INTERNATIONAL STANDARD

ISO 39003

First edition
2023-07

Road traffic safety (RTS) — Guidance on ethical considerations relating to safety for autonomous vehicles

Sécurité routière — Recommandations relatives aux considérations éthiques en matière de sécurité pour les véhicules autonomes

Reference number
ISO 39003:2023(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 241, *Road traffic safety management systems*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

0.1 General

A long established and commonly held view is that the single most significant factor in road traffic safety are the actions of the driver. However, the road transport system is a complex socio-technical system which places high demands on humans to negotiate. Crashes occur since human beings, due to finite cognitive capacity and physiological limitations, cannot always cope with these demands. To increase road safety, the road transport system therefore should be designed to support the road user to cope with this complexity and to mitigate the effects of crashes when this is not possible.

Autonomous vehicles (AVs) have the potential to replace the human driver and to increase road safety by reducing the risk of human error in daily operations. This will probably take a long time and in the meantime the capabilities of technology and humans should be combined and integrated in such a way that the strengths of both are utilized efficiently. Important safety improvements were made over the last century, but automated driving technology provides new opportunities to improve safety even further.

AVs are unlikely to gain widespread acceptance until the travelling public feels assured of their safety and security, not only of passengers but also other vehicles and vulnerable road users. This includes a behaviour of the AV according to the desires and requirements of society. However, despite contemporary expectations and a very optimistic view on technology, humans have an extremely valuable faculty that machines will probably never possess—ethical decision making and judgement. For real traffic situations the human does not always have the prerequisites to take a rational decision based on ethics since the time frame is often too narrow. For that reason, many "decisions" are made instinctively without the possibility of making a well-reasoned and balanced ethical decision. Defining this element and imparting it on machines is critical for the success of AVs. This can only be achieved by ensuring that AVs are equipped with driving action policies that align with the general ethical beliefs, needs, and desires of humanity on a global level, subject to local specific nuances.

To achieve the imprint of global and local ethical considerations in AV design, there is a need for a framework of ethics involving the necessary stakeholders of different areas. To that objective's end, it is important to develop ethical standards for AV behaviour. While there are few standards available, or under development, that address the engineering and technological aspects of AVs, there are no International Standards that address aspects concerning the general topics of driving policy and ethical behaviour, which are also important. Driving policy means a general approach of how an AV makes a decision and performs manoeuvres. Ethical relevant behaviour represents positive or potentially negative impact on road users and especially the vulnerable ones as well as the public space at large.

The objective of this document is to lay out a framework for the development of a standard for ethical and societally accepted driving policy.

0.2 The concept of autonomous vehicles

AVs have the objective to substitute driver, including tasks, decisions, and responsibilities. Hence the driver behaviour model proposed by Michon (1985) and applied to AVs in Reference [26] (Figure 1) can be a start to design and operate AVs. The driving task consists of three levels: the strategic level concerned with the higher-level trip goals (e.g. route choice), the tactical level concerned with the manoeuvring decisions, including negotiations and interactions with other vehicles, and operational level concerned with the execution of these tactical and operational behaviours at the level of vehicle control. There is no strictly hierarchical relationship among these levels.

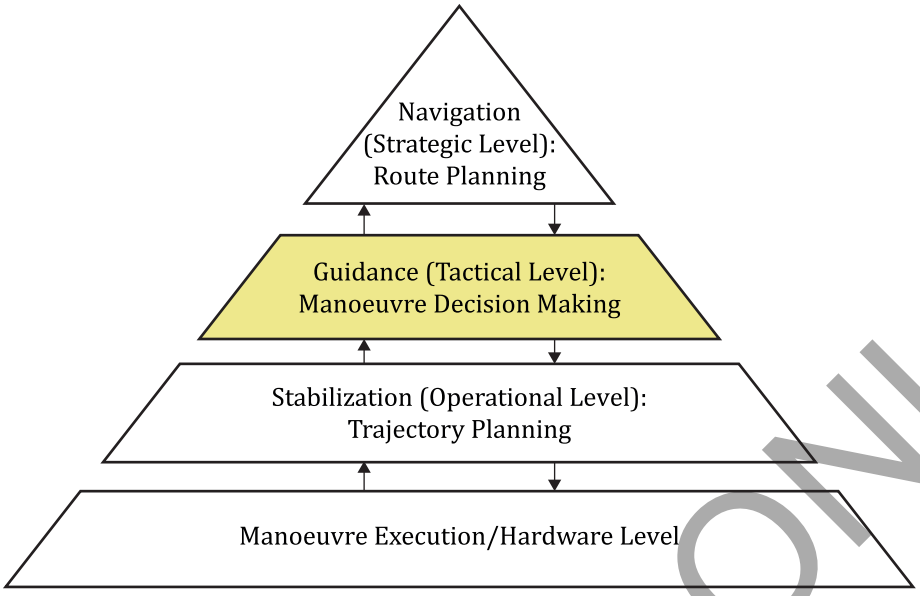


Figure 1 — Hierarchical model of driving task (source: Reference [26])

Beside this consideration which has a timely sequence, the AV will ideally be designed thoroughly to ensure safe and secure operation, and the decision base for the desired behaviour established. Therefore, all necessary information would be available, and the situation will be captured and understood while different perspectives will support the final decision.

Figure 1 is a generic decomposition of how an AV works. Figure 2 is a more focused and detailed description of the decision making module (where the ethical considerations resides) Hence a functional decomposition into the following six layers (Reference [27]) (Figure 2) can provide the first step into solution space.

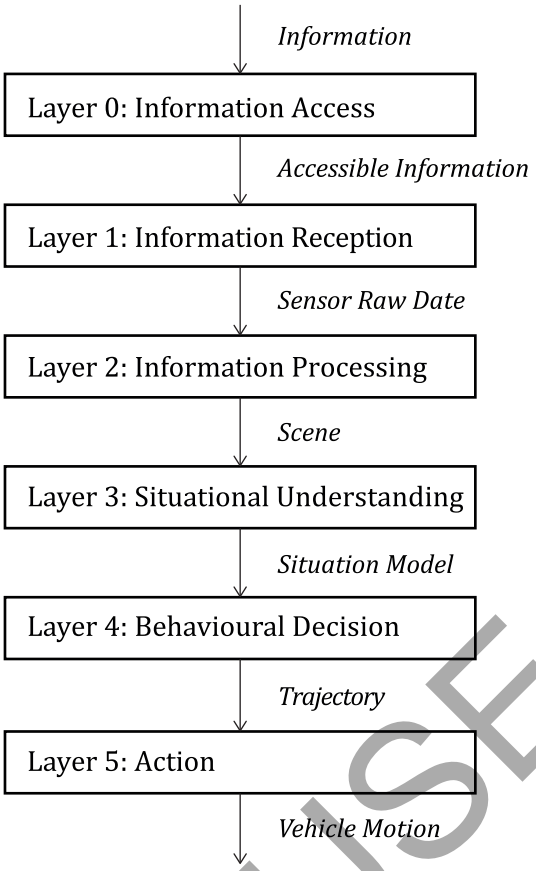


Figure 2 — Decomposition layers (source: Reference [27])

— Information access

Information available and accessible.

— Information reception

All necessary information can be captured.

— Information processing

Captured information contain all required classification and identification for further process.

— Situational understanding

Based on information, situation is captured and understood.

— Behavioural decision

Based on designed or trained situational awareness, the desired behaviour is chosen.

— Action

Vehicle transforms whished behaviour into action.

Within this framework, technical realization can differ while the focus of the situation decision will be comprehensible, accountable, and comparable among different designs. A thorough development will benefit transparency for action and to take decisions. The work described in this document focuses primarily on Layer 4 – behavioural decision.

0.3 The concept of ethical considerations for autonomous vehicles

Philosophy helps people question, understand and make sense of the world so they can act properly in it. This means that decisions and actions are not only intrinsically fair, but that they are also performed in a way that is balanced with respect with other's needs, the societal needs and humanistic values, as well as with respect to the physical world. It is widely accepted that philosophy can be divided into theoretical philosophy and practical philosophy. Theoretical philosophy consists of subdisciplines such as ontology, epistemology, logic, and philosophy of mind, amongst others. Practical philosophy consists of subdisciplines like ethics, political and social philosophy, and aesthetics.

Within ethics there are three main branches: normative ethics, applied ethics as well as meta-ethics. Normative ethics is the study of ethical action and determining standards for decision making and conduct (e.g. consequentialism, deontology, virtue ethics, see Annex A). Applied ethics is the application of standards of ethics to real life situations (e.g. biomedical ethics, AI ethics, political ethics, see Annex B). In this document, ethics refers to both normative and applied ethics but not to meta-ethics (which is more concerned with the meaning of ethical concepts such as 'good' or 'bad' or the nature of moral judgments. Applied ethics is a flexible and practical way to address ethical considerations in the development of new technologies because an applied ethics field, such as AV ethics, can borrow from more than one normative school of ethics. Therefore, the framework offered in this document, although based on a principle-based approach, can use ideas from deontology and virtue ethics, for example, to help solve problems. This is a balanced approach; it does not condone one type of normative ethics over another, it offers a range of perspectives that will help the designer/developer in choosing the best (or better) decision possible for specific situations.

In summary, ethics is the study of how to choose to act in situations in order to make good and rationally justifiable decisions. To make decisions, it is necessary to clarify what matters and to understand what is considered good, bad, right and wrong. Therefore, ethics may be viewed as a tool that helps us create the difference between a "good" decision and a "bad" one, but also to justify this decision on rational and intersubjectively acceptable grounds. This is of great importance in the development of AVs because choices made during the design and development of AV systems determine its "driving behaviour" and how it caters to its passengers and interacts with other road users (e.g. vulnerable road users). "Driving behaviour" is what was designed for and programmed into the machine, "conduct" is what actually transpires as a consequence of applying the driving behaviour to the real world.

This document offers a particular framework for AV ethics which is intended to support the practical integration of ethics into the AV development process. The framework suggested here builds upon Socrates conception of the hierarchical nature of philosophy for practical use which has three levels (see Figure 3). The base level is the conduct that transpires as a consequence of decision and actions (behaviour), the second level is how these behaviours are governed. Namely what kind of policy and arrangements are in place to make decisions. The top level is purpose, which behaviour and conduct will normally align with, in order to achieve the goals and objectives of the entire endeavour. Finally, knowledge is the "tool" for making sure that the purpose is reasonable and balanced and that the governance and behaviour/conduct are indeed feasible and appropriate.

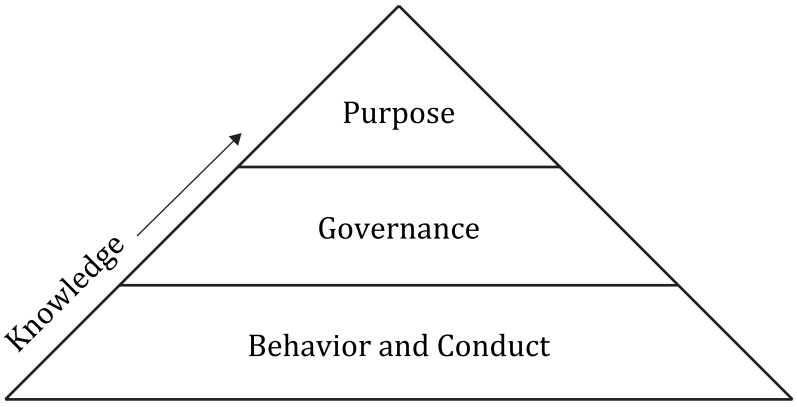


Figure 3 — Socrates' hierarchical nature of philosophy

Figure 4 is a graphical depiction of the framework advocated here for AV ethics. It shares similarities with other ethical approaches for emerging technologies, it is also unique as it has been created specifically for the AV context. Essentially, the AV ethics framework provides guidance for reflective and critical decision making and is composed of four main elements: purpose, values, principles, and process. The framework approach provides structure and guidance, yet it is flexible enough to accommodate multicultural perspectives.

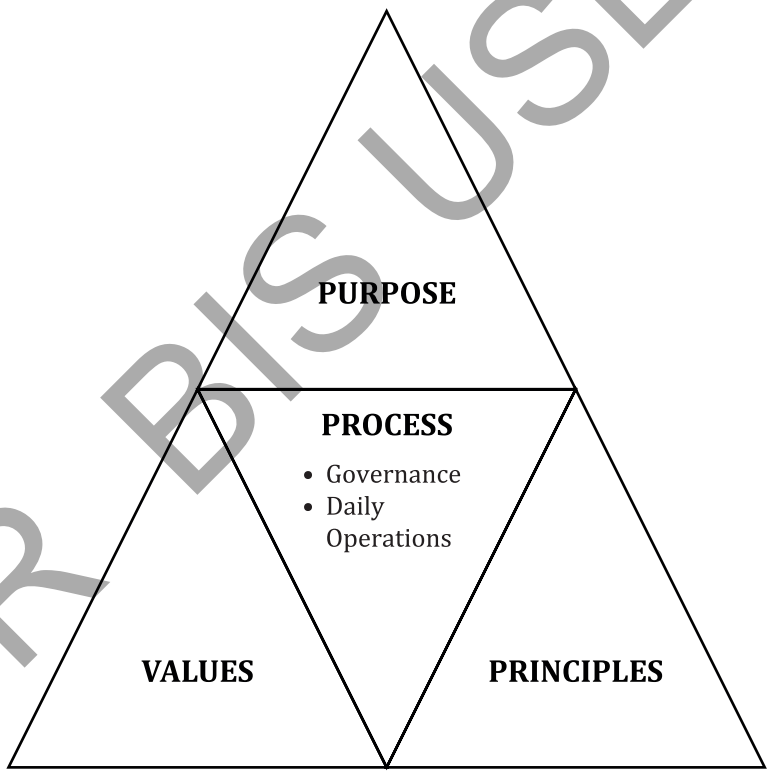


Figure 4 — The AV ethics framework for the integration of ethical considerations into the design and development of AVs

The purpose element refers to the overall goal of integrating ethics into AV design and development. The purpose is the "why" aspect of the framework or the reason for the existence of this document. Purpose offers high level (or meta) guidance when decision making, i.e. will a particular decision lead to an increase in road traffic safety or not. As this is an international standard, the values element provides guidance on what is important to the world's population. Therefore, universal values such as the UN universal values (e.g. human dignity) are recommended in this framework. The principles element gives more detailed guidance on decision making by identifying the boundaries of a good

decision and a bad one as well as directing the attention to specific areas (fairness in AV behaviour and conduct in the context of other vehicles). The process element refers to two levels of process activities, governance, and daily operations, which are required to integrate ethics into AV development. Table 1 summarises these four elements.

Specific examples for the values, principles, and process elements are provided in the AV ethics framework section. The purpose is not an example but a strong recommendation as it contributes to the overall goal of the ISO traffic management safety series of standards. The values, principles and processes presented here are relevant and universally acceptable examples but may be changed according to the requirements and needs of specific countries, societies and organizations. The users of this document may decide on which values, principles and processes they will use for their work. Finally, it is acknowledged that an ethics framework is only useful if it is adhered to by all involved and is well integrated into the AV system development process.

Table 1 — Framework elements and recommended designations

Purpose	Overarching goal: increase safety in road traffic systems for all traffic participants (ISO TC 241)
Values	Value set: UN universal values recommended
Principles	Principle set: artificial intelligence for people principles recommended (with other options listed)
Process	Governance (e.g. integration of AV ethics into existing governance measures)
	Daily operations (e.g. ethical evaluations, operationalization of principles)

Road traffic safety (RTS) — Guidance on ethical considerations relating to safety for autonomous vehicles

1 Scope

This document gives guidance on ethical considerations with regards to road traffic safety of autonomous vehicles (AVs).

It is applicable to vehicles in level 5 mode according to SAE J3016 in 2022, as part of its report.

This document does not apply to the technical method used to control the decision-making process, nor does it give any guidance on the desired outcomes of those decisions; it gives guidance on ethical aspects for consideration in the design of decision-making process.

