

THE | AUTONOMOUS

Chapter Event Safety & Artificial Intelligence

co-hosted by

FIVE
AI

EXECUTIVE SUMMARY

On June 5th, 2020, The Autonomous together with Five hosted a virtual Chapter Event on “Safety & Artificial Intelligence (AI)”. The event featured six presentations and two panel discussions. A moderator managed the interaction between the audience and the speakers. The audience participated by submitting numerous questions that were answered by the presenters and by completing a post-event survey. The event focused on two main topics: (i) Neural Network Verification and Validation and (ii) Assuring Safety in AI components. This report summarizes the presentations, panel discussions, the Q&A, and the results of the post-event survey.

Focus I: Neural Network Verification and Validation

The topic aimed to answer the following questions: “How can we perform effective validation of AI components?” as well as “How can we determine that a testing set appropriately covers the ODD?” Three high-quality keynotes from industry and academia were presented on this matter, and 14 highly relevant questions were thoroughly discussed - some of which are:

- What is the most promising technique for verification of AI-based systems?
- What are examples of performance measures for AI, especially in the case of AV?
- What can we learn from the way that Silicon Valley approaches V&V?

Furthermore, a post-event survey resulted in the following data:

- When it comes to ranking the techniques for V&V of AI components, formal verification takes the first place with 35% of the votes, followed by automated validation of newly collected data with 30%, precision/recall with 20%, and introspective coverage with 15%.
- 82% consider adversarial attacks as important to worry about in practice, whereas 9% do not.
- Concerning V&V of AI, participants think that current standards lack in defining (i) measurable performance indicators, (ii) systematic approaches for incremental coverage of ODD, (iii) and suffer from being too generic – hence not addressing the characteristics of technologies used in V&V.

Focus II: Assuring Safety in AI Components

“What are the approaches for assuring safety in AI components?” was the main question addressed in this session. Once again, three high-quality keynotes were presented, and 11 technical questions were passionately discussed – a portion of which are:

- How could the baseline for safety be defined?
- How can we reliably detect the “edges” of ODDs, especially where they include weather and previously unknown events in the driving domain?
- Are there any approaches to a better understanding of how the AI-based algorithms work?

The post-event survey resulted in the following data:

- Participants concluded that the main challenges in assuring the safety of AI components are: measuring its integrity, ensuring its deterministic behavior, and sufficient explainability.
- Concerning scenario-based testing, 59% believe that it should be the dominant method for testing an AD system. Whereas 36% do not think so, and 5% have provided no answer.
- Experts advised the following for incorporating an AI component into a safety-critical system:
 - Limiting the use of AI to essential functional areas only;
 - Designing safety mechanisms to monitor the AI component at runtime (i.e., on public roads);
 - And others like: use of diverse implementation and measuring its functional integrity.

*“Solving the safety challenges in AI and Automated Driving requires **further serious research and honest work** (away from any populist marketing disruption).*

--Quote from a participant

BACKGROUND AND EVENT DETAILS

The Initiative

For all actors involved in the development of autonomous mobility solutions, who position safety as a fundamental value of their products - **The Autonomous is a knowledge ecosystem** - that generates new knowledge and technological solutions to **tackle key safety challenges** that shape the future of safe autonomous mobility. Complementary to standardization organizations that establish uniform engineering or technical criteria, methods, and processes, The Autonomous will develop **Global Reference Solutions** for autonomous mobility that conform to relevant standards and facilitate the adoption of these solutions on a grand scale. The benefits The Autonomous will provide to the partners of the ecosystem are:

- Development of safe and best-in-class AD solutions thanks to the wisdom of the crowd;
- Reduction of potential product liability risk by (i) tightly working with government and regulatory institutions and (ii) developing common basis for regulatory bodies;
- Reduction of development costs by (i) developing modular and reusable Global Reference Solutions and (ii) sharing the development efforts;
- Reduction of risk of wrong development by joint definition of state-of-the-art and state-of-practice;
- Accelerating the learning curve by collectively learning from individual failures and field observations.

Towards this vision, in 2020, The Autonomous is hosting a series of workshops - **“The Autonomous Chapter Events”** - to facilitate discussions among experts and take the first steps towards the targeted Global Reference Solutions. The second Chapter Event titled **“Safety & Artificial Intelligence”** was hosted by The Autonomous, together with **Five**.

