Admission Certification for Digital Pilots in Public Traffic

Bernhard Nessler[1,2*], **Lukas Fischer**[1], **Jorge Martinez-Gil**[1], **Roxana-Maria Holom**[3], **Karoly Bosa**[3], **Karl-Heinz Kastner**[3], **Eva Tatschl-Unterberger**[4], **Hannes Watzinger**[4], **Johannes Weissenböck**[5], **Thomas Doms**[5], **Christoph Schwald**[5]

[1]Software Competence Center Hagenberg GmbH (SCCH), [2]Johannes Kepler University Linz, Institute for Machine Learning, [3]RISC Software GmbH, [4]DigiTrans GmbH, [5]TÜV Austria

{bernhard.nessler, lukas.fischer, jorge.martinez-gil}@scch.at, {eva.tatschl-unterberger, hannes.watzinger}@digitrans.expert, {roxana.holom, karl-heinz.kastner, karoly.bosa}@risc-software.at, {johannes.weissenboeck, thomas.doms, christoph.schwald}@tuv.at

Abstract

Up to this day the practical economical application of automated driving stays behind the high expectations that were raised by the great successes of Deep Learning. There is no doubt that fully autonomous (level 5 automated) driving in all weather and road conditions is a much more challenging problem then initially expected. Yet there is another obstacle that hinders developers and users to exploit the existing state of the art for applications with limited scope and capabilities and this is a complete lack of viable procedures for the admission of autonomous vehicles (AVs) without safety driver to any contact with the public traffic. Arguably this lack itself is rooted in the skepticism of the society and the missing technical competences in the public administrations.

In this paper we propose and advocate to install a comprehensive independent testing and certification procedure for the admission of digital pilots without safety driver for specific use cases in limited conditions. The successfully certified application - this is the promise of the process - is then finally approved for public road traffic even without a safety driver in the vehicle.

It is only by providing this clear perspective of a driverless application that makes the development of specific use cases economically feasible.

1 Introduction

The digitalisation and automated driving assistance undoubtedly have already contributed to making the public traffic safer and more efficient than it has ever been. Automated navigation with real-time information about the current traffic on the route and the various assistance technologies relieve the driver from many worries and tasks. Nevertheless, one of the most essential goals of this development, the driverless automated vehicle is still lagging far behind the high expectations that have been propagated and proclaimed during the last decade. In fact, the possibility of sending AVs around on public roads without the need of a human driver is the key to the expected transformation of the public traffic and a transformation of the way how we organize the transportation of both human and goods in the 21st century. The famous "last mile"-problem for deliveries is only one representative example that is waiting to be solved.

Driver assistance systems do allegedly already contribute to safer and more relaxed driving. But it is only the fully autonomous digital pilot that will enable the transformatory power and the high financial savings that are necessary in order to take broad advantage of automated driving in our national and international economies. In addition, sustainability goals, like car-sharing of energy efficient electric cars that automatically go and recharge themselves, are hindered by the lack of a trustworthy digital pilot that safely maneuvers the vehicle in absence of any human safety driver.

Currently, many possible customers that would have valuable use cases for automated driving vehicles are reluctant to invest into that development. We think that this is mainly due to the uncertainty, if and how a working vehicle would be admitted to the public traffic. Governmental administrations have hitherto failed to provide a clear framework and regulatory pathways on how a digital pilot could be tested, validated and finally admitted to the public streets.

We would like to highlight the variety of generalizations needed for the different economically valuable use cases for digital pilots. It may be technically simple to build an automated farming vehicle that can find its way from the farmer's home to the work site, but because that route goes over several kilometers of public road, such a vehicle cannot be sold or operated because driverless vehicles are in general currently not allowable on public roads. Even the possibility that the vehicle encounters pedestrians in the field is currently prohibiting the admission.

Another such example could be a last mile application in rural environments, where villages that have no own station of public transport or limited bus connections that travel just twice a day, could be connected to the next hub with a simple AV that is just traveling back and forth on one route. Both these use cases lose their economical value as soon as a

*Contact Author

safety-driver is required in the AV as per current regulations. It is obvious, that in view of the risk that even a demonstrably working application would not be admitted to a driver-less working mode, such economically valuable use cases are not even developed in the first place.

By definition, both these use cases fall under the category Level 5 automation due too the fully automated driverless operation mode. Obviously, their requirements are far away from the broad generalization that "true" Level 5 driving necessitates. It is even conceivable to apply specific additional road markings, specific new traffic signs or even specific new traffic laws that regulate how humans have to react on visible AVs. This is all comparable to existing public transport regulations, as we do already have, e.g. extra bus lanes, or the right of way for public buses exiting the bus stop.