This document does not set requirements for the outcomes of ethical decisions, nor does it offer guidance on methodology. It only details aspects of the behaviour of AVs for which considerations may be made by the designer/manufacturer to ensure that key aspects are not overlooked or disregarded.

This document does not offer the technical precision to prescribe the required controls but would, rather, offer a set of "protocol guidelines" that all decision makers regarding automated driving could choose to self-certify against to assure that the desired necessary ethical considerations were addressed during design and effectively controlled.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 39001, *Road traffic safety (RTS) management systems — Requirements with guidance for use*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 39001 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
autonomous vehicle
AV
vehicle equipped with an *automated driving system (ADS)* (3.2)

3.2
automated driving system
ADS
system that allows a vehicle to fulfil the requirements for a SAE J3016 level 5 vehicle

3.3
conventional vehicle
vehicle designed to be operated by a conventional driver during part or the entirety of every trip

3.4
AV ethics
branch of applied ethics specific to the ethics of *autonomous vehicles (AVs)* (3.1)

Note 1 to entry: AV ethics is concerned with the ethical decision-making process of the designers and developers of AVs while creating and programming AVs. Further, it helps designers and developers of AVs define the conditions of a good choice and determine which of the options available is the most appropriate one.

3.5
ego vehicle
subject-connected and/or automated vehicle, the behaviour of which is of primary interest in testing, trialling or operational scenarios

3.6
driving action policy
algorithmic statement that is evaluated TRUE before an action is taken

Note 1 to entry: In essence, the driving action policy constrains the planner from executing a (unethical) manoeuvre. Many people could consider these as "driving action policies".

4 External factors affecting autonomous vehicle safety

4.1 General

The following are seven categories of external factors relevant to AV safety:

a) other fully AVs (SAE level 5);

b) other highly or partially AVs (SAE level 4 and below);

c) other conventional vehicles;

d) other mechanical transportation;

e) vulnerable road users;

f) animals;

f) road traffic environment.

NOTE For definition of categories a) to c), see SAE J3016.

4.2 The road environment

In addition to road users and vehicle factors, road and environment conditions are also considered as contributing factors to crashes. Hence, it is important for AVs to be aware of environmental factors such as road and topographical conditions and road type (e.g. single-or dual carriageway) throughout the entire journey route.

Attention should be given, but not limited to the following items.

a) Road conditions:

- road surface (e.g. potholes, slippery, greasy),
- geometric features (e.g. sag curves, crest curves, lane width), and,
- road construction work.

b) Road furniture along the route (e.g. type of guardrail used, road signage, road markings).

c) Road environment:

- topographical condition (e.g. flat, undulating, hilly, mountainous),
- signage (e.g. inadequate, confusing or blocked road signages),
- lighting,
- haze,
- thick fog,
- weather condition (e.g. snow, ice, heavy rain, flood, landslide, crosswind),
- animal crossing, and,
- parked vehicles and potential impact on sightlines.

d) Traffic volume and condition (e.g. during peak hours, festive seasons).

The nature and magnitude of traffic plying on the road plays a significant role and gives rise to a number of problems if traffic is not properly regulated and controlled due to non-adherence to the traffic rules which are common features in many developing countries. Therefore, an understanding of traffic characteristics is a key component to build into autonomous driving systems. If traffic is allowed to mix with non-motorized traffic (NMT), it would be difficult to ensure that there would not be any likelihood for occurrence of road crashes. There are many complex issues like the presence of NMT in the traffic stream to be addressed in a holistic manner in context of ethic consideration for AVs.

5 Interested parties in AV design and operations

5.1 General

There are four categories of interested parties involved in the design and operations of AVs. Each one is examined below.

5.2 Producers – Manufacturers, designers and their suppliers

Manufacturers of AVs designing the vehicle carry responsibility toward the users of the vehicle as well as and towards other road users (vulnerable road users, bystanders, other drivers) for the product and toward the regulator and operational agencies such as the Department of Motor Vehicles (see C.1 for a discussion of the concept of responsibility and accountability in AV design.). Manufacturers share this responsibility with their suppliers. This document, which discusses ethical considerations in the design of the AV behaviour, provides guidance to manufacturers and their suppliers as to how they develop ethical relevant components (e.g. vehicle behaviour, interaction with other road users) and how to document the decision process and decisions accordingly. In particular, the focus is on ethical considerations in designing vehicle behaviours and decision making.

5.3 Distribution chain – Distributors, sellers

The role of a distributors is to facilitate the bulk transfer of completed vehicles from manufacturers to sellers. They should address the safety ethical considerations and the guidance within this document and ensure that no part of the distribution process can compromise any safety features or design of the vehicle, as manufactured.

The role of a seller is to facilitate the sale of vehicles to users. They should be responsible for ensuring that all vehicles they sell address the guidance within this document and that the prospective purchasers are made aware of the relevance of meeting the guidelines in this document (e.g. education about AV technology and its societal impact).

5.4 Purchasers, owners and operators

Transport authorities generally attempt to optimize the effectiveness of the transport system, maintain the infrastructure, to improve traffic safety and enhance sustainable development, both for passenger and freight traffic. In addition, digitalization and the transformation of transport with connected AVs emerges as an important area to achieve the long-term policy goals set by the authorities, in the changing environment.

If producers and users work in close coordination, it would be an opportunity of AVs for gaining better knowledge on the network condition and implementation of more sophisticated network management strategies. It is contemplated that connected AVs will be in a process to exchange/share a lot more information about their environment than non-connected and non-autonomous vehicles. This would provide an impetus to the users to make use of more traffic data to better adjust their network operation.

5.5 Government agencies and other interested parties

Road safety is a shared responsibility among stakeholders. Government agencies, private sectors and non-profit organizations significantly influence the automotive market and should ensure that all vehicles made available are safe and fit for purpose and operated safely.

Government represents interests of all citizens and will always play an important role in effectively communicating with the producer, supply chain, operators, end users and other related industries, interested parties and the public in bringing the best possible positive impact on safety.

In order to ensure the ethical behaviour of AVs, the government should consider including ethical consideration and assessment in regulating safety requirements.

6 Governance, assessment and evaluation

6.1 General

Organizations claiming compliance to this document should undertake an AV ethics assessment where appropriate and desirable, the outputs of the assessment can be used as a demonstration of AV ethical compliance to interested external entities.

It is recommended that, if an internal authority is undertaking the assessment, the level of independence from the organization commissioning the assessment and the organization delivering the automotive product or service is clearly explained. It is generally accepted that in using an independent body, the greater the level of independence the higher the degree of objectivity achieved.

The assessment approach should be structured and systematic. It is envisaged that a meaningful report will be produced as an output of the assessment activity.

6.2 Ethical reference for assessment

It should be clearly stated before the work commences what ethical framework is being used to conduct the assessment. The assessment framework should be stated in any reports produced. It is possible that frameworks used in the development of the vehicles may be different to those used in the ethical assessment. Different organizations could have different ethical perspectives. It is not envisaged that different ethical starting points are necessarily contradictory or conflicting. It is recommended that this process be documented with consideration given to checklists or other methodology.

6.3 Additional standards

Where industry-specific, national and international standards are used in the undertaking of the assessment these should be stated. The scope and degree of utilisation of each standard should be indicated in any reports produced.

6.4 General

The scope could be an ethical evaluation of the processes contributing to a specific product or service, or, the ethical processes of the organization in general. The suggested areas of ethical assessment are as follows:

— higher organizational level;

— development organization level;

— specific development and implementation processes;

— post implementation checking against ethical criteria.

It is envisaged that the assessment scope will cover one or more of these areas. It is also envisaged that the scope may cover additional areas. Whatever the case, it should be clear what the scope limits are, for example, what elements of the organization, product, or service are being evaluated.

6.4.1 Higher organizational level

This organizational level is envisaged as the overall business. The assessment should evaluate the general ethical policy. The aim here is to determine the consistency of the ethical approach flowing from the higher organization to the product development. The assessment should include, but not be limited to, an evaluation of the following areas:

— any company-specific or company adopted ethics framework, its applicability to AV ethics, and its distribution to the internal organizations involved in the development of AVs;

— the representation and championing of AV ethics on the company board and the degree of involvement with AV ethics at the board level;

— the provision for and support of AV ethics within company policy including:

  — the provision and effectiveness of AV ethics training;

  — the provision, status, and influence of AV ethics specialists;

— the facilitation of AV ethics through project budgets, planning, scheduling and human resources;

— the requirements placed on suppliers providing equipment or software that facilitates vehicle autonomy, and the means of acceptance of these items. Measures that prevent the introduction of hazards owing to the lack of ethical consideration. This includes development and support tools.

6.4.2 Development organizational level

This organizational level is envisaged as the development suborganization that is responsible for designing and implementing high-level vehicle behaviour. That is behaviour involving the control of the vehicle within traffic and can be considered the tactical behaviour. The assessment should include, but not be limited to, an evaluation of the following areas.

— Non-ethics specialists: methods to ensure those involved in the design and implementation of high-level vehicle behaviour have the appropriate understanding, skills and training in AV ethics.

— Diversity awareness: methods to reduce bias leading to ethical ‘blind spots’, from the selection of personnel performing key roles.

— Ethics specialists: management processes to ensure sufficient ethics specialists with appropriate authority to provide subject matter support and understanding to the non-specialists.

— Working environment: policies to ensure that the working environment, including working practices, are conducive to producing AV ethics developments; confirming the effectiveness of these policies on the actual working environment including levels of compliance.

— Organizational structure: hierarchy designed and appointments tailored to ensure that individuals within the organization have adequate support and defined routes of redress, ensuring that individuals can perform at their best in the context of AV ethics.

— Openness and focus: practices to ensure a working culture where ethical concerns and matters causing ideological conflict are addressed openly and objectively without adverse impact to the individual. A working culture that encourages ethical thinking and AV ethics to complement any focus on product delivery.

— Outsourcing: policies and practices to ensure that elements of associated work undertaken outside the development organization are operating under the same ethical considerations. Where this is not possible, it is important to ensure that there are measures to prevent the introduction of ethical risks via this route.

NOTE 1 It is important to be aware that ethical bias can be introduced because of the lack of diversity in the developmental organization. A particular profession, for example, can be a de facto preserve of individuals with a specific world view. This can be incompatible with the ethical interests of the wider society.

NOTE 2 It is important to be aware that the organizational customary culture and working practices can have an unwanted impact on the implementation of AV ethics.

6.4.3 Specific development and implementation processes

This subclause deals with the processes that directly contribute to the operating behaviour of the vehicle under development. At this level the developer is making decisions about and creating elements of the vehicle behaviour that affect AV ethics.

The assessment should include, but not be limited to, an evaluation of the following areas:

— vehicle behavioural design:
  — the level of AV ethical engagement in the design of vehicle behaviour;
  — the level of AV ethical consideration in the design verification activities;
  — confirmation of ethical considerations modelled or otherwise evaluated prior to implementation;
— behavioural system architecture:
  — the level of system evaluation in terms of AV ethics;
  — behaviour explicitly devoted to AV ethics, for example, there may be distinct ethical elements or AV ethics are embedded within other behavioural elements;
  — ensuring the approach has been evaluated and there a clear rationale for the approach;
— implementation planning:
  — the level and nature of ethical engagement in the planning of implementation activities;
  — the level and nature of communication of ethical considerations in briefings and meetings;
  — the level and nature awareness and dialog;
  — the level and nature of specific ethical activities;
— implementation:
  — at the point of implementation, the level and scope of discretion given to the individual or team in ethical decisions;
  — the level of ethical guidance material provided to the implementer, including codes of practice;

  - the level and nature of supervision by ethical specialists (6.4.2). At the point of implementation, the cognisance of ethical considerations provided in this document;

- updates:
  - the level and nature of AV ethical engagement post development;
  - the process of making in-service changes to vehicle behaviour and associated review and approval by ethical specialists.

6.4.4 Post implementation checking against ethical criteria

It is envisaged that approaches to AV ethics may include post implementation activities. It is also envisaged that these activities may serve to provide mitigation for lower levels of engagement at the design and implementation stages. Checking for AV ethical concerns may occur throughout the lifecycle to ensure a sustained focus during the various stages of development. Activities to ensure AV ethical consideration of the vehicle or product, should be undertaken prior to commercial delivery and public use. Checking for objective ethical fitness should be undertaken concurrently with other acceptance activities. The assessment of the AV ethical checking activities should include, but not limited to, an evaluation of the following areas.

- Coverage: the adequacy of checking activities in ensuring AV ethical considerations are addressed, the completeness of process, the compliance to process, the degree of objectivity in the checking process, for example, what is the organizational independence of the checking authority from the project management?
- Management of verification failures: the adequacy of the control and addressing of process failures.
- Ethical review: the adequacy of the objective ethical checking process in ensuring the vehicle or product is fit in terms of AV ethics for operation in its designated environment. The completeness of process. The identification and mitigation of failures. The appropriate authority and scope allocated to the checking personnel, for example, any authority to specify changes prior to product release.

6.5 Conducting the assessment

The assessment methodology adopted should include the following elements:

- an assessment plan and schedule;
- a deviations and action log;
- an assessment mandate with roles and responsibilities;
- evaluation methods and guidelines;
- an assessment recording process;
- review meetings at discrete developmental stages of the product or service;
- a report of AV ethical assessment;
- activities in support of the report:
  - presentation;
  - workshop with affected parties;
  - training for affected parties where appropriate.

In reviewing and evaluating ethical processes it is recommended the assessment considers the following:

— evaluation reports;

— related corporate policies and mission statements;

— traceability of AV ethical requirements through to implementation and verification;

— control of deviations and concerns;

— accountability of decisions made;

— credibility of AV ethical verification process and results;

— concerns and mitigations;

— management of complexity;

— integration of behavioural elements from different sources;

— degree of compliance to ethical frameworks and standards.

6.6 Expression of results and conclusion

It is important that the level of compliance is clearly understood by the report audience. Areas of weakness should be explained with a justifiable rationale. There is no recommendation for how results are to be expressed, this may be purely narratively, using numerical indicators, graphical methods of indication, a combination of methods, or some other method. Any use of indicators should be meaningful, and the context explained. The results should be useful and understandable to the report audience. A timeline for the delivery of the results and conclusions should also be defined, in order to ensure that they are communicated in a timely fashion. The results should be useful and understandable to the report audience.

The conclusion should be clear about weaknesses found during the assessment. Where personal judgement is used this should be indicated. Strengths should be given the correct weighting. It is recommended that the overall tone is one of encouragement.

Recommendations should be clearly differentiated. This is to enable a systematic response in addressing the findings.

7 Operationalization of ethics - discussion on values and ethics to consider

7.1 General

This section builds upon the elements of the AV design and ethical considerations discussed in 0.2 and 0.3 respectively. The objective of this subclause is to provide examples in three areas:

a) provision of examples for the values and principles elements of the AV ethics framework;

b) ethical considerations during functional implementation;

c) ethical considerations applied to operational situations.

7.2 Ethical framework for the design of AVs (driving action policies and ethical design)

This subclause discusses the integration of the general ethics framework discussed earlier (and shown in Figure 2) in the context of one aspect of AV design, namely its intended behaviour in terms of the "decision making" processes and selection of action. The actual conduct, or that which transpires as a consequence of the decision and resulting behaviour, is beyond the scope of this document.

Use of the framework outlined in this document will help to improve road traffic safety by making AV behaviours inspectable, internally coherent, and well matched (correspondence) to the principles and value of humanity in general and different social norms that may somewhat vary between cultures. Examples being prohibiting infringement and/or pre-emptive behaviour toward other vehicles in a queue, and the development of a standard protocol for vehicle behaviour at an unsignalized intersection. This fits into the larger purpose of ISO, which is to offer guidance on improving engineering system via international efforts.

The suggested AV ethics framework discussed earlier in 0.3 (Figure 4) is now expanded to apply directly to the design and ethical considerations of AV behaviour and conduct. The expanded framework consists of five key elements – purpose, ideals, governance, maxims and driving action policies that generate behaviours, and actual conduct. This subclause discusses the framework to illustrate a suggested process and a formal method for producing AV behaviours that are based on driving action policies. This framework also provides guidance as to what an AV should do in ambiguous situations when there are no specific driving action policies. In that case, the AV is expected to employ the principles and even values and definitely make sure that they are not violated per the selected action. The framework is also useful when there is a design flaw or an unanticipated situation (Reference [5]) Figure 5 is the expanded framework that includes all the necessary content for the task for defining and evaluating driving action policies.