Event Details

Presentations

Focus I: Neural Network Verification and Validation

- Dimensions of AI Systems Validation | David Hand | Imperial College London
- Characterize Your Perception to Simulate for Safety | Iain Whiteside | Five
- Modular Verification of (Non-modular) AI-based Systems | Yoav Hollander | Foretellix

Focus II: Assuring Safety in AI Components

- Assuring the Safety of AI-based Autonomous Driving | Simon Burton | Bosch & University of York
- Safety Cases and Safety Performance Indicators for AI Driven Vehicles | Mike Wagner | ECR
- AI Safety from the Perspective of the DARPA Assured Autonomy Program | Sandeep Neema | DAPRA

Event Statistics

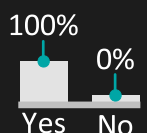
Facts

- 413 event registrations
- 249 different companies attended
- Livestream:
 - 374 unique views
 - 250 concurrent viewers
- 25 questions thoroughly discussed

Feedback

- 45 participants provided feedback

- Did the event meet your expectations?



- How would you rate the event? ★★★★★

Technical Report

Chapter Event Safety & AI

Edited by

Iain Whiteside, Ayhan Mehmed

The Autonomous
July, 2020

Contents

| | | | |
|-------------------|----------|--|-----------|
| 1 | | The Initiative | 4 |
| 1.1 | | Vision | 4 |
| 1.2 | | Mission | 5 |
| 1.3 | | Approach | 5 |
| 1.4 | | Roadmap | 6 |
| 2 | | Chapter Event Safety & AI | 8 |
| 2.1 | | Scope and Topics | 8 |
| 2.2 | | Event Statistics | 9 |
| 3 | | Focus I: Neural Network Verification and Validation | 10 |
| 3.1 | | Talk 1: Dimensions of AI Systems Validation | 10 |
| 3.2 | | Talk 2: Characterize Your Perception to Simulate for Safety | 11 |
| 3.3 | | Talk 3: Modular Verification of (Non-modular) AI-based Systems. | 12 |
| 3.4 | | Panel Discussion on Neural Network V&V | 14 |
| 4 | | Focus II: Assuring Safety in AI Components | 18 |
| 4.1 | | Talk 4: Assuring the Safety of AI-based Autonomous Driving - Technical, Management and Governance perspectives | 18 |
| 4.2 | | Talk 5: Safety Cases and Safety Performance Indicators for AI Driven Vehicles. | 19 |
| 4.3 | | Talk 6: AI Safety from the Perspective of the DARPA Assured Autonomy Program | 20 |
| 4.4 | | Panel Discussion on Assuring Safety in AI components | 22 |
| 5 | | Survey Results | 25 |
| 5.1 | | Contributors | 25 |
| 5.2 | | Subject: General AD | 26 |
| 5.3 | | Subject: The Autonomous | 28 |
| 5.4 | | Subject: Neural Network Verification and Validation | 30 |
| 5.5 | | Subject: Assuring safety in AI components. | 33 |
| Appendices | | | 38 |
| A | | List of Abbreviations | 38 |
| B | | Compliance Guidelines | 39 |
| C | | Standard Settings Guideline | 41 |
| D | | Acknowledgments | 45 |
| E | | Feedback | 46 |

1 | The Initiative

As autonomous mobility is moving closer to becoming a reality, safety and trust concerns prove to be the main hurdle in the way of reaching broad acceptance. OEMs and technology suppliers (Tier 1, 2 & 3, and others) cannot overcome the safety challenge and the necessary investment costs with a “go-it-alone” approach. Therefore, the autonomous mobility industry and other relevant institutions need to come together and show significant efforts in prioritizing and ensuring safety on all technological levels, as well as set common technical and legal standards. Towards this, TTTech Auto initiated The Autonomous - an open platform that brings together actors from the autonomous mobility ecosystem to align on relevant safety subjects.

