data-driven components Multiuser explainability	Medium	New types of artefacts Explainable models without affecting accuracy Intelligent data logging Multistakeholder communication
Diversity, nondiscrimination, and fairness	Avoid unfair bias Accessibility and universal design Stakeholder participation	High variability of agent attributes Same safety level for all road users Data bias	Low	Adaptive behavior Unbiased perception Unbiased human–vehicle interaction Unbiased service provision
Societal and environmental well-being	Environmental well-being Work and skills Society at large	Multidimensional problem High uncertainty	Medium	Well-being criteria Lower uncertainty estimates Reskilling New vehicle interiors and uses
Accountability	Auditability Risk management	External auditing Liability (burden of proof) New risks	Low	Independent external auditing New liability approaches Alleviating the burden of proof Balanced insurance and liability costs

HMI: human–machine interface.

The “Maturity” column refers to the scientific and technological state of the art. Low: basic principles formulated and exploratory proof of concept; Medium: emerging research lines and experimental validation in the lab; High: consolidated research line and experimental validation and demonstration in relevant environments.

The interpretation of this table can be derived from the text. Further details and references are available in Fernández-Llorca and Gómez.⁷

include drivers and passengers of other conventional, assisted, automated, and autonomous vehicles; vulnerable road users (VRUs), such as pedestrians, cyclists, and wheelchair users; and users of personal mobility devices, such as electric unicycles, scooters, Segways, carts, and hoverboards.

In addition to the different mentioned users, human-centered approaches in autonomous vehicles should also be adapted to the specific level of automation (that is, the type of in-vehicle user) and incorporate multiple dimensions, perspectives, characteristics, and types of agents and interactions that need to be jointly addressed. When the objectives of different agents coincide, optimal and efficient solutions can be obtained. However, there are cases in which different types of agents (for example, an autonomous vehicle and a VRU) may have conflicting objectives and interests. For example, the well-known “cross-walk chicken problem”⁸ may be intensified in the case of autonomous driving, as VRUs’ perceived risk of crossing may become practically nonexistent if people trust that an autonomous vehicle will ultimately stop.

This could lead to abusive behavior by VRUs, which, if widespread, could significantly slow down the travel time of autonomous vehicles compared to conventional ones. This conflicting situation can be addressed only by the implementation of several strategies, including educational and even punitive measures (let us not forget that autonomous vehicles will have multiple data captured by sensors that could serve as a basis for proving potentially dangerous behaviors by different types of agents). In the following, we consider the seven key requirements, keeping in mind these mentioned perspectives.

HUMAN AGENCY AND OVERSIGHT

This first requirement relates to the need for AI systems to support human autonomy and decision making by facilitating users’ agency, allowing human oversight, and fostering fundamental rights. Human agency in autonomous vehicles is directly related to the principle of human autonomy. It affects both user acceptance and safety through disuse and misuse, respectively. New agency-oriented in-vehicle and external human–machine interfaces (HMIs) need to be developed to ensure a suitable agency level. To this end, new multimodal approaches are needed to assess and calibrate the sense of agency.⁹

When considering the different involved stakeholders, human oversight mostly relates to in-vehicle users, and it depends on the automation level. In addition, oversight may be also

exercised, to a certain extent, by external road users, as their behavior may influence an autonomous vehicle’s behavior (including a potential risk of abuse in the interaction, due to the fact that autonomous vehicles will always eventually stop). For the interaction to be adequate, a mutual understanding between the autonomous vehicle and the users with whom it interacts should be established. This includes the design of adequate mechanisms to represent and communicate the operating status of the autonomous vehicle to different user types (inside and outside the vehicle), which is related to the transparency requirement and is a key area of future research. Finally, we can expect that adequate oversight by drivers (assisted and automated) and passengers (autonomous) will require new skills, acquired either before interactions or developed during exposure and use.