Thus, we propose to engage in a certification process for the admission of digital pilots to the public traffic in well-specified use cases. We advocate to create a public admission certification process in corporation with a notified body such as the TÜV. It is understood that this certification process has to be authorized by corresponding local and national laws. The admission test itself should then be very similar to a certification audit of any other artificial intelligence (AI) system with the special point that the acceptable level of errors is very low, as we will discuss in the following section.

2 Related Work

2.1 Costs and Harms

The use of AI contributes to safer driving, improving overall traffic flow and thereby reducing infrastructure, air pollution and accident costs. An article from the National Highway Traffic Safety Administration (NHTSA) highlights the following benefits of AVs: NHTSA estimates that human error is the cause of 94% of all serious traffic crashes. As argued critically in [Koopman, 2018], by far not all of these 94% will be avoided by AV and also digital pilots will make mistakes. By defining strict enough criteria for the admission we can create at least the fair expectation that the number of crashes attributed to the digital pilot will be below 20% of that attributed to human drivers. In the following, we thus assume that certified AVs will reduce the 94% by 80%, resulting in a total 75% drop of the accident rate. NHTSA calculated $242 billion (USD) in damages to the U.S. economy caused by crashes in 2010, and another $594 billion in damages caused by deaths or injuries following traffic crashes. In total, traffic crashes in the U.S. cause about $836 billion in damage annually [NHTSA, 2015].

In terms of the Austrian economy, which is equivalent to about 2.34% of the U.S. economy using gross domestic product (GDP) by purchasing power parity (PPP), traffic accidents cause total annual damages of about $19.56 billion (€17.25 billion). The study by [Chen *et al.*, 2019] again only considered accidents in relation to Austrian GDP in its analysis and calculated costs of around $5.36 billion (€4.4 billion). Higher economic costs of around €9.7 billion per year were calculated in a 2017 study by [Sedlacek and Mayer, 2017] for the Federal Ministry for Climate Protection, Environment, Energy, Mobility, Innovation and Technology (BMK), which included human suffering, loss of performance potential and property damage linked to accidents.

Based on the €9.7 billion and with the assumption that AVs can prevent 75% of caused traffic accidents, an annual damage reduction for Austria of about €7.2 billion and for Upper Austria of about €1.2 billion would be possible.

Costs caused by accidents (Billion €)	**US**	**EU**	**AT**	**OÖ**
Based on [NHTSA, 2015]	737.4	538.0	17.25	2.9
Based on [Chen *et al.*, 2019]			4.4	0.7
Based on [Sedlacek and Mayer, 2017]			9.7	1.6

Table 1: Summary of the costs caused by accidents

2.2 RAND Report

The RAND report [Kalra and Paddock, 2016] searched the answer for the following question, which is of public concern: "How many miles would be enough to test AVs before they are allowed on the road for consumer use?" For this, it replied to the following three questions first.

Q1: "How many miles would AVs have to be driven without failure to demonstrate that their failure rate is below some benchmark?" This question was answered by reframing failure rates as reliability rates and using success run statistics based on the binomial distribution [O'Connor and Kleyner, 2012; Kleyner, 2014]. The formula for **Q1** in Table 2 provides a lower bound on the number of failure-free miles, n with the confidence level C, where R stands for reliability and can be interpreted as the probability of not having failure in any given mile.

For instance, to demonstrate that fully AVs have a fatality rate of 1.09 fatalities per 100 million miles (R = 99.9999989%) with a C = 95% confidence level, the vehicles would have to be driven 275 million failure-free miles.

Q2: "How many miles would AVs have to be driven to demonstrate their failure rate to a particular degree of precision?" If the assumed failure rate is λ_* [Mathews, 2010], then **Q2** in Table 2 implies the number of miles that must be driven, where δ is the desired degree of precision and $z_{1-\alpha/2}$ is derived from the precision of failure rate estimate described with the width of $100(1-\alpha)\%$ confidence interval (CI) [DeGroot and Schervish, 2012].

If it is assumed that a fully AV fleet had a true fatality rate of 1.09 per 100 million miles, then based on this information it can be determined, that approximately 8.8 billion miles of driving would be required to estimate the fatality rate of the fleet to within 20% ($\delta = 0,2$) of the assumed rate using a 95% CI.

Q3: "How many miles would AVs have to be driven to demonstrate that their failure rate is statistically significantly lower than the human driver failure rate?" The two equations given for **Q3** in Table 2 determine the required failures (and miles), where λ_0 is the human driver failure rate, λ_{alt} is the

	Applied Eqs.	Outcome
Q1	$n = \frac{\ln(1-C)}{\ln(R)}$	$\geq$ 275 million failure-free miles
Q2	$x = \frac{(\frac{z_{1-\alpha/2}}{\delta})^2}{\lambda_*}$	$\approx$ 8.8 billion miles
Q3	$x = (\lambda_{\text{alt}} \frac{z_{1-\alpha}}{\lambda_0 - \lambda_{\text{alt}}})^2$ $n = \lambda_{\text{alt}} (\frac{z_{1-\alpha}}{\lambda_0 - \lambda_{\text{alt}}})^2$	$\approx$ 5 billion miles
$\mathbf{Q_{orig}}$	$n = \lambda_{\text{alt}} (\frac{z_{1-\alpha} + z_{1-\beta}}{\lambda_0 - \lambda_{\text{alt}}})^2$	$\approx$ 11 billion miles

Table 2: Summary of approaches applied by the RAND report and their derived findings

AV failure rate and α is the significance level, with which the null hypothesis $\lambda \geq \lambda_0$ should be accepted or rejected.