1.1 Vision

*Create a safer, more livable,
and more sustainable future.*

— The Autonomous

For all actors involved in the development of autonomous mobility solutions, who position safety as a fundamental value of their products - **The Autonomous is a knowledge ecosystem** - that generates new knowledge and technological solutions to **tackle key safety challenges** and to shape the future of safe autonomous mobility. Complementary to standardization organizations that establish uniform engineering or technical criteria, methods, and processes, The Autonomous will develop **Global Reference Solutions** for autonomous mobility that conform to relevant standards and facilitate the adoption of these solutions on a grand scale. The benefits The Autonomous will provide to the partners of the ecosystem are:

- Developing safe and best-in-class solutions for Automated Driving (AD) challenges thanks to the wisdom of the crowd;
- Reduction of potential product liability risk by (i) tightly working with government and regulatory institutions and (ii) developing a common basis for regulatory bodies;
- Reduction of development costs by (i) developing modular and reusable Global Reference Solutions and (ii) sharing the development efforts;
- Reduction of risk of wrong development by joint definition of state-of-the-art and state-of-practice;
- Accelerating the learning curve by collectively learning from individual failures and field observations;

Furthermore, the work products of The Autonomous are expected to serve as further input to existing standardization activities and may also result in new standardization projects.

1.2 Mission

Towards the above-defined vision statement, The Autonomous will:

- Provide a diverse and balanced knowledge ecosystem for autonomous mobility;
- Set the stage for open discussions on main technical and architectural questions where controversial approaches can be freely discussed;
- Act as an interface between industry requirements, standardization, regulation bodies, and academic research in safe autonomous mobility. Collectively identify important gaps in the field and focus the efforts;
- Build consensus on major safety solutions within the automotive industry;
- Generate high-quality know-how and Global Reference Solutions compliant to relevant standards in autonomous mobility;
- Facilitate the adoption of the Global Reference Solutions on a grand scale by placing them into applicable standards as solutions compliant to their requirements.

1.3 Approach

Current Approach

The development approach of automotive systems has remained unchanged over many years. Generally speaking, a car manufacturer (OEM) and its suppliers (Tier 1, 2 & 3, and others) cooperate and then compete with other manufacturers in providing better solutions and products (see Figure 1). This approach has worked well for developing standard, well constrained, and deterministic automotive embedded systems like Anti-Lock Braking System (ABS), Engine Control Units (ECU), and others.

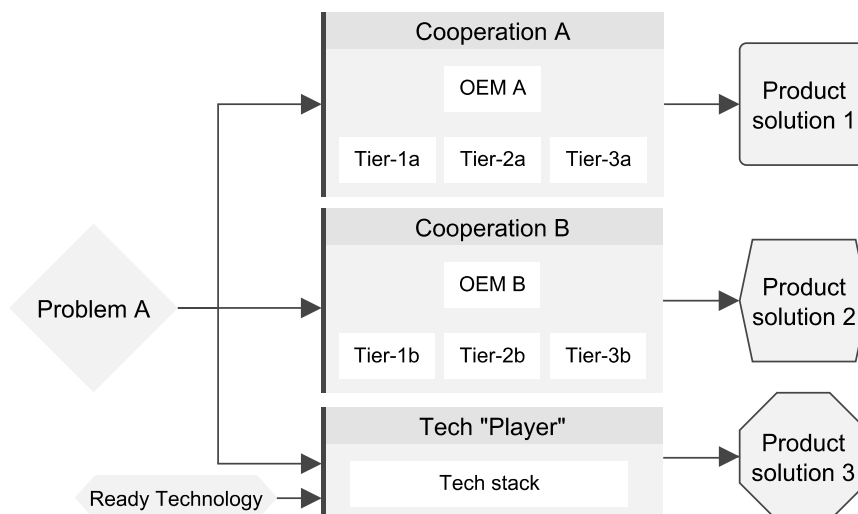


Figure 1: Current development approach of automotive systems.

However, the approach is sub-optimal when it comes to the development of upcoming SAE Level 3 - Level 5 Automated Driving Systems (ADS). The rationale for this is (i) the novelty and high complexity of the AD systems, (ii) the unprecedented high development costs, and (iii) the difficulty in aligning different technical solutions on a common state of the art.

Proposed Approach

To reduce the development cost, a shift from many interdependent cooperation groups (where cooperation groups compete with each other on providing a better solution for a given problem) to a single, broader, and more diverse knowledge ecosystem where partners collaborate towards a single shared goal is necessary (see Figure 2). Such an approach will enable (i) the development of safe and best-in-class products, (ii) an ecological and sustainable development, and (iii) faster development autonomy. Furthermore, in addition to car manufacturers and technology suppliers, The Autonomous also invites stakeholders from governmental, academic, regulatory, and standardization institutions in order to ensure an integrated view.