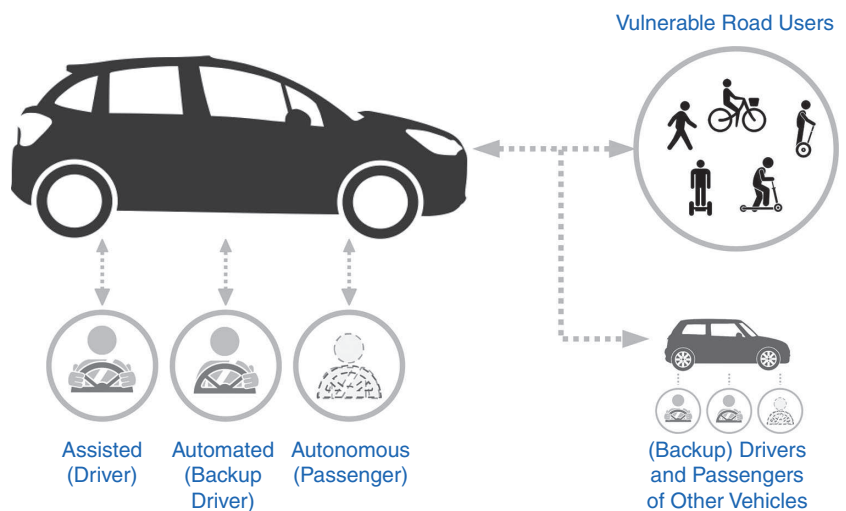


FIGURE 2. The trustworthiness, multiuser perspective, and human–vehicle interaction channels. Requirements for trustworthy autonomous vehicles must consider users inside a vehicle and other external road users. Adapted from Fernández-Llorca and Gómez⁷ with permission.

TECHNICAL ROBUSTNESS AND SAFETY

This requirement is closely linked to the principle of prevention of harm. Technical robustness requires that AI systems behave in a reliable way, minimize unintentional and unexpected harm, and prevent unacceptable harm. In addition, the physical and mental integrity of humans should be ensured. This requirement has a strong impact on user acceptance, as safety and cybersecurity are fundamental variables highly correlated with intent to use.¹⁰ Resilience to adversarial attacks and the security of autonomous vehicles must be designed to be foundationally secure (that is, security by design) through a heterogeneous and constantly updated approach.¹¹ This should include multiple types of defensive measures, including cryptographic methods; intrusion and anomaly detection mechanisms; countermeasures against adversarial attacks (for example, redundancy and hardening); fault-tolerant, fail-x (for example,

fail-aware, fail-safe, or fail-operational), and self-healing methods; and proper user training.

To establish whether an autonomous vehicle is sufficiently safe (the safety gain Δ depicted in Figure 3), novel methods are needed for a comparative assessment of safety in autonomous vehicles versus human drivers, without requiring endless periods of testing. Expectations of safety gains that are too high could be detrimental to user acceptance. Even small improvements in safety by autonomous vehicles relative to human drivers can save many lives, so public expectations must be appropriately calibrated so as not to delay the adoption of autonomous vehicles and the benefits of the technology.

In recent years, considerable effort has been made by academia, industry, and regulatory bodies to develop new safety test procedures for automated driving systems. There is some initial consensus that future

approaches should be multisystem, including not only physical testing on proving grounds but also extensive use of simulators and real-world driving tests.¹² However, open issues remain, such as the absence of real road agent behaviors in simulation environments; the limited variability compared to the almost infinite variety of scenarios, conditions, interactions, and so on of real driving; and the lack of scenarios to evaluate new trustworthy requirements, such as human agency and oversight, transparency, and fairness.

One of the most important issues for achieving high levels of automation is the ability of the autonomous driving system to reach a minimal risk condition after the occurrence of some performance-relevant failure and upon operational design domain (ODD) exit. This feature is what allows users of autonomous vehicles (SAE levels 4 and 5) to be considered passengers. In fact, the inability to achieve this condition in automated vehicles (SAE level 3) requires a backup driver to resume control in the event of failure and ODD exit. The minimal risk condition will obviously depend on the specific scenario and operating conditions. Considering the enormous variability, the development of specifications for the minimal risk condition is a major challenge, and a clear taxonomy is needed, including new fallback strategies as well as testing procedures to assess their safety and robustness.