This means, if one supposes a fully AV fleet had a true fatality rate that was A = 20% lower than the human driver fatality rate (per 100 million miles), then it would take approximately 5 billion miles to demonstrate that the assumed difference is statistically significant with 95% confidence.

$\mathbf{Q_{orig}}$: From the findings above one can derive how many miles AVs need to be driven to perform better than human drivers with some probability (see the normal approximation for the distribution of fatalities for $\mathbf{Q_{orig}}$ in Table 2).

Accordingly, AVs must be driven more than 11 billion miles to determine with 95% confidence and 80% power (i.e., β = 0.2) that their failure rate is 20% better than the human driver fatality rate. With a fleet of 100 AVs being test-driven 24h a day, 365 days a year at an average speed of 25 miles per hour, this would take 518 years.

In order to make these numbers of test cases viable we will stick to the statistical method of the RAND test but apply it only to virtual tests, i.e. to simulated testing. Note, that the statistical validity for the approval of a specific AV with a specific digital pilot software is only valid as long a the software is not changed during the batch of the test. This fact was a problem for the practicability of this RAND test in the real world, but it is no problem in simulations. And it is understood that we will aim for a compression of the statistics by focusing on critical interactions and corner cases, in order to reduce the amount of simulation time by some orders of magnitude [Els, 2018], a compression method that could also not be applied in reality.

Of course it finally is up to the law makers who will decide about the target values for performance and this will determine the corresponding test length. Such decisions might or might not be based on statistics similar to those used in the RAND report.

2.3 Simulations and Reality Gap

As already mentioned, the 11 billion miles of the RAND Report show that acquiring a valid test statistics in reality is way too expensive. The obvious solution is virtually generated input data for the AVs' sensors, which various traffic simulators can generate. There are commercial products in this area, such as Carmaker (CM) and Virtual Test Drive (VTD), and open source solutions, such as CARLA, Microsoft AirSim, VisSim, CarSim, Gazebo, TORCS, Udacity simulator, and AutonoVi-sim.

The issue with simulation is that it is never equivalent to the reality, hence yielding a reality gap. Machine Learning (ML) models that have been learned using only virtual data usually have major problems when applied in practice. There are differences between the simulated and real environments in visual and physical properties [Bousmalis *et al.*, 2017]. In order to use these simulations for testing AVs this "reality gap" has to be closed. For learning various methods have been proposed to overcome the reality gap, as e.g. domain randomization [Tobin *et al.*, 2017; Borrego *et al.*, 2018; Tremblay *et al.*, 2018]. In this process, all objects used are randomly generated, thereby changing number, shape, position, direction, texture, field of view, lights and noise. In this way, the trained model integrates a wide range of realizations of the environment to which the real world can be added as realization example. Domain randomization is mainly applied in basic object recognition. [Reway *et al.*, 2020] also present a method to measure the reality gap of object detection which is a part of the problem. Note that the validity of the simulations with respect to the reality gap, will have to be be assessed case by case by the expert audit.

2.4 AI Certification & Standards

According to [Winter *et al.*, 2021] the certification process is usually represented as a circular process where it is necessary to pass through a series of stages. As a rule, it is checked whether certain requirements (e.g. technical standards) have been met. An external and independent authority verifies conformity with these conditions during the testing process. Both descriptive and analytical standards are considered during the evaluation. It is worth noting that in most certification processes, confidence in the expertise and sound judgments of the assessors is critical. It is often assumed that simulations alone, whether virtual or physical, are not sufficient in most of cases, so the vehicle and all data generated must be carefully reviewed and checked against the data associated with hazardous situations. Examples include the development of techniques to analyze risks and ensure data privacy against attacks or electronic measures that are not as expected. Therefore, the authors of [Winter *et al.*, 2021] state that the certification is only possible when performing white box testing. It is part of the certification process to get insight into how the AI works in order to conduct a thorough evaluation of it. A final certificate is then valid for three years, for example, after which re-certification is required.

In our specific context, the safety of human drivers serves as a vital benchmark against which AV can be measured. And we note that despite the significant total number of crashes, injuries, and fatalities caused by human drivers, the incidence rate of these failures is minimal when it is related to the number of kilometers driven. So that benchmark is quite a challenge for a digital pilot.