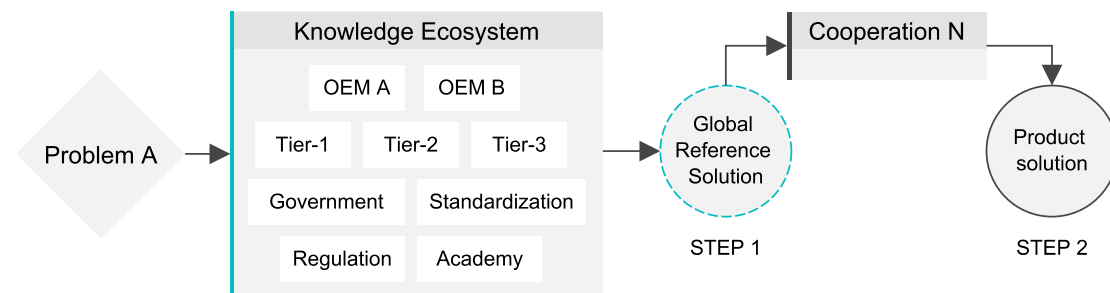


Figure 2: Proposed approach for development of future AD systems.

In “STEP 1” of the proposed approach, the partners of the knowledge ecosystem will work together on Global Reference Solutions that conform to relevant standards. The notion of the Global Reference Solutions is to cover all relevant problems in the development of future AD systems. Hence, more than one reference solution will be available, i.e., ranging from Fail-Operational/Fail-Degraded (FO/FD) architectures to verification and validation (V&V), runtime verification approaches, sensor and sensor fusion configuration, and others. In “STEP 2” of the proposed approach, the partners of the ecosystem will be able to individualize the Global Reference Solution to their needs and therefore keep the competition “alive”.

1.4 Roadmap

In 2020, The Autonomous is organizing a series of *virtual* technical workshops, also known as “The Autonomous Chapter Events”, to facilitate discussions among experts and work towards the target Global Reference Solutions. Figure 3 presents a summary of the Chapter Events planned for 2020. While the scope of the Chapter Events will be

1.4 Roadmap

further broadened by adding other relevant topics, the list below summarizes the current status:

- Chapter Event Safety & Architecture: 2nd of April, 2020 with co-host TTTech Auto;
- Chapter Event Safety & Artificial Intelligence (AI): 5th of June, 2020 with co-host Five;
- Chapter Event Safety & Security: 22nd of June, 2020 with co-hosts Infineon, Secunet, and Integrity Security Solutions;
- Chapter Event Safety & Regulation: 9th of July, 2020 co-hosted with Posser Spieth Wolfers & Partners (PSWP).
- The Autonomous Main Event: 10th of March, 2021 co-hosted with TTTech Auto in Vienna, Austria.

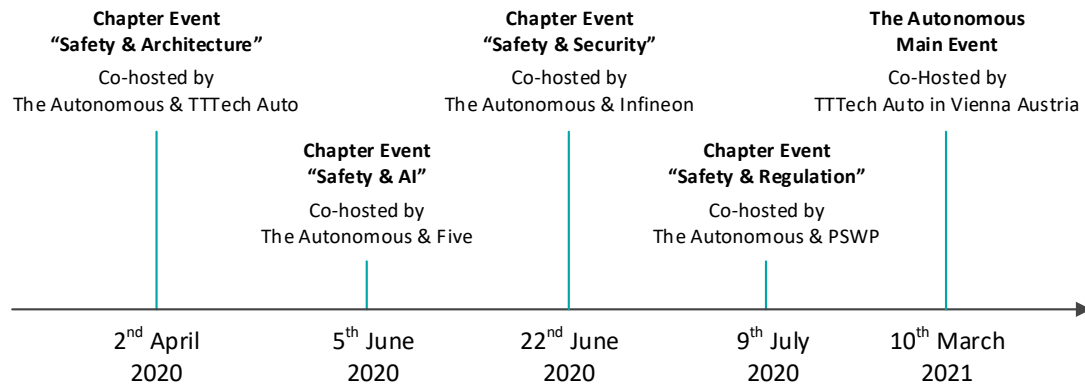


Figure 3: Summary of planned events.