In addition, it is important to note that determining the accuracy of autonomous vehicles is a multidimensional problem, which involves multiple metrics, levels, layers, use cases, and scenarios. Defining holistic metrics and thresholds to assess the

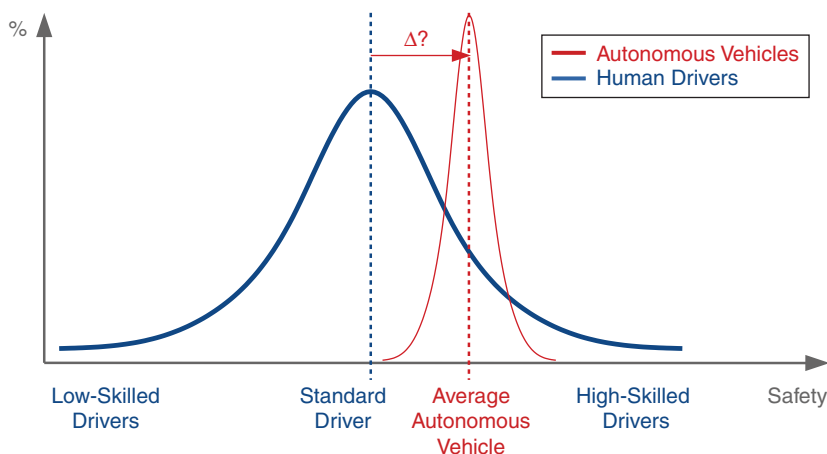


FIGURE 3. The safety distribution for human (low-skilled, standard, and high-skilled) drivers and autonomous vehicles. We assume that AVs will be more consistent across their population than humans. Adapted from Fernández-Llorca and Gómez⁷ with permission.

trustworthiness of autonomous vehicles is a challenging research area and, in some cases, a policy-based problem to be further addressed.⁷ Finally, it is worth mentioning that any substantial change to an AI-based component of autonomous vehicles that may modify the overall behavior, or a significant part of it, must meet all relevant trustworthiness requirements and may need to be retested. This statement is essential for this requirement, but it also applies to all others.

PRIVACY AND DATA GOVERNANCE

This requirement states that privacy and data protection must be ensured in AI systems of autonomous vehicles throughout their entire life cycle. New innovative methods have to be implemented that ensure data protection without adversely affecting the safety of autonomous vehicles. These methods may include techniques for data anonymization and the deidentification of agent-specific data while preserving relevant agent attributes that can be pertinent to guarantee the correct operation of the system. For example, intrinsic attributes of agents may be of fundamental importance when it comes to predictive perception since the behavioral patterns of road agents may depend on certain intrinsic characteristics, such as the sex, age, and group behavior of VRUs¹³ or the type of vehicle.¹⁴

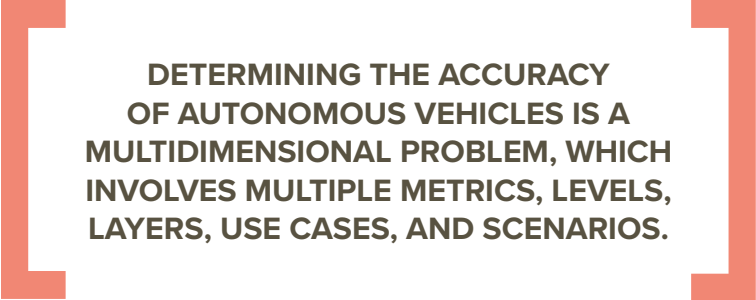
Privacy by design requires the implementation of a set of comprehensive strategies. These include the setup of mechanisms for encrypting data, storage devices, and vehicle-to-everything communication channels; the implementation of unique encryption key management systems for each

vehicle; and the regular renewal of encryption keys.¹¹ These measures are also linked to security and safety requirements.

In terms of personal data processing, there are two aspects to consider. First, for users inside a vehicle (that is, drivers and passengers), safety risks should be the main aspect to minimize when defining which personal data should be shared with other vehicles and infrastructures. It is important to note that many of the privacy issues that arise when using an autonomous vehicle are not very different

decisions made by an autonomous vehicle. Transparency allows the identification of the reasons why an autonomous vehicle decision was erroneous, which, in turn, could help to prevent future mistakes. This requirement includes different aspects, such as traceability, data logging, explainability, and communication strategies.

Traceability, which facilitates auditability as well as explainability, is a well-known challenge for modern conventional vehicles,¹⁵ so in the case of autonomous vehicles, traceability is a problem of considerable difficulty.