With the aim to build a focal point for AI standardization in the broad field of AI, ISO and the International Elec-

trotechnical Commission (IEC) have set up a joint technical committee ISO/IEC JTC 1, Information technology, subcommittee SC 42, Artificial intelligence. Among its many mandates, experts are investigating different approaches to establish trust in AI systems, e.g. ISO/IEC JTC 1/SC 42/WG3 standards [ISO, 2020], ISO/AWI PAS 8800 (under development) [ISO, 2022]. A comprehensive overview of standardization committees and respective work programs on AI certification can be found in [Winter *et al.*, 2021].

3 Methods

3.1 Dimensions for Trustworthy Autonomous Vehicles

The High-Level Expert Group on Artificial Intelligence (AI HLEG) has formulated in its Ethics Guidelines for Trustworthy AI [Commission *et al.*, 2019] the conditions whether an AI system being developed, deployed, or procured can be categorized as a trustworthy AI. The later Assessment List for Trustworthy AI (ALTAI) [HLEG, 2020] derives seven key requirements composed of multiple criteria.

A more recent report from the European Commission [Commission *et al.*, 2021] continues this work and sets a basis for assessing trustworthy AI in the autonomous driving domain by translating the seven key requirements to the context of AV. The goal of the report is to progress towards a general AI trustworthiness assessment framework for AVs. The authors mention the next steps the Commission should do such that the Member States take advantage of the good opportunities to develop new harmonized AV type-approval frameworks. The report describes the main aspects of the autonomous domain and provides an exhaustive analysis of the seven key requirements. There is also an evaluation of the relevance and urgency of each criterion in the context of the AV domain for all types of vehicles, automated (up to SAE Level 3)) and highly automated or AVs (at least SAE Level 4): *critical* in the short term, *important* in the mid term and *impact* in the long term. Although this is a much needed analysis, the provided relevance table is useful during development and internal testing phase, but cannot be considered during the certification process. In this last phase, all key requirements and criteria have to be assessed.

Our goal is to create an assessment framework for fully autonomous digital pilots such that the trustworthiness of that AI-system can be assessed in a holistic way. Therefore, we have evaluated the seven key requirements in this context and split them into two categories. Firstly, we have defined several dimensions which are required to assess each digital pilot individually. Secondly, we have considered the requirements that can be assessed through generic regulatory and standardization processes. Following, we start by describing each dimension and show their relation to the criteria defined by AI HLEG [HLEG, 2020]:

- **Security:** This is one of the main criteria of the key requirement *Technical robustness and safety*, which is mainly linked to the principle of prevention of harm. By *security* we also address the *resilience to attack* criterion, which means testing the exposure of the AVs to threats such as technical faults and defects, outages and potential cyber attacks.
- **Functional Correctness:** This is the main test if the digital pilot correctly respects the functional requirements. This comprises *general safety* with the scope of improving safety of all road actors and its environment. The difficult problem of defining holistic accuracy metrics has to be done once for the overall admission process. The individual digital pilot only has to be tested for its functional correctness versus those predefined metrics. Moreover, the functional correctness comprises the individually defined *reliability* of the digital pilot under test and the *fallback plans* and the *reproducibility* of the behavior, respectively.
- **Traceability:** This dimension asks for information over the process of designing, training, testing, validating and applying of AI algorithms to allow checking their actions and examine the methods by which they have been taken.
- **Explainability:** This dimension assists the empowered user in comprehending how the pieces of information are used and how the vehicle decides. Additionally, this testing dimension is intended to support the approval process in assessing, e.g., the remaining reality gap and gaining the trust in the intended generalizations.
- **Accountability:** This dimension can be put in place by addressing the two sub-criteria: *auditability* and *risk management*. The first one deals with assessing the algorithms, data and design process and therefore, it requires that a digital pilot incorporates mechanisms that allow experts to distill the insights of the operations. An organizational framework and policies have to allow internal, but also independent external auditors to clearly identify which piece of information contributed to which piloting action. The second criterion, risk management, addresses the in-depth analysis of the inter-operation of the AV with its environment in the specific intended use case, such that the negative impacts resulted from the use case are identified and tackled. The dimension of accountability has to make sure that a transparent traceability and auditability are granted for all possible risks.
- **Impartiality:** This dimension addresses the subcriterion *avoidance of unfair bias* and and asks for both internal and external checks to aid in making fair decisions and avoid any discriminatory bias.
- **Data Governance**: These criteria refer at respecting *data privacy* and to not use the vehicle for anything other than what it was designed for.

The following requirements and criteria provided by the AI HLEG are not necessarily to be specifically assessed for each digital pilot, but through generic regulatory and standardized interfaces.

- *Fall-back plans* and *reproducibility* are strongly linked to *reliability* and *traceability* and can be addressed by the specific checking actions.