The target outcome of each Chapter Event is a high-quality content summarized in a report. The current report is a summary of the Chapter Event Safety & Artificial Intelligence. The findings of all the reports will be outlined in The Autonomous Main Event on March 10th, 2021.

2 | Chapter Event Safety & AI

2.1 Scope and Topics

This Chapter Event explored the profound challenges associated with demonstrating – through verification, validation, and safety argumentation – that an autonomous vehicle with AI-based components is acceptably safe for operation.

There is a fundamental conflict between traditional verification and validation — where a well-defined specification forms the basis of any safety argument — and, for example, SAE L4 automated vehicles, where the very complexity inherent in the problem — that which necessitates AI-based components in the first place — results in a lack of complete specification.

Entirely new methods of system architecture, implementation, and verification are required. The scale of this challenge is daunting. To make a start, we have focused on several open questions that are considered crucial to safely unlock the true power of autonomy to transform cities worldwide.

- **Focus I: Neural Network Verification and Validation**
 - How can we determine that a testing set appropriately covers the domain?
 - How can we determine that a testing set has appropriate granularity?
 - How can we do effective validation of AI components during real-world testing?
 - How can we come up with appropriate baselines for the performance of AI components?
 - What are the best methodologies for finding unknown hazardous scenarios?
- **Focus II: Assuring Safety in AI Components**
 - Can a hierarchical ontology of detections be used to make better errors?
 - How can sensor fusion be relied upon?
 - How can one use simulations to test a system with perceptual components?
 - How can a planner best act with the uncertainty information given?
 - Are uncertainties provided by a DNN ever usable?
 - Can runtime monitoring be used as a complementary solutions for assuring AI component’s safety?

2.2 Event Statistics

Figure 4 summarizes the facts about the event and the feedback received from the participants. In particular, 413 registrations were made for the virtual event. The participants were from 249 different companies/institutions. The live stream had in total 374 unique views. Throughout the four-hour event, there were 250 concurrent viewers. Last but not least, 96 questions were asked by the audience, of which 25 were addressed (see Section 3 and Section 4 for the summary of answers). Forty-five participants provided feedback after the event, where 100% of them said “yes” when asked whether the event met their expectations. The participants also rated the event with five-and-a-half stars out of six.

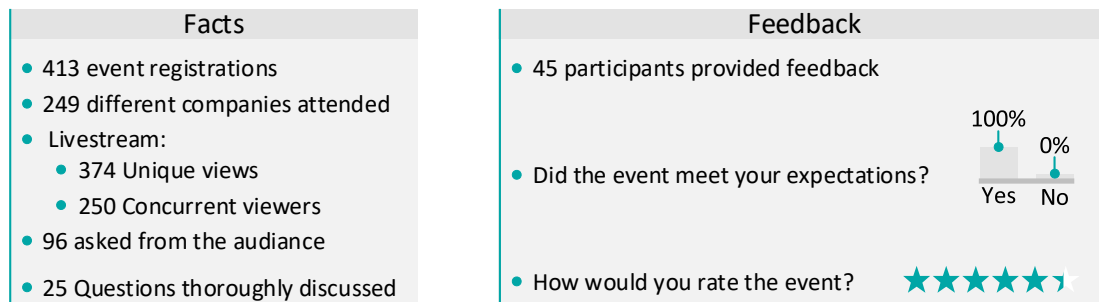


Figure 4: Facts about the event and feedback from participants.

3 | Focus I: Neural Network Verification and Validation

3.1 Talk 1: Dimensions of AI Systems Validation

David Hand

Senior Research Investigator, Imperial College London

Summary

AI systems have the potential to revolutionize society. But this will happen only if the systems can be trusted to do what they are supposed to do, and trust in AI systems can be compromised in several ways. We examine the dimensions in which trust in an AI system can be compromised, ranging from poor high-level specifications of the system's objective to inadequate low-level performance.

Addressed questions

Q1 In your opinion, what is the most promising technique for verification of AI-based systems?