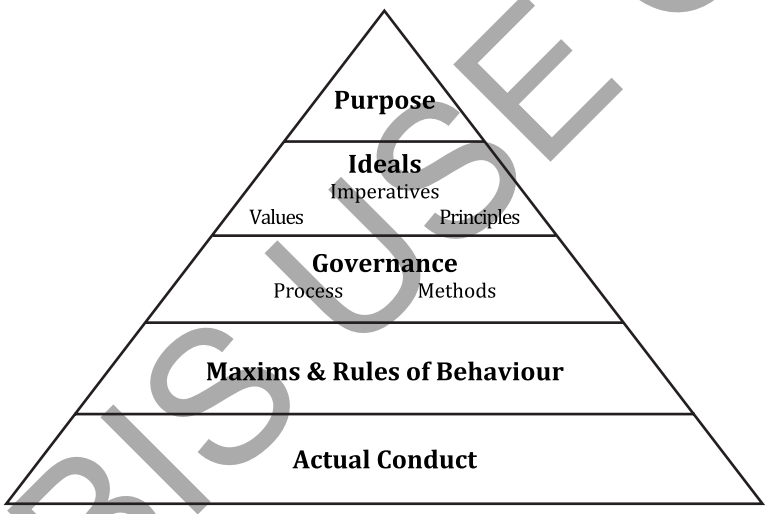


Figure 5 — Modified graph that focuses on implementation of ethical considerations

7.2.1 Purpose

The purpose of the AV ethics framework in this document is given to increase safety in road traffic systems. The purpose element may be adjusted according to the overall goals of the AV developers and relevant stakeholders. The following purposes are higher-level examples of alternative purposes, which may be combined with the goal of increasing road traffic safety.

a) **Moving people to achieve self-realization:**

The purpose of transportation, from this document 's perspective, is to move people or commodities safely, effectively/efficiently, and in a humanistic manner, such that they can achieve their highest individual goals and aspirations for the betterment of society as a whole. Humanity's desire for positive self-realization should be supported by a judicious transportation system that provides "access to resources" to **all**.

b) **Reducing impact**

As such, vehicle design, manufacturing and actual conduct in the public space should be developed and used in such a way to reduce the negative impact on the environment and other road users with considering positive benefits of self-realization.

c) **Enhancing the public space**

Another important purpose for vehicles in general and AVs in particular is to enhance the quality of the public space (shared by, e.g. other people, cars, cyclists and consisting of roads and infrastructure) by proper conduct as well as supporting other road users (e.g. sharing information and assistance) on their respective paths.

7.2.2 Values

Values are independent of purpose, because they are universal and can apply to any endeavour (not just transportation). Values provide a general statement of beliefs. This document recommends a selection of values that are humanistic in nature, judicious, and proper to almost any human endeavour. The consequence of these values for vehicle design and engineering and resultant vehicle behaviour is that they provide context and set as boundaries on design and development of AV systems.

This document recommends the employment of a cross-cultural approach through the internationally recognized United Nation's (UN) universal values of peace, freedom, social progress, equal rights and human dignity.

This document adopts the six UN core values (peace, security, freedom, social progress, equal rights, human dignity) for the purpose. "Peace," whilst important, is not considered relevant for public transportation and two additional values, "privacy and the right for intimacy" and "sustainability" are added. They are modified and ranked as follows:

Value V1 Human dignity

Value V2 Freedom of mobility and making decisions

Value V3 Equity and fairness

Value V4 Social progress

Value V5 Security and safety

Value V6 Privacy, intimacy and human decision-making autonomy

Value V7 Sustainability and proper balance of human-technology-and environmental homeostasis

7.2.3 Principles

Purpose and values offer general ethical guidance but their direct application to daily operations is challenging. Principles provide more detailed guidance in how to implement the purpose and values in the AV development process. Specifically, they are guardrails that designers and developers can use to guide their decision-making process with regard to the driving behaviour of AVs. Principles are applied along the four stages discussed earlier in 0.2: the decision-making logic, the implemented and intended behaviour, the actual conduct of the vehicle on the road, and how this conduct is perceived and observed by other road users.

Bioethics is probably the first sector which considered ethical principles in the context of healthcare and treatment as well as genetic engineering. Since their principles are mature and have endured criticism and modification, they are adopted here. Another sector that has considered principles is AI, and the document adopts an important principle (explicability) from there too. These principles are:

— non-maleficence: first, do no harm ("Primum non nocere");

— beneficence: try to do good;

— justice: every action must relate to others in an equitable and fair manner;

— autonomy: the ability of an individual to make well-informed decisions without coercion and manipulation;

— explicability: no information manipulation and lack of accountability.

These and additional principles, relevant to the vehicle and transportation sector, are detailed below.

Principle P1 Transitoriness, reverence and sustainability ("people are only guests on earth and not masters")

The impermanence nature of being is the keystone principle of the framework. It encourages people to think that in every endeavour and action that they take, they must understand that they are only guests on this planet and should act accordingly. The vehicle should also hold in reverence other road users, the infrastructure, and objects/structures on the road. The vehicle should act in the most sustainable manner possible and avoid unnecessary (carbon, disruption, landscape infringements, etc.) footprint on the environment. For example, a vehicle approaching a non-signalled pedestrian crossing, should slow down a bit even if there are no pedestrians in sight, just as a sign of reverence to the meaning of the crossing. Vehicles should reflect the principle that every action that they take has an impact on the Earth and all its inhabitants and behave accordingly on the road space itself and with respect to other road users.
(This principle supports values V3, V4, V5 and V7.)

Principle P2 Reciprocity and caring ("because first and foremost we are humans")

Vehicle behaviour should not only consider what is effective and efficient (for the network), but also what is considerate and palatable to fellow human beings as other animate or inanimate road users. For example, when two vehicles converge from opposing directions on a single lane road, both vehicles should veer slightly to the side (and drive on the gravel). For effectiveness's sake, it is clear to all that only one can veer to the side and let the other occupy the road.

This principle protects human beings from being degraded to a mere network element, which could happen if the primary goal of a traffic system is to optimize efficiency while ignoring established social contracts that support a mutually beneficial traffic environment. For example, a vehicle trying to get out of a small road and having to wait for 20 min because overall, it is much better for the whole traffic network that it will wait that long. However, the needs of the passengers in the vehicle have relevancy too and is it important to consider these when planning right of way behaviour in diverse situations. The submission **only** to technological imperatives is ethically questionable.

NOTE Rule 6 (German Ethics Commission): the introduction of more highly automated driving systems, especially with the option of automated collision prevention, can be socially and ethically mandated if they can unlock existing potential for damage limitation. Conversely, a statutorily imposed obligation to use fully automated transport systems or the causation of practical inescapability is ethically questionable if it entails submission to technological imperatives (prohibition on degrading the subject to a mere network element).

Principle P3 Cooperation and coordination ("it is not a competition")

Vehicle behaviour should aspire as much as possible to cooperate with other road users and when possible, to coordinate to further support the cooperation. Attempts to compete, game, take advantage of other road users do not lead to a fair transportation system. While at first it may seem that self-gain leads to efficiency, at least from the ego vehicle perspective, the overall efficiency is reduced, and self-gain can always be gamed by a more aggressive road user.

Principle P4 Human autonomy ("freedom should not be taken away")

Autonomy requires that the individuals have the right to decide for themselves whether they want to be subjected to direct or indirect decisions of a technology, that they preserve the decision-making power of being made aware of risks and benefits of a technology. The principle of autonomy implies that the users are free from coercion, that the technology does not impose its decisions. Vehicle automation reduces the options for action for the AV passengers. Autonomy requires that the choice of action left to the AV passenger(s) be considered from the

passenger's perspective and the enablement of the passenger to exercise her/his freedom of decision by providing her/him with meaningful options is balanced with being able to unload the driving of the vehicle. Designing with the principle of autonomy considers the legitimate needs of the passenger (and other humans in and around the vehicle).

Principle P5 Equity and fairness ("what someone does to their neighbour, they do to themselves")

The vehicle should act in such a way that it treats itself with respect and that a similar treatment is given to other road users. That also in the design of the new transportation system one agent is not made favourable and entitled over others. In other words, an AV should be designed in a way that prevents discrimination against certain groups because of personal characteristics**.** The following principle begins to apply "giving to others". (Supports V3.)

Principle P6 Non-maleficence ("first, do no harm" -- "primum non nocere")

In every action that the vehicle does there could be the implication of potential harm to others, either immediately or over time and this could include potential harm to the environment. In practice, what this principle says is that when considering any behaviour, say a given manoeuvre, it is better to do nothing than to perform a manoeuvre that has a potential to harm another road user. For example, if the vehicle planning algorithm is evaluating the manoeuvre of passing a vehicle in the front, and there are some uncertainties about this manoeuvre – the best thing to do is to keep the lane even at the cost of a slower speed and longer travel time.
Another example is a situation when a tailgating vehicle is demanding, by honking and physical movement, that the ego vehicle risk entering the junction (e.g. four-way stop intersection) even. The best thing to do is just to pull to the side and "get out of the game".
(Supports V5.)

Principle P7 Non transgression and no coercion of others ("one must not take what is not designated as theirs")

The vehicle should not impinge on another road user's space (e.g. slot) or even be perceived, in its actual conduct, as doing so. This rules out any "cut-ins" that people do to take someone else's slot in a queue (link to P1 – Reverence and transitoriness).

In situations where behaviour of the vehicle may be legal, but the conduct is a transgression or may be perceived as such by other road users, then this conduct should be avoided or minimized (under the assumption that what a person could do is not necessarily what that person should do). This also includes acts that while at the time seem at first non-transgressive, in retrospect it is found that they were. The vehicle should **not** try to manipulate its way for the sake of the driver/passenger benefit at the expense of others. In the same token, a system, AV, or any transportation app, should not manipulate the driver/passenger for the sake of the vehicle, producer's, or fleet company needs (e.g. revenue stream). (Supports V2, V3, and V5.)

Principle P8 Respect for intimacy ("one must not enter into a space where one is not welcomed")

The vehicle should avoid "entry" into others' sacred spaces, where surveillance and tracking mechanisms (e.g. security cameras, personal photographs, GPS) and the ability to monitor people's (digital) content may lead to manipulation of people's desires and to use their data for unwanted purposes. In the case of AVs, as well as modern cars, it is possible to track one's whereabouts and predict some of their desires and needs. It is also possible to manipulate consumers during their most vulnerable states if the vehicle applies simple tracking and photographing techniques, for the purpose of monetizing this information.

Information retained by the AV should be restricted to only telematic data, essential for postcrash investigation and analysis for improvement of the technology in order to increase traffic safety.

Principle P9 Beneficence ("try to do good")

The vehicle should try, in its behaviour, to support other vehicles and the flow of traffic – even at the expense of its own progress and utility. If all vehicles give the way to others, that may make transportation much better than today.

Principle P10 Explicability to occupants and other road users ("what is not made transparent is concealed")

The principle here is that the vehicle be made transparent and intelligible to humans. It should also be delivered to them in an explicable manner. The vehicle and data stream are designed in a way that behaviour and actual conduct can be accounted for (indicate, for example, the vehicle's principles/logic, belief state, decision sequence, end goal). The principle of explicability demands a theoretical framework of abstraction. That is, what should be provided and what should not. Naturally, such abstraction should not be malicious or include any means of user manipulation for self-gain as discussed in P8 (intimacy).
Supports V1 and V2 non-manipulation – capitalistic surveillance. Understanding what the AV will do and what it will expect from other road users.

Principle P11 Justice and responsibility ("justice is the goal; everyone has responsibility for the world")

What seals the list of principles is justice. The ideas that the way a vehicle behaves should to be fair not solely for its own "benefit," but also for the good of others (e.g. social good and social progress) and that should be taken into account when a route is planned, the concept of justice must be expanded in this context into personal (customer) and fleet (corporate) justice. The first is concerned with the proper and sustainable-based realization of personal resources to achieve self-realization (see the purpose section). The second is concerned with large scale utilization of resources (e.g. fossil fuels and environmental damage) at the expense of the ecology. In other words, the now imperative concept of environmental and climatic justice. The implementation of justice, on a personal, corporate, and governmental dimension, can only come about from assimilation of the concept of innate responsibility towards other human beings and the environment. Of taking actions from a place of understanding one's place and duty towards others and the ecology, it is important that the vehicle acts in a fair way; not only optimizing its own goals at the expense of others.
(Supports V5.)

These principles can also be used to evaluate driving behaviours as well as support the development process. The following subclause will address the use of maxims, which support the further operationalization of principles.

7.2.4 Methods for construction and evaluation of maxims

In this subclause, the concept of a maxim is introduced as a means to translate principles into more tangible guidelines.

It is necessary to adopt a formal approach that will allow development of vehicle behaviours that are intrinsically ethical and are (i) reflective of humanity's shared consciousness and belief set, that they (ii) correspond to the values and principles detailed in 7.2.3. In addition, there must be (iii) coherence between the behaviours, such that one vehicle behaviour does not contradict another vehicle behaviour. Correspondence and coherency determine the integrity of the system of values, principles and how they are embedded in the actual driving action policies that the vehicle executes (References [7] and [8]).

7.3 Background of maxims

For this document, maxims are used as general rules. Additionally, their evaluation method (verification/validation) is used as a way to deliberate AV behaviour and actual conduct. Maxims are pithy and memorable statements about the world and conduct. The formulation of maxims for this document is inspired by the work of Immanuel Kant[9]. See C.5 for a summary of Kant's approach to ethical decision making and a short explanation of his "categorical imperative" approach.

7.3.1 Maxim design and construction

Maxims can be applied not only for evaluation of human ethical behaviour. It is recommended that they be used in the design of decision making for the behaviour of an AV. As such, a maxim is established as a subjective principle of action and is instantiated as part of an agent's decision-making process.

The structure of a maxim has three main parts: (1) the action, or type of action; (2) the conditions and requirements under which it is to be done; (3) the end or purpose to be achieved by the action and/or the motive behind.

The following examples are critical in order to illustrate how maxims can be applied.

EXAMPLE 1 "Whenever anyone (2) needs (3) money and can get some by borrowing it on a false promise, then he/she will (1) borrow the money and promise to repay, even though he/she knows that he/she will not be able to repay." The action (1) is to "borrow money," from a fellow human being; the conditions under which the action is taken is when one is in need (2). The end, or purpose of the action (3), is to "get money"[9].

When this maxim is considered, it is found to not make any sense because it contradicts itself because it nullifies the concept of borrowing on the subject side and the concept of lending to another on the object side. When the maxim becomes a universal law, it demolishes the explicit social contract behind the act of borrowing and lending. When this contract is demolished, it also erodes, implicitly, the concept of reciprocity in human-to-human relations. Similar reasoning leads Kant to conclude that any maxim permitting theft or lying must be rejected.

EXAMPLE 2 Anyone may make it their maxim to (1) increase (3) their wealth (2) by any safe means. Here the action is to increase one's capacity (1). The condition (2) is at "any same means" (safe to the subject, that is). And the end goal of the maxim is monetary wealth (3). It means that they will do anything, as long as it does not jeopardize their personal safety, to get wealth. If they are the strongest, then every means is safe such that there are no restrictions on my means for achieving their goals (e.g. wealth). In contrast to EXAMPLE 1 with the lending/borrowing, this example is not self-contradictory such that it violates the ground on which it stands. If it is universally applied, it will demolish the social contract that sustains us all together and create a society where living is, according to Thomas Hobbes' memorable description, life outside society would be 'solitary, poor, nasty, brutish, and short'.