DETERMINING THE ACCURACY OF AUTONOMOUS VEHICLES IS A MULTIDIMENSIONAL PROBLEM, WHICH INVOLVES MULTIPLE METRICS, LEVELS, LAYERS, USE CASES, AND SCENARIOS.

from those that exist when using a smartphone (for instance, the possibility of location-based tracking). In these cases, the collection and processing of personal data must be subject to some legal basis, such as user consent. Second, for people outside the vehicle (that is, VRUs and other drivers), consent may be impossible to obtain without disproportionate efforts. However, the problem can be effectively circumvented if personal data are processed in real time and if data deidentification is properly implemented.

TRANSPARENCY

The transparency requirement in this context establishes the need for humans to understand and trace the

Despite ongoing efforts, the effective integration of components of data-driven AI systems as traceable artifacts is still an open research question. One of the main enablers of transparency is data logging systems. In the field of autonomous driving, data collection requirements go far beyond event recording. There is a need for the continuous logging of input and output data as well as intermediate states of the decision-making systems. The bandwidth and storage capacity requirements are so demanding that new mechanisms for intelligent data logging are needed.

Regarding explainability, further research is needed focusing on explainable models and methods¹⁶ and, more specifically, on explanations

to in-vehicle and external road users, that is, explainable human-vehicle interaction through new HMIs and external HMIs. Considering explainability as a possible requirement in future vehicle type approval frameworks will clearly improve safety assessment and facilitate the evaluation of human agency and oversight and transparency. However, it will require new test procedures, methods, and metrics. In addition, the design of human-scale interpretable models must be done without detriment to accuracy, which is one of the most relevant challenges in the field of explainable AI. Last but not least, both in-vehicle users (for example, backup drivers and passengers) and external road users must receive clear information that allows them to understand that they are interacting with an automated or autonomous vehicle. New communication mechanisms are needed for this, including new ways of communicating risks.

DIVERSITY, NONDISCRIMINATION, AND FAIRNESS

Trustworthy systems need to ensure inclusion and diversity throughout their entire life cycle by considering all relevant stakeholders, designing inclusive processes, and ensuring equal treatment in line with the fairness principle. In this respect, and to prevent discrimination, autonomous vehicles should avoid making decisions based on social values and the characteristics of some group of users (for example, dilemmas, age, and gender). Instead of focusing the debate on moral dilemmas whose occurrence is highly unlikely, our proposal is to define an alternative objective. Specifically, we consider

that autonomous vehicles should be designed in a way that they can ensure the same level of safety for any kind of road user. (Here, *safety* refers to the intended functionality, that is, the absence of unreasonable risk due to hazards resulting from functional insufficiencies of the intended functionality.) To do so, they will have to react differently for different kinds of road agents to correct possible potential safety inequalities caused by different behaviors.

In this sense, we may need to model the behavior of different road agents, using, for instance, real-time predictive perception and local patch planning systems, to react in a personalized way. For example, a predictive detection system might anticipate, for a child, a higher likelihood of being at a curb crossing in front of a vehicle than for an adult. The local path planner might then need to reduce the vehicle speed for the child while maintaining it for the adult. This does not necessarily mean that adults are being discriminated against (as long as there is no bias in the data and algorithms used to develop the predictive perception systems) but that the uncertainty in the behavior and possible motion of the child is de facto greater than in adults. Therefore, the autonomous vehicle must adapt its behavior to provide children and adults the same safety level by implementing a more conservative speed profile in one case than in the other one. This is an area of research that is not yet mature, and new approaches are needed.

Datasets used may suffer from the inclusion of bias, incompleteness, and bad governance models, and the way in which AI systems are developed (for example, algorithms'

programming) may also suffer from unfair bias. The available work in this area is still very preliminary.^{17,18} Further research is then needed to identify possible sources of discrimination in state-of-the-art perception systems when detecting external road agents, considering different inequity attributes, such as sex, age, skin tone, group behavior, type of vehicle, color, and so on. Unfair bias may also be present at the user-vehicle interaction level. Universally accessible and adaptable HMIs must also be designed, which is challenging, considering that autonomous vehicles have the potential to extend mobility to multiple types of new users. Finally, as autonomous driving opens up new autonomous mobility systems, services, and products, this requirement also states that new mechanisms should be put in place to avoid any service provision approach that may discriminate against certain user groups.