✔ Answer by David Hand

Different techniques are applied in different application domains. A system that has to make rapid decisions on the fly, such as an automated driving system is very different from a system that has to make a medical diagnostic decision. Medical diagnostic decisions might not be so immediately important: there might be even possibly days to make those decisions. I think that the most promising techniques will very much depend on the application domain. I suppose the bottom line is always a comparison in some way of what a system is doing - its output, speed and so on - with the ground truth.

Q2 What is the Mean Time Between Failures (MTBF) an AI system provides today? Which MTBF should it reach to be allowed for commercial use?

✔ Answer by David Hand

Again, it depends on the application domain. If it's a spam filter, or a fraud detection filter in credit card fraud, it will use a different MTBF from an automated driving system or an aeroplane autopilot. Perhaps you should measure the time between failure not in time but in the number of decisions it makes. I think in some sense, the time and number of decisions interact. We might have a system which fails where the meantime between failures is 10 in terms of numbers of decisions it makes – but if it's a slow-acting system, it only has to make one decision per day, and you had 10 days to sort out a problem, that's one thing, but if it's a system that makes a decision every millisecond, that's an entirely different question.

Q3 How do you think that properly representative data could be measured? Or checked by a regulator?

✔ **Answer by David Hand**

This is a question very dear to my heart. I see so many situations where a perfectly good system uses data which is distorted or corrupted in some way, leading the system to fail, and then, of course, the system gets blamed. But that's not appropriate, it is the data collection process where the blame should lie. I think the answer to the question is again "*it depends on the application domain*". In many areas, yes, the regulator should monitor the quality of the data, ensuring the data coming-in from a variety of sources, perhaps is validated. In another case, perhaps especially systems that have to make decisions very rapidly on the fly, there is no role for the regulator in the same sense.

3.2 Talk 2: Characterize Your Perception to Simulate for Safety

Iain Whiteside
Principal Scientist, Five

Summary

We outline some of the major challenges for building safe Automated Vehicles, with a particular focus on those challenges associated with AI components. From those challenges, we outline a new paradigm that we call Understand and Explore that we believe can help solve these challenges by enabling both more efficient V&V of AV and a more detailed understanding of the weaknesses of a given AI component. We briefly describe an approach in this paradigm that we call PRISM and elaborate on the benefits that it provides over traditional approaches.

Addressed questions

Q4 In your opinion, what are the gaps existing in the emerging standards as SOTIF, SaFAD (ISO TR 4804), UL 4600?

✔ **Answer by Iain Whiteside**

That is a good question. Actually, I am a fan of these emerging standards. They are doing a great job of explaining what the basic requirements are. I do think there are potentially a few gaps in terms of how to justify that your nominal behavior is safe. However, in my opinion, the most significant gap I see is the difference between what one has to do versus how to achieve it. There are many gaps in those technical mechanisms to achieve a lot of these goals stated in UL 4600, for example.

Q5 Do you see runtime simulation deployed on the vehicle capable of making a real-time prediction of AI intentions? And, in case of predicted malicious intentions, to have mechanisms in place that switch to a highly safe fail-operational driving mode?

✔ **Answer by Iain Whiteside**

This is definitely an approach, and it is an approach explored by Five in the past, so it is a viable thing to do for making predictions. On the point about malicious intentions, yes. It is not just malicious intentions that you need to model, however. You would be predicting the intentions of vehicles that are following the rules of the road. However, you also have to be able to predict, for example, vehicles making illegal U-turns, that are breaking the road rules. Moreover, as you enforce more stringent safety levels, you need to be able to demonstrate safety against less frequent events, where you are more likely to have to incorporate illegal and malicious maneuvers in your safety case.

Q6 How are we going to distinguish between data sets used to train an AI and data sets used for its verification?

✔ **Answer by Iain Whiteside**

I think it depends on what you mean by "distinguish". You would actually be using data throughout the whole distribution of your training and your testing. But I would say there is also a third category of data that is equally important, which is the sort of data that you get when you are running on the road, since the lifecycle of an AI component in an automated vehicle includes when you put it in the real world and do your validation. This means that it is essential to distinguish between the data that comes from your validation and the data that you have trained and verified your model on. It is this validation data that enables you to understand where there were gaps in your training data. So absolutely, both of those data sets need to match the distribution of your operational design domain and need to be able to be representative. Then, as you collect more data in the field, you revise those assumptions.