In the context of AVs, speeding and being aggressive to other road users (e.g. inching toward vulnerable road users or making aggressive moves) belongs to EXAMPLE 2. There is a third class of maxims that is relevant here. Specifically, those maxims that take advantage of equally strong road users who follow the rules to the benefit of the perpetrator. For example, consider the following: a driver makes it their maxim to (1) pass other cars on the right (2) every time that it is possible for them (3) to reach the destination in minimal time. Here the driver takes advantage of the fact that everyone else does not pass on the right so that they can accomplish their maxim.

The same applies to a maxim of being opportunistic in a queue: (1) getting in front or "game" another car ahead in the queue (2) when it is safe under the assumption that the other driver will be scared and yield (3) to maximize driving efficiency so as reach the destination in minimal time. The same applies to drivers who take advantage of others in a US four way stop intersection and those who turn on a red signal, under the assumption that opposing traffic will indeed stop on red.

7.3.2 Evaluations of maxims

The recommended evaluation process is inspired by Kant's method but extends it in the context of technology and the notion of reflection. Maxims are a step toward developing "driving action policies." Appropriate driving action policies (that will be consistent with the values and principles) will be derived from each maxim. The steps of the recommended evaluation process are the following.

a) A check of if the maxim is a self-contradictory attribute, of what the maxim builds on as necessary for its own sake (e.g. the lending example). For example, a maxim that will lead to a vehicle behaviour that usurps other vehicles' right of way, nullifies the notion of right of way in the sense that if everyone will do it, there will not be right of way.

b) A check of if the maxim takes advantage of others nominal and predictable actions to accomplish its own purpose. For example, a vehicle impinges on others' space under the assumption that they will not do the same to it. A vehicle that violates a red light does this on the assumption that other cars will not do the same.

c) A check of if the maxim, when it is applied universally, leads to a lawless way of using the roads and/or an environment that is not pleasant to participate in.

d) A check of if the maxim contradicts another maxim within the rest of the system of maxims.

e) A check of if there are any contradictions with the principles and values for AV defined earlier. If for example, a maxim violates the notion of autonomy (P4) or limits social progress (V4) then it may be necessary to change the maxim.

f) A careful analysis of the maxim to see what will happen if this maxim becomes a universal law that is obligatory to all and indeed covers all foreseeable situations. For example, considering the personal maxim "I ought to save that puppy from that oncoming truck", a universal law is generated from a maxim by applying it to the entire rational population. For example, "Every rational person must save puppies from oncoming trucks." if this method is applied on this maxim it is evident that while it does not self-contradict itself, or violates any principle, applying it to every rational person is probably a bit presumptuous and perhaps dangerous (how close and fast is the truck behind?). This is in conflict with the principles of autonomy and non-maleficence.

g) A check to see if the maxim is acceptable when and if the acting driver were to hypothetically "switch" sides with the other road users that will bear the consequences of the action. That is, if the acting driver becomes the recipient of the maxim's behaviour (see Reference [10]).

h) Reflection: does the behaviour designed into an AV and the consequential conduct which will be perceived by other road users, reflect the values of responsible and considerate humans and engineers, regardless of what others have defined as values? The behaviour being portrayed in the design, and the conduct that will be exhibited by the design, is reflective of human self-perception and ambition. This step is an example of the application of "Virtue ethics" to this process.

The method suggested herein recommends the use of these eight checks for every implementation of decision making in AV behaviour that has implications for the occupants of the vehicle, other road users around the vehicle, and humans in general. Initially, a maxim is developed for the class of situations under consideration that provides the general constraints for AV behaviour. Then, the maxim is checked for internal contradiction, taking advantage of others (nominal and rule following behaviour), for internal consistency with other maxims (coherency), contradiction with the system of values and principles, and it is attempted to assess the universalization of the maxim, and finally it is checked whether the maxim is faithful to our own understanding of proper conduct.

There are two more categorical imperative from Kant, both are important to this document. Kant's second categorical imperative instructs to "act as to treat humanity, whether in my own self or that in another, always as an end and never as a means only." This categorical imperative rules-out any mistreatment of another. For the purposes here it covers aspects that are more associated with the use of automation information, digital content, and AI algorithms. The point is that transportation consumers, whether riding in AVs or other modern cars, should not be subject for such manipulation that are basically technological methods (or "means") for others' monetizing and unscrupulous gains ("end"). It also considers others as members and entities that are worthy of values, which means that they must have autonomy and the right to choose that should not be compromised by anyone of power, physical advantage, or emotional advantage, or cognitive superiority (supports V1, V6, P4). This categorical imperative is a bastion against abuse of autonomy and advancement of inequality.

The last categorical imperative is not related to ethical stance as the first two. "Act as if you were a law-making member of a kingdom of ends'' has an important meaning: the set of maxims and driving action policies must have internal coherence in it. As a "lawmaker" or in this case a standard maker or standard writers, the set of maxims developed will ideally provide adequate coverage over all possible use cases and that there are no internal contradictions. This was referenced earlier as the third challenge: coherency.

7.4 Driving action policies

Driving action policies can be derived from the maxims and should be consistent with the entire AV ethics framework. This subclause provides a set of example rules that could guide the design and ethical consideration of AV behaviour. The link between the values, principles, and maxims is forged by what are defined here as driving action policies. In a sense, driving action policies constrain the vehicle behaviour and conduct in accordance with some concept of operation.

Naturally, each operator (producer, fleet, etc.) may tailor some aspects for its unique operational concepts (e.g. a taxi fleet may behave somewhat differently than regular AVs, when it comes to picking up passengers).

Driving action policies are formulated by the manufacturer and will, ideally, be evaluated at three main stages along the design and development process.

a) It is recommended that driving action policies are designed and implemented in the actual design stage and comply with the values and principles discussed above.

b) In addition to the development itself, it is best practice to verify the correctness of the correspondence between the values/principles and vehicle behaviour (e.g. for a quality assurance team or for a regulatory agency).

c) Finally, there is the stage of validation. There should be correspondence between the actual vehicle conduct as well as how it is perceived by other road users and the values/principles discussed above.

The driving action policies detailed below constitute the frame concerning vehicle behaviour.

For each driving action policy, the maxim from which it is derived, and the relationship to the principles and values discussed above, are detailed. Driving action policies DR1-DR2 deal with the general frame for driving and DR3-5 are normal in the sense that they deal with regular situations on the road. Driving action policies DR6 and DR7 are "abnormal" in the sense that they discuss situations with outcomes and consequences that are not necessarily predictable and are, in a sense, ad-hoc. Driving action policies DR7 and DR8 are associated with emergency situations on the road, namely giving way to first responders and avoidable collisions.

Annex D gives a possible action plan that may be adopted.

7.4.1 Need versus desire driving action policy

If there is no strong need to take to the road, and thereby pollute the environment and take a valuable road space, perhaps the decision should be not take to the road. There may be other means to achieve the need (internet shopping, virtual meeting, etc.), take shared and/or public transportation.

> **Driving action policy DR1**: "whenever possible one should not take to the road and thereby avoid at all occupying a slot in the road space"
>
> A maxim from which this driving action policy is derived from can be defined as:
>
> **Maxim M1**: "One must act as if one does not exist."
>
> **Principles**: P1 - Transitoriness, reverence and sustainability
>
> P10 – Justice and responsibly
>
> **Values**: V4- Social progress
>
> V7 – Sustainability and proper balance

7.4.2 Once on the road space

If a customer indeed has to make the drive (e.g. passing rule DR1), the vehicle when it enters the road space should minimize any disturbance to the traffic. Namely, the vehicle should minimize lane switching, slowing down below the traffic's speed, unnecessary or risky overtaking of other vehicles, zig zagging to progress overly quickly with resulting disturbance and jeopardy of other road users. Namely, the vehicle should maximize acts that enhance the road space such as being "kind" to other road users (e.g. vulnerable road users), supporting their journey as well as enhancing the public space

with its appearance and actions. A vehicle should attempt to perform at least one such act during a sortie or at any other interval.

Driving action policy DR2: "As a general rule, one should not take any unnecessary action that will cause other road users to change speed, slow down, perform an evasive manoeuvre, or any emergency action due to your actions."

Driving action policy DR3: "Even if there are no other vehicles in vicinity, the vehicle should always be prepared for the sudden appearance of other vehicles with respect to its actions." (Explanation: even driving on a deserted highway in the middle of the night, it is not proper to switch lanes and disregard pedestrian crossings and the likes.)

The maxim from which these two driving action policies is derived from can be defined as:

Maxim M2: "One must act in such a way such as to minimize one's impact on other road users, the public space and the environment."

Principles: P1 - Transitoriness, reverence and sustainability

P3 – Cooperation and coordination

Values: V4 - Social progress

V7 – Sustainability and proper balance

Driving action policy DR4: During every ride or sortie, one should drive in a defensive and cautious manner that does not prioritize personal efficiency.

The maxim from which this "act in kindness" driving action policy is derived from can be defined as:

Maxim M3: "One must act in such a way such as to maximize one's positive impact on other road users, the public space and the environment."

7.4.3 In the lane behaviour (includes braking)

When a vehicle is in the lane, it should maintain an acceptable gap from the preceding car as well as from the car behind and the sides. The emphasis is on the acceptable, which does not only take into account safety margins (e.g. enough braking distance) but also the comfort/acceptance of the people inside the vehicle as well as the acceptance of its behaviour by all other road users. Specifically, an AV should not brake in front of vulnerable road users in a way that the perceived safety margins of vulnerable road users are compromised (except in unpredictable emergency situations). In the same vein, its braking profile should be considerate. Finally, an AV should not "inch" into a vulnerable road user's perceived safe field of travel in a way that is aggressive or feels unsafe.

Driving action policy DR5: "While in the lane, lane changes should not be made unless there is a comprehensible reason, for example, the speed differential is very high or there is a road obstruction."

Maxim M2: "One must act in such a way such as to minimize one's impact on other road users, the public space and the environment."

Principles: P1 - Transitoriness, reverence and sustainability

P3 – Cooperation and coordination

Values: V4- Social progress

V5 – Security and safety

7.4.4 Lane switching

Lane switching is a major cause of accidents. Improper lane switching and lack of turn signalling is the best predictor for a driver with a high risk of being involved in an accident. Improper lane changes also causes traffic slow down and leads to high fuel consumption. In many cases, unnecessary lane switching is done to maintain higher speed (time to destination), but in many cases the time saving is minimal (especially in urban settings or small countries).

Lane switching is needed for exits, entries, obstacles and the likes. Rule DR3 concerns minimizing any lane switching that is not obligatory and this set of rules defines the conditions for making a lane switch (given that it "passes" rule DR4).

Driving action policy DR6: "Lane changes should only be done if there is an empty 'slot' and there are no dangerous (exemptions see DR7) consequences"

Maxim M2: "One must act in such a way such as to minimize one's impact on other road users, the public space and the environment."

Principles: P1 - Transitoriness, reverence and sustainability

P3 - Cooperation and coordination

P5 - Equity and fairness

P7 – Non transgression and no coercion of others

Values: V3 – Equity and fairness

V4- Social progress

V5 – Security and safety

The following driving action policy (DR6) is a relaxation of the strict DR5 and should be used in circumstances of limitation that make DR5 infeasible (need criteria when DR6 is permissible).

Driving action policy DR7-a: "Lane changes should only be made if there is an empty or partially empty 'slot' and causes minimal speed changes and manoeuvring to the vehicle behind or ahead of the slot (except unpredictable emergency situations)."

Driving action policy DR7-b: "Lane changes should only be made if there is a partially-empty 'slot' and minimal speed changes and manoeuvring to the car behind or ahead of the slot (except unpredictable emergency situations)."

Driving action policy DR7-c: "If the vehicle under consideration of DR6, DR7-a, DR7-b cannot follow a predicted path (e.g. for making an exit), the vehicle should rather forfeit the manoeuvre and perhaps continue on another route or make a U-turn in the next exit than cause consequences in terms of speed changes and manoeuvring to the car behind or ahead of the slot."

Maxim M2: "One must act in such a way such as to minimize one's impact on other road users, the public space and the environment."

Principles: P2 - Reciprocity and caring

P3 - Cooperation and coordination

P5 - Equity and fairness

Values: V3 - Equity and fairness

V4 - Social progress

7.4.5 In the presence of the other

An AV is hardly ever independent. Every entity in the world has relations with another and being in an independent state is an illusion (Reference [23]). As such, the relation with other entities in the world is part of living and must be practiced well to achieve harmony with all others. In the same vein, an AV must act in such a way that comes from a place of reverence to all other entities and establishes good and healthy relations, starting from reciprocity with others and escalating to caring and compassion without need for reward.

For example, when an AV is nearing a pedestrian crossing it should behave and conduct itself in a way that signals reverence to vulnerable road users. The vehicle should not harass or inch forward to block vulnerable road users and signal its desire to begin moving. The vehicle should also not be aggressive, in terms of planned behaviour, or be perceived as aggressive, in terms of actual conduct, by any other road users (ORU) or even when there are no such entities in vicinity. The goal is to train the AV to always be in reverence to the entities in the world as a reflection of its internal behaviour and as a sign of taking responsibility toward all entities in the word, animate or inanimate (Reference [24]).

Driving action policy DR8-a: "Should act in a way that acknowledges, communicates with, and respects every entity in the road space". The reverence is practiced by positive interaction (no aggression, rule following, and yielding to the other even when he/she/it acts "illegally" and avoiding confrontation) and responsibility-taking toward other road users even at the expense of inefficiency, ineffectiveness and at times cost.

Driving action policy DR8-b: "The respect for the road space, social norms, and other entities comes from a place of 'duty,' which assumes a role for the AV as a member of the community and an entity that is responsible for the safety and well-being of others." This duty is extended to AV behaviour and conduct, which should always be respectful and proper, even "when no one is watching." That is, even in the middle of the night or when it is not mandatory, the AV should be respectful of all traffic laws, rules, and act in a manner that is "optically" correct.

7.4.6 Road/infrastructure use cases

This subclause reviews several specific road use cases. These are cases that will be as challenging to AVs as they are to non-AVs. The analysis shows that such use cases can lead to unethical behaviour by AVs, not because of any behaviour or conduct, but simply because the use case is challenging. Making the road space amenable to AVs may support unethical behaviour. Such situations should be handled in awareness of local law.

— **Roundabouts (of the congested kind)**

Roundabouts have proven to be safer and cause less carbon emission than signalled or non-signalled intersections, especially in urban and neighbourhood settings where they also cause drivers to slow down and are more respectful of vulnerable road users and cyclists (Reference [12]). However, roundabouts, where some of the entering roads are congested, can easily limit access to vehicles on other entering roads and it is thus practically infeasible to have traffic flow without AVs violating some of the maxims discussed (M3-DR6 and M4-DR7).

— **Four-way-intersection (of the congested kind)**

Four-way intersections are employed in the US in rural areas as well as in neighbourhoods. The right of way is given to the first arrival or the vehicle on the right of the arriving vehicle when it is not clear who has arrived first. During congestion and multiple lane intersections, it is also quite difficult to determine whose turn it is to have the right of way. In order to avoid unethical behaviour, safety problems and potential accidents producers of AVs and/or authorities could provide solutions to determine priority of driving in intersections.

— **Unsignalized pedestrian crossing**

Current behaviours at a crosswalk are strongly governed by cultures. In some countries, drivers are very courteous and tend to yield to vulnerable road users even when the latter are inconsiderate, while

in others, vulnerable road users are courteous and cautious and tend to yield to approaching vehicles. Still, there are many urban environments where neither the drivers nor the vulnerable road users are overly patient which frequently leads to near or even actual accidents by trying to force their way with potentially disastrous outcomes. Since AVs will be risk aversive, this will lead to a potential situation where vulnerable road users, especially in busy and congested situations, will quickly learn that playing the chicken game is a sure winner.

7.4.7 Resolving conflict

Until now this document has dealt with "normal" behaviour and conduct of an AV. That is, those situations and use cases that are standard and predictable. In the following subclauses (7.4.7, 7.4.8, 7.4.9) "abnormal" behaviour and conduct of AV, where consideration must be taken to resolve the situation in a proper ethical manner, by means of individual acts by the AV, negotiations, and help from others, will be addressed.