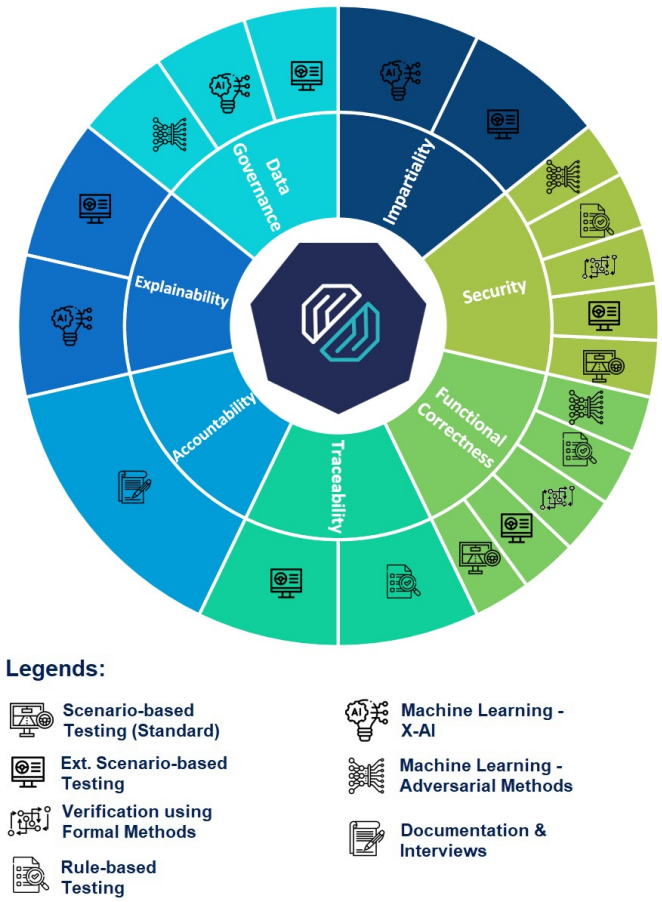

Figure 1: Representation of trustworthiness dimensions for digital pilots (inner circle segments) with associated existing solution approaches (outer circle segments).

- *Communication* is an important aspect for the acceptance of AVs and therefore, it takes a central place in the design of AVs. The design of the communication interfaces is challenging, because of the multi-user perspective. There exist two categories of users: the users inside the vehicle, which can be divided based on the automation Level (Level 1 and 2 - assisted driver, Level 3 - assistant/backup driver, or Levels 4 and 5 - passenger with no driving responsibility), and the external road users. In our target domain of fully automated digital pilots, the communication with a passenger is minimal as compared to other automation levels. The intervention of the passenger is only required for some strategic tasks (e.g. specification of the destination, requesting a stop). There are no shared responsibilities and no handovers have to be defined. The correct communication with the external road user is part of the functional correctness dimension. In the case where the empowered user must understand the decisions/actions of the AV, *explainability* and *traceability* (through data logging) represent the basis for the car-user interaction.

- *Human agency and oversight*: These requirements address the principle of respect for human autonomy and traditionally, the sense of human agency is depending on the level of automation. That is, at a fully automated digital pilot, only specially empowered users (e.g. administrative staff over a remote link) are allowed to override the automation. The passengers are restricted to something like pushing the emergency button and waiting for a respective reaction. Such human oversight principles have to be defined by regulations for fully autonomous (public) digital pilots in general and must therefore not be assessed at each digital pilot individually, except for the functional correctness of the implementation of such rules.

- *Societal and environmental well-being*: This is a user-oriented factor that means that regulations are needed to define the usage of AVs and the potential environmental impact (e.g. emissions, traffic jams, pollution, etc.).

- *Accessibility and Universal Design*: This sub-criterion is part of the key requirement *diversity, non-discrimination and fairness*. The assessment process is following in this case general regulations which are specific for the universal design and indicated by the relevant bodies.

- *Stakeholder Participation*: The list of stakeholders in the AVs domain is extensive and therefore, a clear taxonomy is required. The assessment process follows in this case the procedures formed by public and non-governmental institutions.

3.2 Existing Solution Approaches

For the assessment of the seven dimensions described in the previous section, we have defined six test methods, with which each digital pilot must be evaluated individually. Figure 1 shows the assignment of the trustworthiness dimensions (inner circle) to the test methods (outer circle). It should be noted that none of the methods is sufficient in itself to perform a holistic assessment. The methods include:

Scenario-based testing (standard) uses virtual simulation environments to check safety and reliability according to the requirements for trustworthy AI. In doing so, AVs must pass a variety of different test scenarios, such as lane changes, turns, or overtaking, as far as those are necessary to fulfill the needs of the targeted use-case. With ML supported **mutation tests**, AVs are targeted to their system limits to provoke failures and score the systems with so-called mutation scores. A major difficulty in scenario-based testing is the test case definition and generation, since it is currently not possible to verify that all scenarios for safe operation are present in the test catalog. We expect that the limitation to the specific use-cases simplifies that process significantly.