3.3 Talk 3: Modular Verification of (Non-modular) AI-based Systems

Yoav Hollander
Founder & CTO, Foretellix

Summary

Complex systems are never completely verified. However, complex systems with a large AI component (such as AVs) are even harder to verify, for the reasons we all know: In particular, the AI part is hard to spec, opaque, and resists modular (as in module-by-module) verification. This talk will try to sketch an approach to overcome some of these issues and visit the (uneasy) meeting point between Machine Learning and rule-based systems.

Addressed questions

Q7 From the theoretical point of view, is safety assurance of AI-based systems (for AVs) even possible? At what cost/time?

✔ **Answer by Yoav Hollander**

I think that in a sense – and it may be disappointing to some – no complex system is ever entirely correct, so complete safety assurance is only possible for extremely simple systems. So in a sense, when talking here, we are talking about an optimization game – how much safety, how many risks can we remove from a given amount of time, using bound resources. This is true for both the AI-based system itself or the machine learning part of the system and the full system. I think the best we can do, and many people feel uncomfortable about it, is to optimize the use of resources and track how well we are doing and then decide at some point that this is good enough.

Q8 What can we learn from the way the silicon industry approaches V&V? What are the similarities but also the differences concerning autonomous vehicles?

✔ **Answer by Yoav Hollander**

So we can learn some things, but we cannot trust this too much. As I have mentioned in the discussion, there is certainly this thing about having lots of bugs and needing a systematic way of getting rid of them. If you remember the Pentium bug, you know that you need to have a systematic way of getting rid of those once you have a complex enough system. Some of the techniques used there can be useful here. Another thing you could observe sociologically is the rise of the verification engineer job as a respectable occupation. However, this other world of autonomous vehicles is, in a sense, much more complex. In a sense, when you think back to David’s presentation, he was talking about the issue of how the objective of unspecified systems is adequately formulated. This is an issue here. Trying to bound only the ML system is hard; trying to specify what the whole system does exactly is a big problem. Trying to specify the pieces is very hard. It is similar and significantly harder.

Q9 How do we ensure that we have the completeness in risk dimensions?

✔ **Answer by Yoav Hollander (1/2)**

There are two interpretations of this question. Number one is “*How do we know that our set of risk dimensions is complete?*”. Number two is “*How do we know for a specific risk dimension whether we have validated it?*”. To both, we have to be cautious that there are no complete answers. About the issue of risk dimension – things will happen in the world, and it will suddenly make us realize that there are more risk dimensions.

✔ **Answer by Yoav Hollander (2/2)**

If you have a tsunami and wonder what an autonomous car would do if it is faced with that, you suddenly have a new risk dimension. You have not thought about this risk previously, so you would add this and hopefully generalize it into a more generic risk dimension. So I assume that the whole set will grow as more as there are accidents. For those more specific risk dimensions, there are methodologies to define covered spaces and go systematically over the known problem space. None of those methodologies are perfect, and the game of optimization is – you get better and better, without ever being done.

3.4 Panel Discussion on Neural Network V&V

Addressed questions

Q10 If we are using data sets to train an AI (in conventional speech, a requirement specification), to also test the performance of the system (verification), given the “performance specification data set” is getting wider and wider, assuming implementation is perfect, is verification becoming redundant?

✔ **Answer by Iain Whiteside, David Hand, Yoav Hollander**

Verification is never redundant: you can never be sure a system is working, and novel situations will always be encountered, and systems will somehow be able to cope with those. There is a trend towards more and more validation of automated driving systems because that is the more straightforward thing to do. After all, you go out and drive in the real world. However, if you work hard, you can attempt to have at least a partial specification of AI components to at least make sure that you can get somehow confidence, even if that specification looks more like a model of the errors that it tends to make, that a planning subsystem needs to be resilient to. As there is always a residual risk, validation still plays an important role. As we mature in the industry, verification will become more and more important. If you have a system that has never malfunctioned and appears to never malfunction, the question is, should you, therefore, be reassured it is perfect under all circumstances, or should you be suspicious, you have not tested it to its boundaries. Additionally, the testing dataset that people use often assumes that we know what the full system should do and, in a sense, especially if you put it inside the bigger system. It is almost impossible to test it using just the inputs to the subsystem because to specify what the bigger system does is hard. Even if you had infinite training sets, you would still not be sure that it does the right thing within the bigger system. It is all an aspect of working out exactly what the question is.