There are many situations where there is a conflict between two road users. Resolving the "conflict" is the foundation of many driving laws by providing a priority scheme that determines who has the right of way. The general approach advocated here for minimizing disturbance to other road users is to obey the priority scheme wholeheartedly. When, and if, the priority scheme is unclear, or when another road user is violating the scheme or infringing into the ego vehicle's "legitimate slot", (the road position that a vehicle is entitled to by way of local traffic regulations or accepted custom and practice) then the action should be to yield to avoid conflict.

Driving action policy DR9-a (unilateral): "Any entry into an unsafe situation should be avoided by taking proactive actions well in advance of the impending situation" (including avoidance of congested areas, lane switching to avoid anticipated conflict, or even a-priori routing modifications).

Driving action policy DR9-b (unilateral): "In a situation where another road user is about to create a conflict with regard to the ego vehicle's legitimate slot, the action should be to yield and give another road user the slot" (if another car is tailgating behind, then it is necessary to move ahead or to the side to give them the ego vehicle's legitimate slot).

Driving action policy DR9-c (bilateral; gradual diffusing): "If however the action of yielding the slot and right of way cannot be done immediately, the actions should be to signal intent to yield the slot and execute as soon as possible. (Cannot give the way immediately because of a truck in the adjacent lane, signal that you are intending to get out of the way ASAP.)

Driving action policy DR9-d (bilateral; minimization of risk): "In a situation where another road user is about to create a conflict with respect to the ego vehicle's legitimate slot or right of way, and it is not feasible to use DR7a, DR7b, or DR7c – the driver should take actions to minimize potential risk and impact (e.g. unable to get out of the way due to constraint such as single lane, begin slowing down to reduce potential impact from the tailgating car).

Maxim M5: "One must avoid conflict."

Maxim M5-humans: "An AV or robot should avoid conflict and not get into a conflict with a human being and should recede in the best possible way to avoid harm to human beings."

Principles: **P4d** – Human autonomy

P5 – No maleficence (by a robot)

P7 – Non transgression and no coercion of others

Values: **V4** - Social progress

V5 – Security and safety

7.4.8 Negotiations

Negotiations are means of resolving conflict and can take many forms, but they are always better than any other unconstructive ways of defending a person's own position uncompromisingly or taking advantage of an inequality, or coercion. The act of fair negotiation is a sign of maturity and respect for the other. Moreover, a negotiation that starts from a "level playing field" and remains there is a mark of equity. When the two or more negotiating entities also understand their responsibility toward the other's needs, and are committed to obtaining a outcome leading to a bilaterally fair (perhaps even "win-win") result, then the act and process of negotiation is taken to an ethical realm (Reference [24]). See P11 for more details regarding responsibility.

Driving action policy DR10: "When in conflict, either due to the lack of a formal driving law or due to other road user actions (purposeful violations, confusion, needs, etc.), one should seek to negotiate the situation from an equivalence state while taking responsibility over the process and outcome of the 'other side'."

Maxim M6: "One must negotiate as if it is not know on which side of the bargain one will find itself when the 'dealing is done'."

Principles: **P1** - Transitoriness, reverence, and sustainability

P3 - Cooperation and coordination

P5 - Equity and fairness

P11 - Justice and responsibility

Values: **V1** - Human dignity

V3 - Equity and fairness

V4 - Social progress

7.4.9 AV unable to function as intended

There will be situations that either due to internal malfunction or inability to function or negotiate in a certain road situation the vehicle must stop functioning. Such behaviour is accompanied with an escalation protocol of how the vehicle interacts and communicates this deficiency to the occupants, other road vehicles, and its (fleet) management. How the vehicle executes a manoeuvre to minimize its disruption given the specifics of the road space and other road users, weather, etc. is termed minimal risk manoeuvre (and discussed in detail in ISO/TR 21959-1 and ISO/TR 21959-2,). Sometimes the vehicle also demands support (e.g. manoeuvring right of way) from other road users. There is also at times the need to call on and engage first time responders (Reference [15]; SAE J2990).

From an ethical consideration aspect, every action on the road must be predictable as much as possible, well communicated and coordinated, and when unpredictable, it is indispensable to take extra care to minimize disruption and to overact in communication.

Driving action policy DR10: AV indicates that it requires help from others in dealing with an unplanned situation. When others do provide and yield, it should be ensured that the minimum risk manoeuvre taken for the AV and the corresponding responses from other road users are well communicated.

Maxim M7: "One must act in a way that is unsurprising to others."

Principles: **P3** - Cooperation

P5 - No maleficence (by a robot)

P7 - Non transgression and no coercion of others

Values: **V5** – Security and safety

7.4.10 Yielding to first responders and emergency response vehicles

The last three examples of driving action policies and maxims deal with emergency situations. How an AV could respond to first responders' vehicles, emergency situations encountered on the road space (e.g. a vulnerable road user who is about to be hit by another vehicle), and other dilemmas are addressed below.

The vehicle's behaviour in the presence of first responder's vehicle or personnel should come from a place of having the intent to do the best it can to support not just the first responder but also the situation at large. That is, the vehicle, via its sensors and willingness to act, is trying to do whatever it can to minimize the negative impact. The perspective taken here is that the vehicle's intent to act, its behaviour, supersedes what will be the consequential outcome.

Only if an emergency vehicle demands, an AV could take actions such as infringing on road markings, sidewalks, and taking another vehicle slot. An AV could take a slot from another vehicle in a way that can actually put the AV and others at some possibility of harm (e.g. moving into the opposite lane). The AV could make itself exceptionally communicative and visible. The vehicle, however, should not put another road user, and in particular vulnerable road users, in clear and present danger as a consequence of its actions. The vehicle should start building a corridor for emergency vehicle access in slow-moving traffic, when necessary.

Driving action policy DR11: "The progress of a first responder should not be limited. This may be at the expense of violating traffic laws, in an ad-hoc manner and a reasonable probability of negative consequences to other road users". For example, moving out of the way of an emergency vehicle, by pulling off the road in a safe manner.

Maxim M7: "One must act in a way that is unsurprising to others" to ensure that acts undertaken are reasonably foreseeable by others. Acts that violate the law would be unforeseeable, and should, therefore, not be allowed.

Principles: **P1** - Transitoriness and reverence

P3 - Cooperation and coordination

P11 - Justice and responsibility

Values: **V1** - Human dignity

V4 - Social progress

7.4.11 Protecting other road users

There may be cases, rare as they may be, where an AV sees an imminent harm coming with regard to other road users (e.g. accident about to happen) and can take action to minimize the harm. In cases where the AV can do this without damage to any other road users and itself, it would be expected that the AV perform such action to divert such harm. In cases, where the AV needs to block another vehicle from colliding with a vulnerable road user, for example, there are two main use cases: one, where the expected outcome to the respective AV will not harm its passengers, then such actions are laudable. Two, if the actions will cause harm and potential casualties to the passengers inside the AV, such actions, noble as they may seem, should not be allowed. The relevant issue is how much risk the vehicle will take to protect another and the willingness, or duty, to act from a place of responsibility. This issue in the context of the next use case: unavoidable accidents, will be examined later.

Driving action policy DR12: "Harm to other road users should be limited even if it results in a situation that violates traffic laws."

Maxim M7: "One must act in a way that is unsurprising to others" to ensure that acts undertaken are reasonably foreseeable by others. Acts that violate the law would be unforeseeable, and should, therefore, not be allowed.

Principles: **P1** - Transitoriness and reverence

P3 - Cooperation and coordination

P11 - Justice and responsibility

Values: **V1** - Human dignity

V4 - Social progress

7.4.12 Unavoidable collision with other road users

It may be reasonable to assume that an AV will find itself in a situation of unavoidable collision with another vehicle or a road user, as in manual driving. But the assumption is that the occurrences of such events would be lower than in human driving. Manufacturers of AVs place much emphasis on safety verification and validation of their designs and the behaviour of AV will be on the cautious and defensive side, probably far more than any human driver.

Nevertheless, mishaps are expected to take place, either because of logical design flaws, technological failures (e.g. sensor failure), or improper communication and coordination with other road users. In this case, just as a human driver is expected to do his or her best to minimize negative impact on other human beings, animals, and inanimate objects, an AV should do the same.

Many concerns have been fuelled by the "trolley problem" and similar decision making "dilemma" that can hypothetically arise when an AV is faced not only with an unavoidable collision situation, but also with a limited set of alternative actions. In these dilemmas, the set of alternatives is purposefully reduced to a choice between two or more options: it is important to mention that the trolley problem is a philosophical construct for which there is no "good" answer, and it was intended to fuel philosophical debate (see C.3 for a discussion of the "trolley problem" and its decision space).

The first priority in situations of unavoidable collisions should always be to protect human life even at the cost of property damage or minor physical injuries of other traffic participants (see M5). Specifically, an AV should do whatever it can, including risk of material damage of the AV (but not risk of human life), to minimize harm. An AV should not base its decision on personal attributes as, e.g. age, gender, physical or mental status. One example of several possible solutions how to deal with dilemmas is a decision-making policy from queuing theory called FiFo (first in first out). FiFo directs the agent to take action (e.g. evasive manoeuvre) toward the first entity it detects and does its best to perform the act successfully. It is unlikely that two (or more) agents will arise and be detected at exactly the same time. There is always the first arrival, and he or she must be dealt with as best as the vehicle can perform. The vehicle should also minimize as much as possible decision-making perturbations. (Buridan's principle is discussed in C.2.)

Finally, unavoidable collisions can also consist of quandaries involving "potential property damage versus other potential property damage" here the same approach is recommended. The vehicle should do its best with respect to the objects identified first and not be consumed with calculating utility values of properties.

Driving action policy DR13: "In an unavoidable collision situation, impact should be minimized to respectful entities".

Driving action policy DR13a: assessment of property versus property: FiFo principles

Driving action policy DR13b: assessment of human life versus property or minor physical injuries: priority of protection of human life

Driving action policy DR13c: human life/lives versus human life/lives: based on quantitative criteria when differences in quantities are evident. However, if the quantities are equal or no significant difference exists, FiFo is used. However, this attribute does not include attributes by traffic category (e.g. vulnerable road user).

Maxim M8: "Positive action must be taken to avoid or minimize severe negative consequences of any situation."

Principles: **P1** - Transitoriness and reverence

P3 - Cooperation and coordination

P11 - Justice and responsibility

Values: **V1** - Human dignity

V4 - Social progress

7.4.13 Other issues

Maxims and driving action policies are necessary to achieve an ethical decision-making process.

This document provides a methodology for the design and verification of driving action policies for AVs. The list serves as an example. Creating a complete list of AV use cases for ethical consideration in design and development is beyond the scope of this document.

It is also important to remember that producers, working in different cultures, may have different principles (and perhaps even values) that they wish to define. This is understandable and indeed the framework supports such differences. Naturally, as a consequence the maxims that correspond to these principles will be different. In this context it is worth mentioning that some maxims will be made public by the producer while others, presumably those dealing with the low-level aspects of the decision making and/or manoeuvres will remain proprietary. What is not recommended are situations where there is no harmonization between producers and the resulting confusion of other road vehicles when they encounter different behaviour and conduct from AVs.

8 Framework for rule construction and dealing with violations and deviations

8.1 General

The end results and the output of the ethical framework and methods suggested above are driving action policies that each producer subsumes and codifies in its vehicle.

It not a trivial endeavour to write rules, and especially not for AVs. Rule construction is hierarchical in nature and begins with an understanding of the kind of behaviours that the rules are in place to prevent or promote.

Specifically, rules that can be devised to govern behaviour are categorized according to whether these behaviours are dealing with prevention and self-constraint (e.g. abstinence) or to boost performance. Each of these behaviours can be further dissected, depending on whether the behaviour is either a) predictable or b) unpredictable. For each category, it is possible to observe if the behaviour is i) periodic (much easier to deal with) or ii) ongoing (more difficult). There are different design "tactics" for each subcategory. The intent is that once the behaviour requiring control had been understood, it is possible to build effective, reliable, and practical rules.

8.2 Framework

The hierarchical framework begins with the rule's goals and then begins to parse the goals to practical rules and tactics that can be used to make sure that the rule is consistent with known practices and

conduct that can be expected from an AV that is functioning in the transportation public space. Figure 6 describes the hierarchical framework.

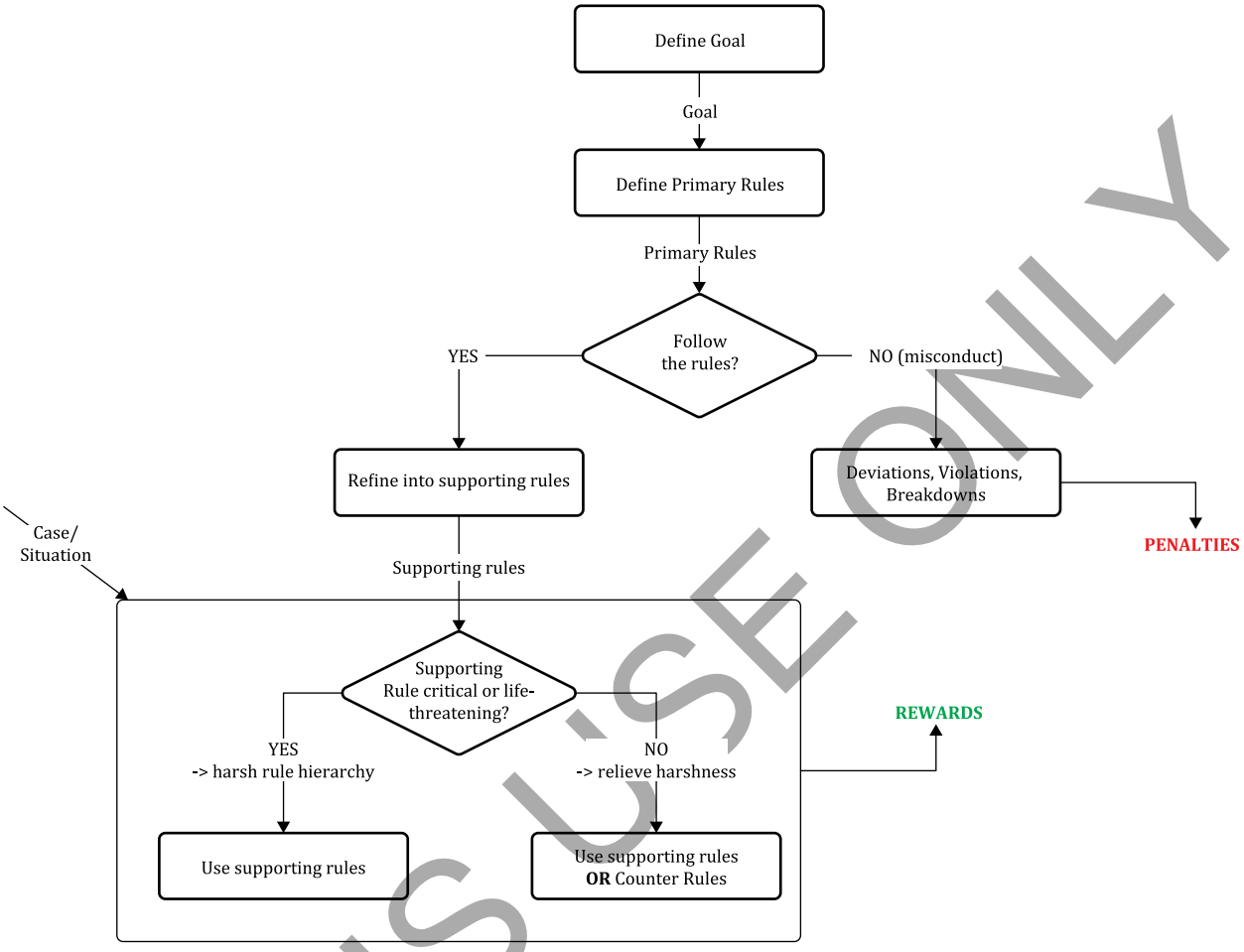


Figure 6 — Hierarchy of rule construction

8.3 Goals

In devising rules, the goal is defined first, or the desired end. Otherwise, there is no criterion to judge the rule and subsequent tactics. For example, a personal goal may be to be healthy and not overweight. But merely translating this goal into the relevant dietary behaviour does not tell us anything about ways and means for fulfilment and achievement of the goal. In the same token, keeping one's weight at 80 kg may sound like a rule, but it does not tell what to do and what not to do, and nothing about the management of the rule over time and in difficult circumstances. Generally speaking, it is preferable to treat a rule as something that describes and constraint behaviour, not necessary just outcomes 80 kg is an outcome, not a rule.