Extended Scenario-based testing extends the standard process through various methods. Examples include root cause analysis, formal rule-based testing, logging & documentation, automatic generation of scenarios with ML, and a fairness pipeline that supports detection and mitigation of bias and explanation of metrics in data sets.

Verification based on formal methods can be used to achieve completeness, while testing can only prove the presence of errors, but not their absence. Formal verification relies on mathematical models to prove or disprove specific specifications and properties. What differentiates it from testing is that the formal verification method is able to find an erroneous state in the system if it exists, and if so it can be demonstrated that this is an issue in the system model. Using semi-formal verification together with testing is also an amenable approach in dealing with ML systems. The goal of these methods is to find corner cases in ML algorithms which may lead the system to an unsafe state.

Rule-based verification is the tool for the verification of applicable laws and traffic rules. Several solutions are proposing the use of rule-based systems as a basis approach to translate traffic rules to corresponding formats that can be further used to check the behavior of driving models. [Deng *et al.*, 2020] detail in their paper a declarative, rule-based metamorphic testing approach called RMT that enables domain experts to specify custom rules derived from real-world traffic rules using a domain-specific language. [Xiao *et al.*, 2021] propose a rule-based approach for transforming traffic laws and other driving rules to formal rules that have a priority structure. Our method of testing is intended to refine these approaches and ensure that the digital pilot emerges unambiguously as the innocent road user in simulated accidents according to the current law. This should allow for a sufficiently fluid, efficient and dynamic driving behavior, in contrast to an over-restrictive traffic rule interpretation.

Machine Learning – XAI is a set of tools and frameworks that help to understand, comprehend and interpret the predictions of ML models. It allows to debug or improve the model performance and thereby helps humans to understand its behavior in a systematical and interpretable manner. [Samek *et al.*, 2019]

Machine Learning – Adversarial methods Adversarial ML aims to exploit models by creating hostile situations utilizing accessible model information. The most common reason for this is to cause a ML model to fail. The vast majority of ML algorithms were designed to work on specific problem sets in which the training and test data were drawn from the same statistical distribution. When such models are applied in the real world, adversaries may produce data contradicting that statistical assumption. This data could be altered to exploit vulnerabilities and compromise results. Evasion, poisoning, model extraction, and inference are the four most common adversarial ML strategies.

Because of the complexity of an autonomous system, the process of ensuring its trustworthiness is considered to be an interdisciplinary challenge [Rajabli *et al.*, 2020]. As mentioned before, none of the methods is sufficient for assessing the digital pilot, if considered separately (e.g.: Formal methods and rule-based approaches as stand-alone solutions are only applicable at component level and not at system level). Therefore, all methods mentioned before are specifically guided by the team of auditors. They do so not just by following standard methods, but they have to be aware of the current state of the art in the relevant technical and scientific fields and do their best in order to challenge the digital pilot and identify its weaknesses. They have to assess every application individually and finally give an informed judgment about its trustworthiness.

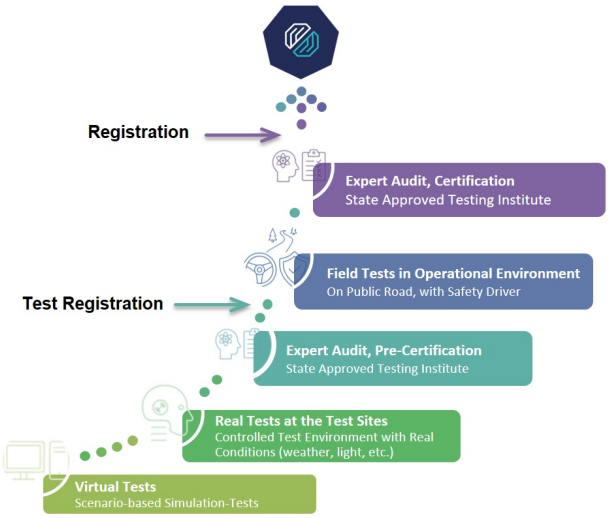

Figure 2: Sketch of the steps required to verify trustworthy AI in the case of autonomous driving.

4 Results

For the verification of trustworthy AI in autonomous driving, a concept of a holistic test process is proposed. This test process consists of five main steps: Virtual Testing, Real Testing, Expert Audit & Pre-Certification, Field Testing in Deployment Environment and Final Certification by Expert Audit.

As shown in Figure 2, these five steps form a sequence similar to a sequence of quality gates on the way to a certificate of trustworthiness. This means that, for example, virtual tests and real tests at the test site can be regarded as prerequisites for the field tests or the final certification step.

In the following, the individual quality gates are discussed in more detail.