SOCIETAL AND ENVIRONMENTAL WELL-BEING

This requirement establishes the importance of the specific social (for instance, social agency, relationships, attachment, skills, and physical and mental well-being) and environmental impacts of autonomous vehicles. Understanding and estimating these impacts is a highly multidimensional and complex problem involving many factors, such as decarbonization, platooning, eco-optimal driving, traffic congestion, shared mobility, vehicles per kilometer, road accidents, user acceptance, and extended mobility, to name a few (more details are in Figure 4), for which we can make predictions based only on yet uncertain

assumptions.¹⁹ New studies and approaches are needed to provide estimates with lower uncertainty and higher precision.

A widely discussed societal impact of AI is the one about jobs and skills.²⁰ In this context, we can consider that automated vehicles (SAE level 3) are not expected to have a major negative impact on jobs. Rather, new tasks and skills for backup drivers will be needed, primarily for tasks related to interaction with the automation system. For high levels of automation (SAE levels 4 and 5), as drivers are not needed within the ODD, the expected impact on work and skills is likely to be negative. However, it may be partially mitigated by the need for non-driving tasks that are less susceptible to automation (for example, loading and unloading goods) as well as by the

new jobs and skills brought by transportation automation.

Another potential social impact of autonomous driving is the fact that the technology brings the possibility to use commuting time for work-related activities, potentially leading to higher productivity and a reduction of time at the workplace, as travel time could be considered working time. In the coming years, we can expect to see new proposals to transform the interiors of autonomous vehicles into places to work, which is a major challenge, especially in shared mobility scenarios. Finally, from a more general perspective, autonomous driving has the potential to transform the way people spend their commuting time on the road, including new possibilities for social interaction, as the interior cabins of autonomous vehicles are likely

to be significantly modified to allow greater flexibility for work, leisure, and social activities.

ACCOUNTABILITY

This requirement is mainly focused on auditability and risk management. First, mechanisms must be put in place to ensure responsibility and accountability for autonomous vehicles and their outcomes, both before and after their development, deployment, and use. As a safety-critical application, and as an additional complement to self-audits, AI systems of autonomous vehicles should be able to be audited by independent external auditors. The main challenge here is to establish the minimum set of requirements for third parties to audit systems, without compromising intellectual and industrial property.

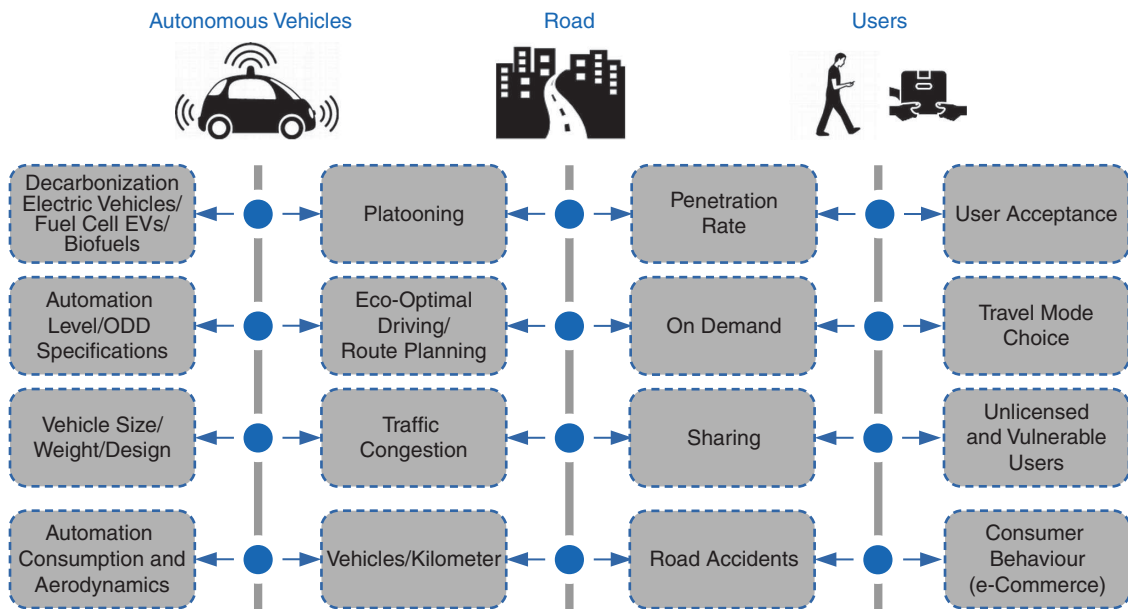


FIGURE 4. The key environmental and social factors in the development and adoption of autonomous vehicles. They are categorized based on three main components: vehicles, road infrastructure, and users. The factors are categorized according to the two most important components. See Fernández-Llorca and Gómez⁷ (reproduced here with permission) for more details.