8.4 Primary rules

Rules that express the constraints, or requirements, on an agent's behaviour and conduct are called primary rules. These primary rules are behaviours to design for successful achievement that will fulfil the goals. It will be necessary to address anticipated conflict, consequential incompliance and devising enforcement strategies. Assuming that primary rules can be feasibly followed, what, then, is the optimal way to formulate them to achieve successful rule following and achievements? Primary rules such as "maintain driving safety at all times," "be defensive when driving," and "be sustainable" are all example of primary rules that require further elaboration. The maxims discussed in 7.3 are example of primary rules.

8.5 Supporting rules

Supporting rules mean a variety of rules that are in place to reflect the inadequacy of primary rules to take care of in-situ circumstances and difficulties. In making a decision to conduct an automatic lane change (ALC), it is conceivable that a system may decide to execute the manoeuvre when the parameters are close to the limit, only to encounter a violation of the rule after leaving the lane. While primary rules have their function as a way to establish the necessary requirements, one has to address supporting rules to help actually achieve them.

For example, rules can be made effective if you easily can tell the difference between compliance and violation. The notion of “bright lines" is commonly used by legal scholars to indicate the absence of ambiguity in rulemaking. A bright line is a discontinuity, a kind of a binary decision point that makes it clear if the AV’s behaviour and conduct resulted in obedience or disobedience of the rule. A clear bright light is the requirement to stop in front of a vulnerable road user stopping (even if there is an opportunity to pass) or not to overspeed despite the open highway.

Generally speaking, there are two main categories of supporting rules: precautionary and disabling for prevention, and reinforcing and enabling for performance.

8.6 Precautionary and disabling rules (prevention)

Precautionary rules provide margins of safety from violation of primary rules by drawing multiple fences, or boundaries, around the unwanted behaviour. They are set up to keep the AV at a measurable distance from violating the primary rules. It is difficult to determine if the AV will or will not be able to pass a vehicle in front in time to make an upcoming exit, it is much easier to rule out such behaviour, for example, when the distance to the exit exceeds 5 km. A disabling rule is an extreme case of prevention. Whereas the precautionary rule merely draws a brighter line at a safe distance away from the activity bade by the primary rule, disabling rules put the prohibited activity altogether beyond reach. For example, designing a rule that forbids an AV to make an unprotected left turn (commonly used in North America) and instead make a series of right turns or avoid such a route altogether is an example of a disabling rule. Again, as with the precautionary rules, there is a built-in mechanism such that at future time when an AV is most susceptible to violating the primary rule, it is de-facto disabled from that behaviour.

8.7 Reinforcing and enabling rules (performance)

The counterparts of precautionary rules are reinforcing rules. Making provision for AVs such as the requirement to give it the right of way in lane changes or other priorities over regular vehicles is an example of an enabling rule.

This document has listed a family of supporting rules to enable rule achievement. The focus has been primarily on management of rules via some manipulation of the context and the situations in which the behaviours occur.

8.8 Counter rules

The hierarchy of primary and supporting rules can be a harsh scheme. In cases where the rules are not critical nor life-threatening, it is possible to relieve the harshness of the rule system occasionally. The rule that one can “earn” oneself an exception can offer a certain relief that will allow flexibility and show that rules can sometimes be bent. Sometimes these exceptions, such as passing a vehicle that constantly changes speed, despite the rule that such passing violates a rule (e.g. not passing when exit is less than 5 km ahead). Such rules, which are qualifiers to the primary and supporting rules, are called counter-rules. Several categories of counter rules are examined below.

8.8.1 Exceptions (prevention and performance)

There are times when it is inappropriate to obey the very rules that were established for an AV. If circumstances are unfavourable to maintain a given rule (e.g. entrance into a congested roundabout),

there must be a set of rules for exceptional cases. Like supporting rules, the rules governing exceptions are least likely to promote a reaction from other road users if they are neat, simple and objectively defined. It is important to note that certain rules are not intended to allow any kind of exception (e.g. it is unlikely that AV would drive into a crossing where there is a vulnerable road user).

8.8.2 Discretionary and compensatory rules

An illustration of the difference between an exceptional rule and a discretionary rule is when inching through a congested roundabout when the AV has been waiting for longer than 5 min is an exception granted a priori, whereas a provision to allow the AV to cross a red light due to an emergency vehicle presence a discretionary choice that depends on in situ decision making that can be quite complex and involve risk taking by the AV.

It is possible to offset the negative public "perceptions" of robotic rules exceptions, by providing a compensatory substitute. For example, if an AV has had to force its way into a congested roundabout and cut in front of a vehicle, the AV can grant that vehicle right of way or allow it to be more expedite after the roundabout by moving out of the way or any other conciliatory manoeuvre.

In summary, counter rules provide an antidote to deal with the rigidness of the rule system.

8.8.3 Misconduct

Violations and deviations stand on the other side of the goals, primary and supporting/counter rules. Violation (Latin *violationem*: to treat with violence, outrage, dishonour), in our context, is an act against the structure of rule, rejecting both the goal and its primary rules. Deviation, from Latin *deviatus*, is to to turn out of the way, to depart from the established path. In the context of rules, it usually refers to the replacement of a supporting rule by another, shortcutting an existing rule or doing it differently than agreed on, but without the intention of violating and rejecting the goal and primary rule.

8.8.4 Violation

Violations are an integral part of any rule. It is anticipated that some AV developers will act against the rules (it may also be the case, perhaps at some future technological horizon, that the AV system itself will produce and execute an action that will violate a rule).

There are several means available to discourage violation, such as:

— making them unattractive in the first place by the assignment of credible penalties;

— making them visible and perhaps even highlight the acts. If the act is caught (on camera) and "published" it can be quite embarrassing.

It is also necessary to keep violations, especially the repeated ones, from causing a rule system to collapse (which can lead to breakdown, to be discussed next). A rule that declares that such repeated violations will be a reason for a fleet not only to stop its business but be placed under the supervision of another business can be a serious deterrent to such violations.

8.8.5 Deviations

Deviation is a lesser failure than a violation. It is about fulfilling the primary rules or goal in a different way. What is unique about intentional deviations is this interplay between intent (to deviate) yet with upkeep of the goal and primary rule. It is anticipated that deviations that will take place in AV systems and the regulator may have to define ways and means to deal with such deviations, in the same manner that courts deal today with deviations in case of a car blocking the road that requires passing over a white line or even responding to an emergency vehicle demand to clear the road.

8.8.6 Breakdowns

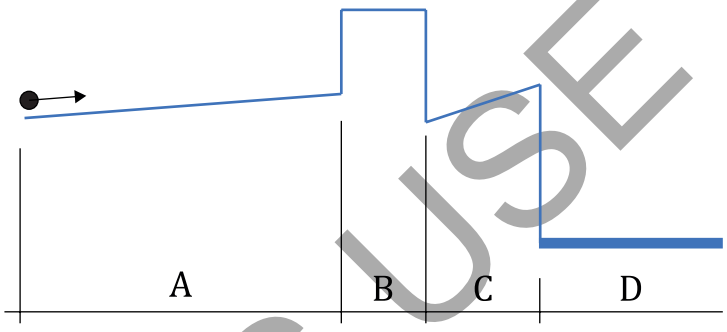
Such behaviours differ qualitatively from deviations and violation in the sense that breakdown is what threatens the full collapse of the original goal. These are cases of AV violence, for example, that cannot be treated as a momentary episode to be penalized according to the rules. This would be a serious offense, from which recovery cannot take the usual course, and serious outside intervention is required.

8.9 Rule strategy

Rules are means of control. Setting a rule will try to control how AV systems will behave at some future time. This document has considered, primarily, the design and conduct concerning a single rule. It is also possible to determine a strategy for the implementation of several rules, perhaps in sequence, to prevent a behaviour or to promote it.

8.10 Boundaries (for prevention of unwanted behaviour)

Herein, are defined the kind of rules necessary for building boundaries prior to the unsafe region or unwanted system behaviour, to include margins, barriers, and buffers. Figure 7 presents an abstract and temporal graphical depiction of the various boundaries that can be "erected."



Key

A margins
B barriers
C buffers
D unsafe region

Figure 7 — Boundaries

8.10.1 Margins

Margins of safety, operational margins, and safety shields are terminology to describe primarily a physical space that provides extra leeway prior to an undesirable behaviour or event. This is a space where safety procedures, driving action policies, placards, and signs limit entry to where problematic behaviour can take place or where unwanted events occur. There are system indications, alerts, and warnings that can be used to signal when an AV is exceeding its margins.

8.10.2 Barriers

Once margins with their warnings and alerts have been surpassed, it is time to consider physical barriers. The recommended approach is for one to anticipate the situations, behaviours, as well as specific events that will happen in the future.

8.10.3 Buffers

In many situations an additional "space," after the breach of hard barrier defences should receive consideration. In medieval architecture, beyond the moat and before the walls of the castle, there is a

physical space called the berm. It provides an additional defence and may give the defenders time to act before the walls are attached. Finally, note that from the moment of breach to the (bad) consequences, there may be differing time intervals that will allow for different rules.

8.11 Promoters (for performance)

The opposite of boundaries to block "bad" behaviour are "promotors" to encourage "good" behaviour. It is desirable to define a rule environment that encourages continual performance improvement and accommodation of the public on part of AVs. When the public is made aware of how ethics rule design and rule following is critical for safe AV deployment, it is possible to anticipate expectations and bring attention to manufacturers to design AVs that follow traffic laws and regulations as well as follow sound AV driving action policies, as discussed in Clause 7. Sometimes there is a beneficial side effect such as extra bonus or new business opportunities when a manufacturer has established itself as one that is cognizant of rule following and is strict in its quality assurance process and responses to violations and deviations (e.g. recalls).

8.12 Further considerations

In anticipation of the arrival and deployment of AVs as the first manifestation of robotic systems in the public space, much consideration should be taken regarding how such rules are constructed and how potential difficulties are taken into consideration in the design. In Clause 7 driving action policies were discussed, as well as a framework for their consistency and coherence. The rules themselves were examined and how to ensure that an AV complies, addressing the difficulties and unforeseen deviations and the likes. Such considerations are critical for a standard on AV ethics as it complements the ethical concepts and expectations with the realities of everyday practices.

9 External/internal design

Ethical considerations during AV design, development, and deployment are not only restricted to the AV driving behaviour or policy. There are also ethical considerations in the design and development of the physicality of AVs and the communication between AVs and the public (e.g. VRUs). These aspects are not included in the driving action policies but addressing them would lead to an increase of traffic safety. Therefore, the operationalization of the values, principles and maxims are extended to the physical and communication elements of AVs. Design considerations are offered here to address some of the pertinent ethical issues in physical and communication design that are related to traffic safety. These considerations are not exhaustive and only provide an example of possible design considerations.

a) The needs and requirements of people with disabilities and other marginalized social groups should be considered when designing AV system behaviour, physical design and communication design. Diverse groups of people interpret the actions of an AV differently. For example, a group of elderly people may be more cautious in their interpretation of an AVs behaviour. An example of this is the braking profile of an AV.

    1) Principles: P2 reciprocity and caring, P3 cooperation and coordination, P5 equity and fairness, P7 non-transgression and no coercion of others, and P9 beneficence.

    2) Values: V3 equity and fairness, V4 social progress.

b) The communication between AVs, VRUs and other road users must be considered. As described above, even the AV's braking profile communicates with road users. Please refer to ISO/TR 23049:2018 for more detailed information on ergonomic guidance on external visual communication between automated to other road users.

    1) Principles: P3 cooperation and coordination, P5 equity and fairness, P9 beneficence, P10 explicability to occupants and other road users, and P11 justice and responsibility

2) Values: V3 equity and fairness, and V4 social progress

c) The communication between AVs and passengers should be considered as should the communication between AV passengers and the outside world. Please refer to the ISO 9241 series for further information on accessibility and the ergonomics of human-system interaction.

1) Principles: P3 cooperation and coordination, P5 equity and fairness, P9 beneficence, P10 explicability to occupants and other road users, and P11 justice and responsibility

2) Values: V3 equity and fairness, and V4 social progress

d) Identification of biases and prevention of their spread to AV design should be considered in order to make the space within the AVas inclusive and as safe as possible. For example, designers and developers are made aware of possible existing biases in ergonomic data (e.g. crash test dummies) and they make sure that those biases are mitigated for the physical safety testing and when creating digital human models for simulating various scenarios (e.g. crash scenarios)

1) Principles: P4 human autonomy, P5 equity and fairness, P6 non-maleficence, P7 non-transgression and no coercion of others, and P10 justice and responsibility

2) Values: V1d human dignity, V2 freedom of mobility and making decisions, V3 equity and fairness, V4 social progress, and V6 privacy, intimacy, and human decision-making autonomy

10 Sustainability

From the perspective of a sustainable traffic system, it is clear that the changes induced by AVs will affect economic, environmental and social issues. The framework described in this document may support the identification and assessment of ethical issues related to sustainability and further supporting key sustainable development goals (e.g. SDGs 9 "industry, innovation and infrastructure", 11 "sustainable cities and communities", and 12 "responsible consumption and production").

For example, different sustainability issues can be addressed by the intrinsic technical attributes of the vehicle itself on a micro level, and the use of ethical principles in their development, e.g.:

— increased road safety through minimization of human driving errors;

— increased road safety through crash avoiding technologies;

— lowered fuel consumption through electrical drivelines, lighter vehicles, technology enabling smoother driving patterns (e.g. keeping the speed limits and less braking), etc.;

— increased effectiveness of the road transport system through crash avoiding technologies, technology enabling efficient driving patterns like platooning and route optimization, etc.

These technical attributes are expected to contribute to sustainable and accessible mobility on a macro level, especially if ethical considerations have been taken into account during design and development. Contributions may include enabling ride-sharing, lessening congestion, increasing accessibility for different groups of people which cannot use the road transport system fully due to age and disabilities, lowering transportation costs, more efficient use of time, etc. However, some of these attributes may have opposite effects due to changes in consumer decisions on, e.g. living and workplace locations and daily activity and travel patterns. Such behavioural changes are likely to dominate the sustainability implications of AVs. The lifecycle of the vehicle itself, from extracting raw material to reverse logistics, also has an impact on sustainability and must be considered in the design and development process.

It is important to consider the micro and macro levels of impact when addressing sustainability issues. The use of the framework in this document and the implementation of ethical principles when designing the technical and other attributes of the AV systems may support development teams in mitigating negative effects (short-term and long-term) of the technology from a sustainability perspective and promoting positive ones.

11 Review and re-evaluation following controls system updates

AVs are, as well as other vehicles, subject to continual review, updating, design revision and amendment.

Also, the systems within the vehicle, which are subject to safety ethical considerations, are under permanent review of the vehicle producer.

The vehicles, which are allowed to be operated in public traffic, must have passed a certification procedure to verify that they are built in accordance with the actual regulations.

After-market modifications like updates or changes of systems must pass certain certification procedures before they may be implemented in the vehicles. This includes the systems and sub-systems, which are relevant for the safety ethical aspects of the safe operation of the vehicle (e.g. “interpretation of traffic rules” and “decision system”).

Non-manufacturer modifications must be subject to acceptance by the relevant authorities.

Annex A
(informative)

Overview of ethical philosophy related to AV

The outgrowth of this document is guidance as to how an AV should conduct itself (the engineering term used here is "behave") and how this individual behaviour is governed across vehicles. Specifically, to ensure that not only are traffic safety and effectiveness maintained, but that the operation is also ethical (e.g. equitable, fair, and accommodating in accordance with human values). In addition, to make sure that the deployment of AV will produce a utility that is worthy, acceptable, and meaningful for society as a whole. Finally, at the end of this document, it is explained how to really know that the ethical guidelines written are indeed useful.