4.1 Virtual Tests

The first stage of the certification process consists of scenario-based simulation tests. These tests are the basis of the testing suite for the obvious reason that these tests are relatively cheap and non-destructive. Depending on the targeted use-case the intensity of these tests may vary, for very limited use-cases a nearly exhaustive testing of all intended scenarios might even be possible. The stochastic validity, i.e. representative power of these tests is assured by an online random generation of the testing scenarios, which makes training-to-the-test a futile approach. Different sets of test scenarios and, in particular, critical scenarios, the edge cases, which can be extracted from traffic data recordings, serve as the basis for the distribution from which the tests are drawn. These virtual

test are also intended to simulate crashes and accidents of all forms. The automated rule-based testing has then to identify if the digital pilot would legally be to blame for its driving behavior. There are countless situations where accidents cannot be avoided even by an efficiently driving digital pilot. So the test is failed only if the digital pilot is found guilty for the accident. Nevertheless the accident avoidance behavior provides interesting insights into the holistic trustworthiness of the digital pilot.

4.2 Real Tests at the Test Site

The number of testing scenarios that can be played within a testing lab is very limited. Therefore these real tests can never be representative for all scenarios, not even in limited use-cases. The aim of these tests is to verify that the digital pilot exhibits the same reactions in the real tests as it does in the virtual simulation for the same scenarios. Thus, this is the first tool in the chain that is necessary to gain the trust that the reality gap is closed. Any deviation in behavior between the real scenarios and their counterpart in the simulation raises concerns that have to be discussed in the following Expert Audit session for the pre-certification.

4.3 Expert Audit, Pre-Certification

The Expert Audit is where the "natural intelligence" and the know-how in the scientific state-of-the-art of the auditing scientists come into play. The functional correctness was proven in the extensive virtual tests. However, the validity of these tests concerning the reality gap has to be established by expert insight into the modules and the inner working of the digital pilot. Apart from the usual catalog-like audit questions and measurements, the expert has to decide if the internal design and the training procedures led to a robust and trustworthy digital pilot. The pre-certification audit serves as a permission to carry out field tests in the operational environment and also decides about the distribution of test-scenarios that should be staged in the field tests.

4.4 Field Tests in the Operational Environment

The aim of the field tests is to expose the digital pilot to randomly selected scenarios with random environmental conditions in real road traffic and to evaluate recordings and logs of the inner working details of the digital pilot. All road sections that are part of the intended operational environment can be used as a test track. To ensure safety, the vehicles to be tested must be continuously monitored by a safety driver during the field test, who must be able to intervene and take control of the vehicle at any time. If the construction of the vehicle prohibits the transport of persons then the AV in the field test has to be closely observed and controlled with a suitable remote control.

Note that the trustworthiness of the digital pilot has in principle already been demonstrated in the tests prior to the field test. The field test is mainly intended to confirm the hitherto gained results and insights. The field test is not statistically sufficient for a positive conclusion (as we know from section 2.2). But even a single fail or critical deviations from the expected behavior in the field Tests could lead to a rejection of the admission. The assessment what failing or critically deviating precisely means is highly non-trivial and is decided upon deep inspection of the test protocols in the final Expert Audit.

4.5 Expert Audit, Certification

As the final stage in the certification process, the results of the field tests are analyzed and compared with the expectations from the pre-certification audit. This is the ultimate test to establish the trust that the reality gap between the virtually simulated functional tests and the real world behavior is sufficiently closed. Note that the field tests alone do not by themselves serve as a positive proof for the functional correctness as the probability of experiencing difficult edge-cases is minimal according to the RAND report cited in section 2.2 above. This is why the in-depth analysis of the protocols of the field tests and their comparison with the simulated behavior are the most important outcome of the field test.

Based on the pre-certification and the confirmation of the expectations by the assessment of the field test, the holistic evaluation of the system against the seven defined dimensions of trustworthiness is finished.

5 Discussion & Conclusion

In this work we have highlighted the need for a publicly accepted certification procedure for digital pilots. We have specifically advocated to create an admission certification that is suitable to allow automated vehicles to public traffic without safety driver in special restricted use-cases. For every use-case a special assessment and individual tests might have to be designed and eventually even special local traffic signs have to be installed.

In order for such a process to work, we have discussed the necessary testing dimensions and an overall testing scheme that includes AI experts, scientists knowledgeable in the state of the art of AI, whose duty is to intentionally challenge the pilot under test with specific tests until those experts are convinced of the trustworthiness of the device.

6 Acknowledgments

The research reported in this paper has been funded by the State of Upper Austria within the strategic program upperVision2030 (TWN, Certification) and by BMK, BMDW, and the State of Upper Austria in the frame of SCCH, part of the COMET Programme managed by FFG. We thank the projects S3AI (FFG-872172) and ELISE (H2020-ICT-2019-3 ID: 951847).

References

[Borrego *et al.*, 2018] João Borrego, Atabak Dehban, Rui Figueiredo, Plinio Moreno, Alexandre Bernardino, and José Santos-Victor. Applying domain randomization to synthetic data for object category detection, 07 2018.