ABOUT THE AUTHORS

DAVID FERNÁNDEZ-LLORCA is currently a scientific officer with the European Commission Joint Research Center, 41092 Seville, Spain, and a full professor in the Department of Computer Engineering, University of Alcalá, Alcalá de Henares 28805, Spain. His research interests include trustworthy artificial intelligence for transportation, predictive perception for autonomous vehicles, human–vehicle interaction, end user-oriented autonomous vehicles, and assistive intelligent transportation systems. Fernández-Llorca received a Ph.D. in telecommunication engineering from the University of Alcalá. He is the editor in chief of *IET Intelligent Transport Systems* and a Senior Member of IEEE. Contact him at david.fernandez-llorca@ec.europa.eu.

EMILIA GÓMEZ is the principal investigator on the Human Behavior and Machine Intelligence project of the European Commission Joint Research Center, 41092 Seville, Spain, and a guest professor at Pompeu Fabra University, Barcelona 08002, Spain. Her research interests include music information retrieval, where she develops technologies to support music listening experiences. Gómez received a Ph.D. in computer science from the Pompeu Fabra University. She is a member of the Spanish National Council for AI and the Organization for Economic Cooperation and Development One AI expert group. Contact her at emilia.gomez-gutierrez@ec.europa.eu.

It is important to note that the same requirements and expertise needed to audit AI systems of autonomous vehicles would be also needed for victims and insurers to claim liability in accidents involving autonomous vehicles. Due to the high complexity and opacity of these systems, proving defects and fault for victims

would be a very complex and costly process. Shifting the burden of proof to the manufacturer would make these systems more victim friendly. It is also a more economically efficient approach, as manufacturers will always have a much more favorable position to access and properly interpret the data. It is then necessary to update and harmonize the existing national and international regulatory frameworks for product liability, traffic liability, and fault liability. The adoption of autonomous vehicles will entail new types of risks, including risks that are unknown at the time of production and may emerge only after market launch. Policymakers and stakeholders must define new balanced policy


frameworks to better accommodate insurance and liability costs between consumers and victims, on the one hand, and autonomous vehicles providers, on the other.

One of the most important changes when thinking about trustworthy systems is to go beyond classic requirements of safety and accuracy toward a more comprehensive human-centric framework that also considers criteria such as human agency and oversight, security, privacy, data governance, transparency, explainability, diversity, fairness, social and environmental well-being, and accountability. Developing and implementing these requirements is a challenge for any application domain. It is probably still too early to know to what extent attempts to generate a horizontal sector-independent set of requirements will be effective, but for applications as complex as autonomous driving, it is reasonable to think that such a comprehensive approach will need to be tailored to the specifics of the sector.

As discussed in this article, the application of the requirements for trustworthy AI systems for autonomous vehicles involves addressing multiple problems of different natures, some of them still at a very early stage of scientific and technological maturity, bringing new research and development challenges in different areas. The introduction of not only safety criteria but also requirements related to fundamental human rights in future type approval procedures for autonomous vehicles is a considerable challenge, but it will clearly serve as an accelerator and driver for the development and adoption of a technology

DISCLAIMER

The views expressed in this article are purely those of the authors and may not, under any circumstances, be regarded as an official position of the European Commission.

that can change transportation as we know it. 

ACKNOWLEDGMENT

The authors acknowledge main funding from the Human Behavior and Machine Intelligence project of the European Commission Joint Research Center. Other research grants that have partially contributed to this work were provided by Community Region of Madrid (S2018/EMT-4362 and SEGVAUTO 4.0-CM) and Spanish Ministry of Science and Innovation (DPI2017-90035-R and PID2020-114924RB-I00).