Attention is focused in the early parts of this document on AV behaviour (conduct) as well as on the rules and principles for making sure that their overall behaviour (governance) is ethical. (The utility of AV deployment is addressed toward the end of this introduction as well the assurance that the guidelines presented here are indeed valid.)

With respect to these two issues, a branch of philosophy called "ethical philosophy" is a reasonable starting point. This philosophy has occupied itself for generations (starting primarily with Socrates and Plato in the occidental tradition and the Bhagavad Gita and the Code of Hammurabi and law in the oriental and Middle Eastern traditions respectively) -- with how an individual (agent) should conduct itself and behave and the cumulative (emergent) behaviour of many other agents. Drawing upon these insights enables questioning how robotic agents such as AVs should behave.

Broadly speaking, moral philosophy/ethics can be divided into several schools of thought, most prominently.

a) **Consequentialism:** Consequentialist ethics postulates that the normative properties of an action depend on the consequences they bring about. Consequentialist ethics assumes that humans are goal-directed and that aim at producing certain results to achieve their goals, hence the emphasis on the consequence. The most widely known case of consequentialism is utilitarianism, which was developed by Jeremy Bentham and John Stuart Mill. According to Bentham, a morally right action is the one that brings "the greatest happiness for the greatest number of people". Some of the issues surrounding consequentialism are the definition of the benefit ("which consequence should we attempt to achieve?") and inclusiveness ("consequences for whom?").

b) **Deontological ethics**: While consequentialism focuses on the consequences of an action, deontological ethics focuses on the actions themselves, more specifically on the duties. Immanuel Kant is recognised as the most famous proponent of this ethical theory. Kant proposes that a morally right action is the one that follows a universally defined and unconditionally valid concept, which Kant calls a categorical imperative. While deontological ethics emphasize living according to some universal moral laws or rules, it does not specify them. From a deontological point of view, something is moral not because of its consequences, but is moral because the motive or intent is "good." Kant's approach to ethics begins with an analysis of "ulterior motives." Something could look good, and really be bad; and vice-versa, something could look bad, and really be good. Kant then proceeds to analyse the acts of so-called "good samaritans" to see why they do good things for complete strangers. What matters is whether or not the good samaritan is truly, or formally, doing the good thing out of the kindness of their heart – or whether they expect payment, glory, or the return of a favour. The question then becomes: "Under what circumstances will people sincerely do good with no expectation of benefit?" Kant says the answer is when people are "doing their duty" and the concept of duty becomes an important part of his ethical framework.

c) **Virtue ethics**: In contrast to the consequences of a behaviour or duties, virtue ethics emphasizes the moral character of actions, how virtues are acquired, and applied to behaviour. It has its origins

in the work of Plato and Aristotle. Instead of asking the question, “What should I do?”, virtue ethics asks itself, “What would a virtuous person do in this situation?”. It is concerned with overall moral character and making decisions that demonstrate one's virtues. Modern proponents of this ethical school often ask the question, “Would I be happy for my decision to appear on the front page of tomorrow’s news?”. Virtue ethics shares the conflict problem with deontological ethics.

With the rapid advances in the development of autonomous technology and artificial intelligence, several international organizations adopted a rights-based approach in recommendations on the values and principles for autonomous systems and artificial intelligence based on treaties, such as International Human Rights Law or EU Charter of Fundamental Rights. These include but are not limited to the European Commission's European Group on Ethics in Science and New Technologies, AI4People, Declaration of Montreal, European Commission's High Level Expert Group on Artificial Intelligence, the COMEST Report on robotic ethics and the UNESCO AI Ethics. In Reference [21] it is concluded that the principles covered in various documents correspond to the four core principles of bioethics - that is, beneficence, non-maleficence, autonomy, and justice, which is complemented it with a fifth one - explicability, that is considered indispensable for the developing technologies.

Annex B
(informative)

Sustainability issues

From the perspective of sustainable transportation system, it is absolutely clear that the changes induced by AVs will affect economic, environmental, and social issues. When endeavouring to look from the concept of sustainability that balances towards developing better society by managing the impact of transport on environmental, economic, and social issues, some of the examples can be cited below with respect to its number of implications.

The economic implications of the AVs would gradually bring systemic changes induced by their adoption. Tourism and construction activities would be guided by AV-induced urban sprawl that would result in housing and road sustainability. In the retail sector, service activities (e.g. by making delivery of goods and meals more convenient) can also be boosted. Driverless AV trucks result in cheaper freight transport, lowering prices of goods. Labour markets will be affected with the development of driverless AV, extending the effects of information technology on places of work. A number of jobs will be lost due to this impact. There will be a direct effect on the impact on employment loss due to widespread AV adoption. It would accelerate the elimination of most driving jobs in trucking, taxi, and public transit sectors—a significant stranding of the labour force.

It can be realised from the fact that adoption of AV along with the extension of information technology would create more jobs while on the other hand, there would be loss of jobs in other sectors. An extensive work to understand and appreciate the above issues is required in a holistic manner so that the problems with respect to sustainable issues can be addressed more rationally and scientifically.

Annex C
(informative)

Responsibility and accountability in the context of AVs

C.1 Overview of responsibility and accountability

A hierarchical framework for understanding responsibility and accountability and how it relates to control authority and the ability to make ethical decisions in the context of AVs is presented in this subclause. In particular, the notions of control authority, supervisory authority, responsibility and accountability which are key foundation when it comes to the design of fully automated control systems, are discussed.

Control authority

An agent's control authority relates to its ability to control an ongoing process as well as intervene and change behaviours when necessary. In the case of an AV, control authority is in the "hands" of the machine, based on the instructions and computer code developed by engineers.

Supervisory authority

Human operators who supervise an AV or a fleet of AVs are given authority to make decisions and strategic plans and at times even goal modification. Supervisory authority is granted to an agent by some higher, or superior, entity such as a ranking official, a group of official, or some other entity that "owns" the endeavour. Currently, supervisory authority is only granted to humans.

Accountability

An agent is liable and its actions must be answerable for what it is entrusted with and bow to some post-hoc judgement (from ld French *aconter* "to enumerate; reckon up, render account."). The basis of accountability is an obligation to comply with the agreement either written or implied. Such organizational accountability is generally well defined, as well as the administrative authority derived from it, including its boundaries (Reference [28]).

As such, accountability concerns obligations to comply with rules as well as duty to fulfil the agreement or expectations. Immanuel Kant's conception of one's duty, as described in his categorical imperative is quite fitting here[9]. A human agent can be held accountable and perhaps also reprimanded for his or her dereliction of duty. There is a duty that comes from a place of obligation and rule following and there is a duty that comes from a place of commitment. The former is rule based and the latter extends beyond the standard rules.

Along these lines, those who design the control authority of AVs are held accountable toward those who granted them with task of designing the system. Namely, the system architects and engineers who built and made the decision about its behaviour are accountable toward the chief engineer of the firm, CEO, and also toward the OEM who manufactured the AV. The main point here is that when authority is bestowed to an agent, accountability toward the granting agent is important.

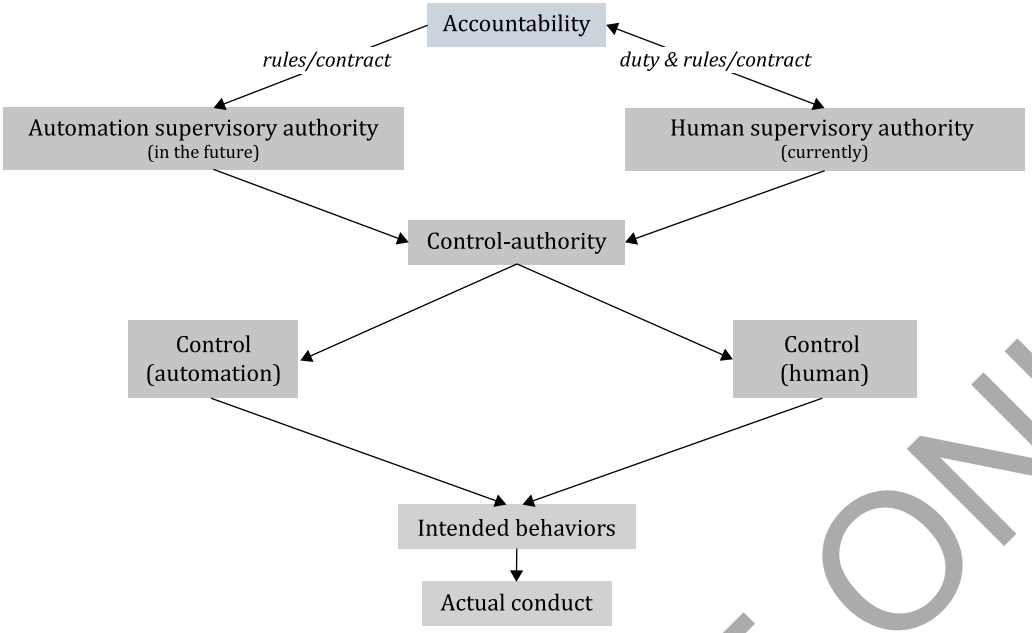


Figure C.1 — Control- and supervisory-authority and relations to accountability

Figure C.1 shows the relations between control- and supervisory-authority and their relations to accountability. When a human is in control of the system, he or she is able to produce behaviours (e.g. manoeuvres) based on their human abilities (driving the vehicle). The automation controls the system and produces behaviours based on instructions and software code, some of it pre-defined and some is computed in situ. Whether human or machine, the control authority is granted by a supervisory authority that currently is assumed by human operators, for example, in the control centre of an AV fleet. (In the future there may be designs where the supervisory authority is granted to an automated machine.) The supervisor, be it a human or machine agent, is held accountable in accordance with the rules and contracts that are in place. In the case of human supervisor, he or she is also bound by a "duty" that comes with the authority given to them by the person or entity they are accountable toward.

Responsibility

Although accountability seems as a reasonable closure to the hierarchical framework presented here (see References [29] and [30]), in fact, in its rule following and duty perspective, it is somewhat limited. This is because it does not take into account the possibility of the ability to go beyond duty and contract/rule obligations. While precise rule following may be as far as can be expected from a computer, most people expect more from those who are given authority over their lives. For example, pilots and sea captains are held to standards that extend beyond rule following and accountability toward their supervisors. Instead, they have a sense of commitment toward their constituents (e.g. passengers) by professional and humanistic standards. The flying public, for example, has an implicit expectation that their flight crew will do whatever is needed and even beyond what is expected to secure their safety, and airline pilots take this commitment with grave seriousness; some hold this commitment to a level of conviction.

When it comes to AVs, it cannot be assumed that automated agent will be go beyond its rule following authority to help the passengers in time of dire straits. However, such actions, albeit remotely controlled, are expected from human operators who are managing the AV from afar. Moreover, machine agents, on their own, cannot address the kind of conflicts and dilemmas that involve humans' well beings; because solving these dilemmas require circumventing rationality, optimality, and shunning rule following. Although AVs are built by humans, for it, the passengers are, at the end of the day, objects. It seems that it is impossible, at least on the current technological level, to attribute to autonomous systems any kind of responsibility for human beings.

Virtue and higher calling

Virtue is the willingness to take one's responsibly toward others to an extent that is beyond what's expected by one's accountability. This responsibility includes not only the understanding of purpose and meaning of a given situation, but also its moral context (References [16] and [17]). Unlike accountability, such responsibility cannot be formally defined, because it is unbounded, and presupposes any free choice. This responsibility can be extended by humans toward other humans and perhaps also beyond to include animals and inanimate entities. It has no rational basis, can only be understood emotionally and at times is realized in non-meditated moral decisions with very difficult implications (see the discussion concerning the acts of Captains Evans and Weiland in C.2). The willingness to take these virtuous actions demand faith, not in the rudimentary sense, but in the sense of faith in what is sometimes defined by modern philosophers as the "higher good" or" higher calling" .

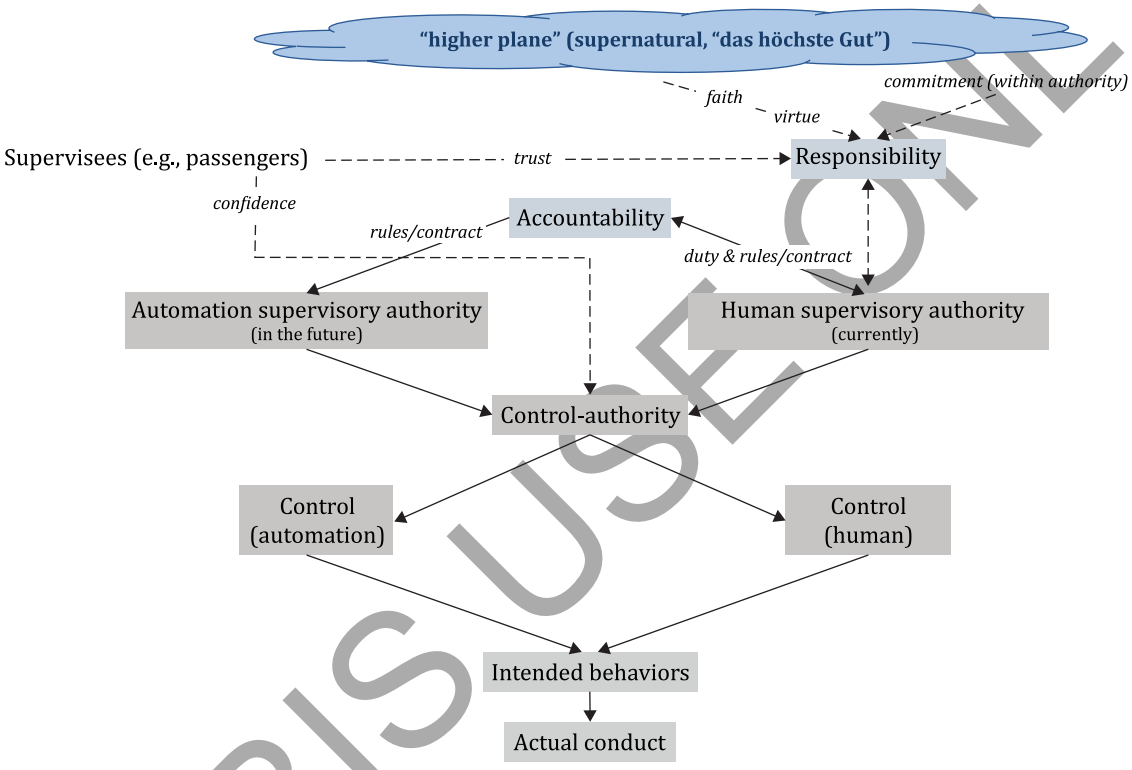


Figure C.2 — The accountability-responsibility framework

Figure C.2 shows the entire framework for accountability and responsibility of both humans and machine agents. Note that while machine agent can only be accountable to those that granted it control authority, the human agent possesses, in addition to accountability, also responsibility. The limitation of accountability is represented by the fact that customers of an AV for example can only have confidence in the behaviour of the machine. Trust, on the other hand, can only be with respect to humans, who can act beyond the rules when the situation demands it. Responsibility, of the humankind, receives its mandate from two sources: the more rudimentary one is from the pilots' "commitment" (towards airline passengers, e.g.), which is bound by their authority and the second, which is more sublime, is from the notion of the "higher good", which is boundless.

C.2 Buridan's principles as applied to AV decision making

There is an age-old hypothetical situation, dating back to Aristotle, about a decision-making problem of not being able to choose between two or more alternative, resulting in constant perturbations between the two alternative and a consequential (and sometimes dangerous) standstill. The paradox is named after the 14th-century French philosopher Jean Buridan, whose philosophy of moral determinism it satirizes. The example is of an ass that is equally hungry and thirsty, who is placed precisely midway between a stack of hay and a pail of water. Since the situation assumes the ass will always go to whichever is closer, and both are set at an equal distance, the ass dies of both hunger and thirst since it cannot make any rational decision between the hay and water.