[Bousmalis *et al.*, 2017] Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, Sergey Levine, and Vincent Vanhoucke.

Using simulation and domain adaptation to improve efficiency of deep robotic grasping. *CoRR*, abs/1709.07857, 2017.

[Chen *et al.*, 2019] Simiao Chen, Michael Kuhn, Klaus Prettner, and David E Bloom. The global macroeconomic burden of road injuries: estimates and projections for 166 countries. *The Lancet Planetary Health*, 3(9):e390–e398, 2019.

[Commission *et al.*, 2019] European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics guidelines for trustworthy AI*. Publications Office, 2019.

[Commission *et al.*, 2021] European Commission, Joint Research Centre, D Fernández Llorca, and E Gómez. *Trustworthy autonomous vehicles : assessment criteria for trustworthy AI in the autonomous driving domain*. Publications Office of the European Union, 2021.

[DeGroot and Schervish, 2012] Morris H DeGroot and Mark J Schervish. *Probability and statistics*. Pearson Education, 2012.

[Deng *et al.*, 2020] Yao Deng, Xi Zheng, Tianyi Zhang, Guannan Lou, Miryung Kim, et al. Rmt: Rule-based metamorphic testing for autonomous driving models. *arXiv preprint arXiv:2012.10672*, 2020.

[Els, 2018] Peter Els. How much testing will prove automated cars are safe?, "https://www.automotive-iq.com/autonomous-drive/articles/", 2018.

[HLEG, 2020] AI HLEG. The assessment list for trustworthy artificial intelligence (altai) for self assessment. https://altai.insight-centre.org/, 2020.

[ISO, 2020] ISO. "iso/iec tr 24028:2020, information technology – artificial intelligence – overview of trustworthiness in artificial intelligence". https://www.iso.org/standard/77608.html, 2020.

[ISO, 2022] ISO. "iso/awi pas 8800, road vehicles — safety and artificial intelligence (under development)". https://www.iso.org/standard/83303.html, 2022.

[Kalra and Paddock, 2016] Nidhi Kalra and Susan M. Paddock. *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* RAND Corporation, Santa Monica, CA, 2016.

[Kleyner, 2014] Andre Kleyner. How stress variance in the automotive environment will affect a 'true'value of the reliability demonstrated by accelerated testing. *SAE International Journal of Passenger Cars-Electronic and Electrical Systems*, 7(2014-01-0722):552–559, 2014.

[Koopman, 2018] Phil Koopman. A reality check on the 94 percent human error statistic for automated cars, "http://safeautonomy.blogspot.com", 2018.

[Mathews, 2010] P. Mathews. *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Mathews Malnar and bailey, Incorporated, 2010.

[NHTSA, 2015] NHTSA. The economic and societal impact of motor vehicle crashes, 2010 (revised)1. *Annals of Emergency Medicine*, 66(2):194–196, 2015.

[O'Connor and Kleyner, 2012] Patrick O'Connor and Andre Kleyner. *Practical reliability engineering*. John Wiley & Sons, 2012.

[Rajabli *et al.*, 2020] Nijat Rajabli, Francesco Flammini, Roberto Nardone, and Valeria Vittorini. Software verification and validation of safe autonomous cars: A systematic literature review. *IEEE Access*, 9:4797–4819, 2020.

[Reway *et al.*, 2020] Fabio Reway, Abdul Hoffmann, Diogo Wachtel, Werner Huber, Alois Knoll, and Eduardo Ribeiro. Test method for measuring the simulation-to-reality gap of camera-based object detection algorithms for autonomous driving. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1249–1256, 2020.

[Samek *et al.*, 2019] W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, and K.R. Müller. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Lecture Notes in Computer Science. Springer International Publishing, 2019.

[Sedlacek and Mayer, 2017] I. Sedlacek, N.and Steinacher and A. Mayer, B.and Aschenbrenner. Unfallkostenrechnung straße 2017. https://www.bmk.gv.at/themen/verkehr/strasse/verkehrssicherheit/unfaelle/ukr2017.html, 2017. Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK), Wien.

[Tobin *et al.*, 2017] Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *CoRR*, abs/1703.06907, 2017.

[Tremblay *et al.*, 2018] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization, 2018. cite arxiv:1804.06516Comment: CVPR 2018 Workshop on Autonomous Driving.

[Winter *et al.*, 2021] Philip Matthias Winter, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, and Bernhard Nessler. Trusted artificial intelligence: Towards certification of machine learning applications. *arXiv preprint arXiv:2103.16910*, 2021.

[Xiao *et al.*, 2021] Wei Xiao, Noushin Mehdipour, Anne Collin, Amitai Y Bin-Nun, Emilio Frazzoli, Radboud Duintjer Tebbens, and Calin Belta. Rule-based optimal control for autonomous driving. In *Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems*, pages 143–154, 2021.