REFERENCES

1. "Artificial intelligence for Europe (Communication from the Commission)," European Commission, Brussels, Belgium, COM 237, 2018.
2. "Ethics guidelines for trustworthy AI," High Level Expert group on Artificial Intelligence, European Commission, Brussels, Belgium, B-1049, 2019.
3. "The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment," High Level Expert group on Artificial Intelligence, European Commission, Brussels, Belgium, B-1049, 2020.
4. "Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts," European Commission, Brussels, Belgium, COM 206, 2021.
5. "(R) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," SAE International, Warrendale, PA, USA, J3016 202104, 2021.
6. S. Liu and J.-L. Gaudiot, "Rise of the autonomous machines," *Computer*, vol. 55, no. 1, pp. 64–73, Jan. 2022, doi: 10.1109/MC.2021.3093428.
7. D. Fernández Llorca and E. Gómez, "Trustworthy autonomous vehicles," Publications Office of the European Union, Luxembourg City, Luxembourg, EUR 30942 EN JRC127051, 2021.
8. A. Millard-Ball, "Pedestrians, autonomous vehicles, and cities," *J. Planning Educ. Res.*, vol. 38, no. 1, pp. 6–12, 2018, doi: 10.1177/0739456X16675674.
9. J. Silva, "Increasing perceived agency in human-AI interactions. Learnings from piloting a voice user interface with drivers on Uber," in *Proc. Ethnographic Praxis Ind. Conf. (EPIC)*, 2020, pp. 441–456, doi: 10.1111/1559-8918.2019.01299.
10. K. Garidis, L. Ulbricht, A. Rossmann, and M. Schäh, "Toward a user acceptance model of autonomous driving," in *Proc. 53rd Hawaii Int. Conf. Syst. Sci.*, 2020, pp. 1381–1390, doi: 10.24251/HICSS.2020.170.
11. "ENISA good practices for the security of smart cars," European Union Agency for Cybersecurity, Athens, Greece, 2019. [Online]. Available: <https://op.europa.eu/en/publication-detail/-/publication/49449381-47cd-11ea-b81b-01aa75ed71a1/language-en>
12. "Future certification of automated/autonomous driving systems," United Nations Economic Commission for Europe, Geneva, Switzerland, UNECE WP29 GRVA, Informal document GRVA-02-09, 2019.
13. R. Quintero Mínguez, I. Parra Alonso, D. Fernández-Llorca, and M. Á. Sotelo, "Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1803–1814, May 2019, doi: 10.1109/TITS.2018.2836305.
14. H. Corrales-Sánchez, N. Hernández-Parra, I. Parra-Alonso, E. Nebot, and D. Fernández-Llorca, "Are we ready for accurate and unbiased fine-grained vehicle classification in realistic environments?" *IEEE Access*, vol. 9, pp. 116,338–116,355, Aug. 2021, doi: 10.1109/ACCESS.2021.3104340.
15. S. Maro, J.-P. Steghöfer, and M. Staron, "Software traceability in the automotive domain: Challenges and solutions," *J. Syst. Softw.*, vol. 141, pp. 85–110, Jul. 2018, doi: 10.1016/j.jss.2018.03.060.
16. D. Omeiza, H. Webb, M. Jirotko, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10,142–10,162, Aug. 2022, doi: 10.1109/TITS.2021.3122865.
17. B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," in *Proc. Workshop Fairness Accountability Transparency Ethics Comput. Vis. (CVPR)*, 2019. [Online]. Available: <https://arxiv.org/pdf/1902.11097.pdf>
18. M. Brandao, "Age and gender bias in pedestrian detection algorithms," in *Proc. Workshop on Fairness Accountability Transparency Ethics Comput. Vis. (CVPR)*, 2019, pp. 1–4.
19. Z. Wadud, D. MacKenzie, and P. Leiby, "Help or hindrance? The travel, energy and carbon impacts of highly automated vehicles," *Transp. Res. A, Policy Pract.*, vol. 86, pp. 1–18, Apr. 2016, doi: 10.1016/j.tra.2015.12.001.
20. S. Tolan, A. Pesole, F. Martínez-Plumed, E. Fernández-Macias, J. Hernández-Orallo, and E. Gómez, "Measuring the occupational impact of AI: Tasks, cognitive abilities and AI benchmarks," *J. Artif. Intell. Res.*, vol. 71, pp. 191–236, Sep. 2021, doi: 10.1613/jair.1.12647.