The situation of Buridan's ass has been used in control theory. For example, consider a driver approaching an intersection and trying to make a decision whether he or she has enough time to cross before a train arrives. If the driver perturbates between a go or no-go decision, thus accelerating and braking and then accelerating and braking on and on, there is a trajectory that can lead to the car ending up in a standstill on the tracks. Similar situations have been observed in negotiations and turn taking in traffic.

The possibility of a “Buridan’s ass” like perturbations when an AV is confronted with two equally important alternatives and when there is competition between the two possible manoeuvres in the spaciotemporal dimension of the problem should be considered cautiously. In 7.4.12 there is discussion of the application of the first come first serve principle to avoid a Buridan’s ass decision making perturbations.

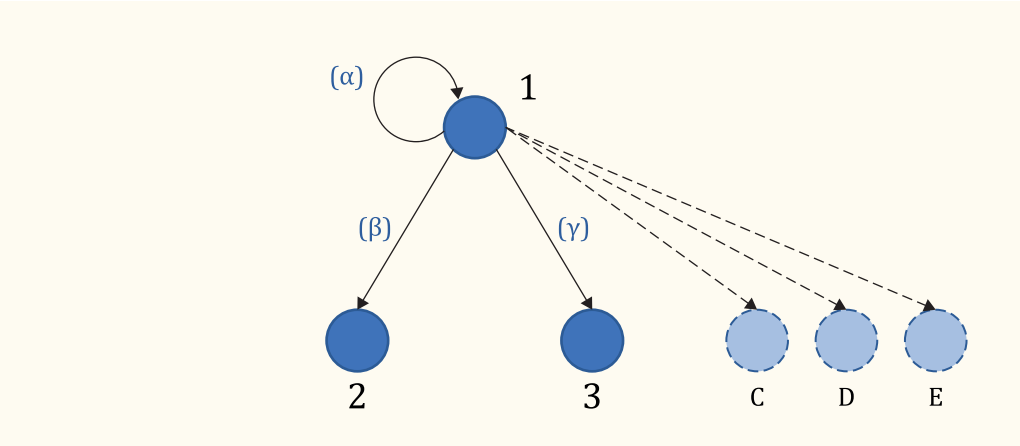
C.3 The trolley problem

The entrance of AV into the public discourse, brought with it some fear and concern regarding the reliability of these systems - partly due to lay people’s unfamiliarity with the technology and fear/anxiety from robots. Much of these concerns have been fuelled by the so-called “trolly problem” and similar decision-making “dilemma” that can hypothetically arise when an AV is faced with an unavoidable collision situation and the AV will have to decide who will die and who will live. To be fair, the probability of an AV encountering this situation is so slim that the interest in the problem is to a large degree academic only as there is hardly any evidence to such dilemma situations in the 120 years of car- bus- and trolly-based transportation systems. The majority of human drivers have never faced the decision of “who will be killed today,” but nonetheless in the context of AV there is a conception that such problems will somehow arise. The emphasis of this academic problem as having a corollary in practical AV deployments has been used by some people to disqualify AV.

While the problem is indeed academic, there is a chance, in the context perhaps of future robotic systems that may govern human life that a similar situation may occur. Therefore, the deployment of AVs as the first manifestation of robotic systems that can harm people, does deserve some consideration and attention placed on these dilemmas. Moreover, since, deserving or not, the issue has caught the public’s attention, however, this issue should be addressed head on because it has become a mainstay.

The concept of what eventually will be termed the “trolley problem” was first conceived by Philippa Foot in a 1967 paper as part of her “doctrine of double effect.” (Reference [31]). She presents the problem as “the driver of a runaway tram which the driver can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track the driver enters is bound to be killed”. From Foots’ perspective “The question is why we should say, without hesitation, that the driver should steer for the less occupied track, while most of us would be appalled at the idea that the innocent man could be framed.” Interestingly, proceeding her short discussion of the trolly problem is a case of the pilot who has to make a decision which area to hit after having some catastrophic failure. Foot’s concept, which is only a few sentences in her paper, was taken by J.J. Thomson made into a full paper entitled “The Trolley Problem.” Over the years, various variants developed over the original problem to include not only tram and trollies but also situation involving terrorist blackmail demand and medical triage and organ harvesting decision making (“cut up Chuck''). These problems are not in practice, but they allow us to frame our thinking, and this is another reason why they deserve some attention. It is not clear who first made the link to AVs.

The hypothetical “trolley problem” poses three basic decision-making options: either the own vehicle will act and will then be damaged/destroyed, or the first external agent will be damaged/destroyed (agent A), or a second external agent (Agent B). If the own agent decides not to sacrifice itself, it can potentially be made to choose among the two (or more) agents to kill. Figure C3 is a graphical description of the problem.



Key

1 own agent
2 external agent A
3 external agent B

Figure C.3 — Decision making for trolley problem

To the best of our knowledge, the only practical cases that are somewhat related to the trolley problem have occurred in aviation, specifically military aviation: On June 16, 1953, Captain Francis T. Evans, an American fighter pilot, tried to land his F-86 Sabrejet after loss of hydraulic power. He needed to eject, but straight ahead in front was an elementary school where dozens of children were out on the playground. He decided to enter a 45° dive so the plane would hit the ground in front of the playground. Captain Evans had managed to crash the jet 30 feet (approximately 10 m) before the playground, which protected everyone from the flying pieces of the jet. No child, teacher or school employee on the ground was injured in any way. His attempt to eject during this manoeuvre was fatal.

On August 8, 1954, William H. Weiland, also a US Air Force pilot, encountered a catastrophic failure in his F-84 G Thunderjet fighter while flying over suburban neighbourhoods in Long Island, New York. He could have bailed out and parachuted away from his disabled aircraft but decided to continue to fly the plane and make a forced landing on a street between a row of houses. He managed to land the plane perfectly between the houses, but the aircraft caught on fire. Five bypassers were injured and one died a few days later in the hospital. He himself could not get out of the burning plane.

These two cases can be mapped using the model in Figure C.3. In both cases, the pilot's option was to either eject safely with the unknown consequence of hitting people on the ground (event β), or act to minimize the consequences to people on the ground by taking a calculated personal risk (event α). Captains Evans and Weiland did not intend to die and in this sense, these two cases differ from the theoretical trolley problem. Their intention was to take personal risk to minimize potential harm to others.

It is suggested here that the discrete decision-making form of the trolley problem is not necessarily the problem of the AV as it is not about pitting one life versus another's or several others' lives. The real AV issue is the willingness or unwillingness to take the good Samaritan stance and the measured risk that the AV is willing to take to minimize harm to other road users. This differentiates the AV ethical problems from the dilemmas.

Use case 7.4.12 discussed decision option (α) and dismissed it: suggesting that the vehicle should do its best, including the risk of material harm to its own outer shell, but short of expected physical damage to its occupant(s). In the hypothetical case where there are two external agents that the vehicle may end up hitting (agents B and C in Figure C.3) the question becomes "who should a vehicle cause harm to?" (e.g. one person to the left or five people on the right). One approach to "solve" this problem is to employ utility calculations of (expected) value: one life sacrificed versus five lives saved, one man with a short life expectancy versus another will full life expectancy, etc.

C.4 Addiction by design

Addiction by design is a term used to describe a design approach that aims to create products and services that ensure addiction to them when used. The term was popularized by the anthropologist Natasha Dow Schüll[19] and her research on the coercive design of machine gambling. The principles behind addiction by design have been integrated into most of social media and digital game development. Typically, algorithms are configured to persuade users towards continuous use and lengthen the amount of time spent on the product, service or device. Designing to ensure addictive behaviour is unethical. The designers and developers of AV systems using this document, particularly when designing interaction between the AV and its passengers, should not encourage negative addictive behaviour.

Corresponding values: V3 equity and fairness; V6 privacy, intimacy and human decision-making

Corresponding principles: P4 human autonomy; P7 non transgression and no coercion of others; P11 justice and responsibility

C.5 On Kant's ethical approach and its applications in AV ethics

Immanuel Kant has used maxims to evaluate ethical decisions (e.g. behaviour) and their outcome/consequences (conduct). Kant developed a set of "categorical imperatives" as a framework through which a maxim can be tested for determining whether the actions they guide are right, wrong or permissible. Kant starts off with an ethical imperative that is, for him, the foundation of all ethical behaviour: "Act only according to that maxim by which you can at the same time will that it should become a universal law." Namely, when one develops a maxim for (ethical) conduct, one should assess the implications of this maxim, as if everyone should act in accordance with this maxim (universality). So, for example, if a personal maxim says that "while I am driving, I must take positive action to save a dog that crosses in front of my car.", a law is then generated from such a maxim by applying it universally. For example, "Every driver must take positive action to save a dog that crosses in front of his or her car."

Kant's method requires the use of this process every time an ethical decision needs to be made or a design of AV behaviour, whether in a general situation or a specific (road) use case. The process starts by coming up with a private or producer-specific maxim for the situation, then universalizing that maxim and lastly by verifying if that universal law is something that should be followed to the letter. When all possible situations are considered, realistically and practically, it may be found that the above maxim with the dog is perhaps not the best to instantiate as a universal law. It is understood that there are situations when a driver should not take positive action for a dog, e.g. when driving on a narrow mountain road with a deep ravine to the side, or that sometimes saving a dog who has entered a high-speed highway can lead to consequences that are problematic.

Annex D
(informative)

Action plan – Example

Ethical considerations may be overwhelming. To avoid paralysis of action, the action plan below is deliberately simplistic to start the process.

a) **Simplifying ethical to equal and precautionary**

Value differences are ignored and people and objects are all assumed to be equal and no touch is accepted. If not all can be saved, strict sequence of appearance and following action determines the outcome. This is also realistic to expect from a manual driver.

The precautionary principle is followed. Safety first is set and developed without gambling. the brake pedal is developed first and then the accelerator pedal.

b) **Starting with 'stop objects'**

Stop objects are defined, objects that the AV should not run over. Initially, the AV is made to stop and break for all, undefined, objects. This will make the AV not moving at all. It grants that nothing unexpected can be run over.

The break function includes determining the intensity of the break. It starts with break intensity set to maximum, which is equivalent to the where it leads to the minimum stop length.

c) **Defining 'go objects' and 'go situations'**

Go objects are defined, objects that the AV is allowed to disregard. For example, non-dangerous particles on the pavement and static or dynamic objects on the sidewalk without probability to enter in front of the AV.

A go situation would be when all objects in its field of vision (field of sensing) are recognized as go objects. Now the AV can go. But this situation will change as soon as the AV starts moving. Hence a frame-by-frame development of go situations will follow until the AV can reach B along a straight line. Continue until the AV can reach sufficient Bs.

d) **Defining 'turn situations'**

The conditions that must be fulfilled to make the AV turn (in addition to go) are defined, for example, at road bends, intersections and when passing other vehicles. It must be specified exactly what to do in each situation, such as increase or decrease the turn, frame-by-frame. Situations are added until the AV can reach sufficient Bs with combinations of go and turn.

e) **Documenting the considerations**

Each above development step will probably have its precautionary or other ethical considerations on a more detailed level. These are documented continuously for learning and transparency.

f) **Do-good (benevolence) development**

The above, no-harm philosophy is a challenge. It will require its time and may be sufficient. But in case do-good (benevolence) principles will come in question, as a next step, such development may involve the following:

1) ranking according to predictability (P) × consequence (C) situations, like vulnerable road users appearing in front of the ego-vehicle, threatening vehicles from various directions and prioritization of other traffic depending on situations.

2) instructing the AV to go, break or turn depending on the P × Cs.

Additional sensors may be needed to sense vehicles behind, and other objects elsewhere than immediately in front of the AV in the decision algorithms.

Bibliography

[1] SAE J3016, *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*

[2] ISO 26000, *Guidance on social responsibility*

[3] ISO/TR 4804, *Road vehicles — Safety and cybersecurity for automated driving systems — Design, verification and validation*

[4] ISO/PAS 21448:2019, *Road vehicles —Safety of the intended functionality to the list*

[5] Barshi I., Degani A., Loukolupolus L., Mauro R., 2014). Designing procedures for normal and emergency situations. NASA Technical Memorandum #3475.)

[6] Raworth K., Doughnut Economics: Seven Ways to Think Like a 21st-Century Economist. Random House Business, New York, 2017

[7] Mosier K.L., McCauley S.T., *Achieving coherence: Meeting new cognitive demands in technological systems.* In: Adaptive perspectives on human-technology interaction, (Kirlik A., ed.). Oxford University Press, New York, NY, 2006, pp. 163–74.

[8] Brunswik E.1943). , Organismic achievement and environmental probability. Psychological Review, **50**, pp255–272.

[9] Kant I., 2015). Critique of Practical Reason. Translated by Mary Gregor. Cambridge University Press.

[10] Rawls J., 1971) A Theory of Justice. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

[11] King S.B.2006). , An engaged Buddhist response to John Rawls's the law of peoples. Journal of Religious Ethics **34**(4). Pp. 637 – 661 (DOI: 10.1111/j.1467-9795.2006.00288.x)

[12] Buckley C., 2021). How Roundabouts Help Lower Carbon Emissions. [1)]› 2021/11/20

[13] Heymann M., Degani A. 2019). Autonomous vehicle interactions with other road users: Conflicts and resolutions. *Proceedings of the 10th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, June 24-27, Santa Fe, AZ.

[15] Automated Vehicle Safety Consortium, 2020). Best Practice for First Responder Interactions with Fleet-Managed Automated Driving System-Dedicated Vehicles (ADS-DVs). SAE Industry Technologies Consortium.

[16] Klein M., 2005). "responsibility". In Honderich, Ted. Ed. Oxford Companion to Philosophy.

[17] Eshleman A. 2009). "Moral responsibility" in *The Stanford Encyclopaedia of Philosophy (Winter 2009 Edition),* ed. E.N. Zalta. Available at http://plato.stanford.edu/archives/win2009/entries/moral-responsibility

[18] a man, being just as hungry as thirsty, and placed in between food and drink, must necessarily remain where he is and starve to death.— *Aristotle, On the Heavens 295b, c. 350 BC*

[19] D Schüll N. D. 2012) *Addiction by Design: Machine Gambling in Las Vegas*, Princeton University Press. http://www.jstor.org/stable/j.ctt12f4d0

1) www.nytimes.com

[20] Floridi L., Cowls J., Beltrametti M., Chatila R., Chazerand P., Dignum V. et al., AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds Mach. 2018, **28** (4) pp. 689–707

[21] Beauchamp T.L., Childress J.F., Principles of biomedical ethics. 5th. Oxford University Press, New York, 2001

[22] SAE J2990, *Hybrid and EV, First and Second Responder Recommended Practice*

[23] Mori, 1972 Mori, M. (1974/1981) The Buddha in the Robot: A Robot Engineer's Thoughts on Science and Religion. Terry, C.S. (Translator). Kosei Publishing Company

[24] Levinas E., 1968/2019). Nine Talmudic Readings. Indiana University Press,

[25] Michon J.A., 1985). *A Critical View of Driver Behavior Models: What Do We Know, What Should We Do?.* In: Evans L., Schwing R.C., (eds) Human Behavior and Traffic Safety. Springer, Boston, MA. https://doi.org/10.1007/978-1-4613-2173-6_19

[26] Ulbrich S., Maurer M., 2013). Probabilistic online POMDP decision making for lane changes in fully automated driving. 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). pp. 2063-2067.

[27] Amersbach C., Winner H.2017). , Functional Decomposition: An Approach to Reduce the Approval Effort for Highly Automated Driving.

[28] Grossi, D., Royakkers, L. & Dignum, F. Organizational structure and responsibility. *Artif Intell Law* **15**, 223–249 (2007). https://doi.org/10.1007/s10506-007-9054-0

[29] Flemisch, F. & Winner, H. & Bruder, R. & Bengler, K.. (2016). Cooperative Guidance, Control, and Automation. 10.1007/978-3-319-12352-3_58.

[30] Danaher, John (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. Philosophy and Technology 29 (3):245-268.

[31] Foot, P. (1967). The Problem of Abortion and the Doctrine of the Double Effect. Oxford Review, 5, p.1-7.

ICS 03.220.20

Price based on 44 pages