Q11 How can we quantify system robustness against novel and unexpected situations and is there a minimum coverage required before pushing them out into the real world?

✔ **Answer by Yoav Hollander, Iain Whiteside, David Hand**

This issue of the long tail of the novel and unexpected situation is probably one of the most annoying and dangerous for these systems. There is a tendency, I think in regulatory bodies and standards to start thinking in the terms of coverage, which I think is a good idea. It's reasonable that at some point, some regulatory bodies will say okay, here is the kind of coverage you should have, maybe even formalize what it means. Historically, the automotive industry comes out with ever-more-strict standards. So there is the 2022 version of a standard and then there is the 2025 version of a standard which is more strict. So the demand is probably going to be higher and higher as we are going on and learn more. The quality of what is covered is another important dimension, since combinations of the parameters that specify a logical scenario may not all make sense in a given concrete instantiation. There are also similarities to extreme value theory in statistics where one might try to develop analogous models in this area for how often unexpected events are going to happen when something that you have not seen before is going to occur and then try to build a model for how often you think that will occur and so decide where you put your boundary. This is especially important since the majority of verification for AV is moving to simulation and, in order to make that simulation salient, you need to sample from models of real world driving. Since you can't drive enough in the real world, we need mechanisms for sampling the tails of the distributions.

Q12 What are examples of performance measures for AI, especially in the case of AV?

✔ **Answer by Yoav Hollander, Iain Whiteside, David Hand (1/2)**

There are two kinds of performance measures for any system. There are context-specific and high-level system measures, which will include, for instance, how quickly a system will make a decision, how quickly you can update a system when you provide new information or you want to modify it. Then there are low-level system measures, where performance is, for instance, the proportion of correct classifications, the area under the ROC-curve, and precision-recall, for example. These are all aspects of how many decisions the system is making correctly or incorrectly—today's discussion sort of mixes these high-level measures with the low-level. We need to be able to integrate those with the more context-specific high-level aspects of performance. It is also absolutely critical to have a set of measures. Those measures should be specific to the problem domain. For AV, this means understanding how well, for example, your classifier performs with objects close by vs. far away. It also means understanding how much of your training and testing data for ML systems are covering the variety of your ODD e.g., data in fog and rain.

✔ **Answer by Yoav Hollander, Iain Whiteside, David Hand (2/2)**

At the high level of the whole system of which the ML system is a part, we also need to have context-specific performance measures, often called KPIs. KPIs, such as the time-to-collision, or a min/max acceleration (i.e., to have to stay within some) limits are also interconnected: there is no one significant number for each of them.

Q13 Generally speaking, the life of a system is split into two phases: the development phase and the phase when it is already on the public road. We have been talking thus far about approaches applied at the development phase that help us verify as best as we can an AI-based system to ensure that the system is safe when deployed on the public roads. However, we already talked that we cannot verify a complex system 100% at the development phase, so there will always be a residual risk when we deploy the system. What is your opinion on the use of runtime monitoring for the detection of unsafe operation of the AI-based system when being deployed on public roads?

✔ **Answer by Yoav Hollander, Iain Whiteside, David Hand**

If we think back for six months, for instance, the nature of people who have used the roads and how they have used the roads have changed a lot. Our system might have been great in December for the way people are driving. However, when people stopped using the road because of the lockdown and had occasional races because people had the opportunity to race 130 mph through towns, things started to change. The point is that the world is non-stationary. It changes in strange and unexpected ways, and we have to be aware that the system has to respond to these partly unexpected changes, so we have to monitor it continually. Most of the autonomous vehicles manufacturers can have assertions in the software and the ability to log failures of those assertions. However, failures of individual assertions do not mean that the whole thing has failed, but maybe just a subsystem had an issue, and then the rest of the world compensated. They use that to estimate what is bad. They use that to find unusual circumstances when they do offline checks. One thing that has, perhaps surprisingly, proven to be very useful is, as you simulate things offline, you take those assertions as that sort of natural monitoring that happens and use the triggering of those assertions as a way to introduce some surprises in your scenario e.g., have a car come in from the left from out of sight. This can be powerful, since there is this well-known fact that, in many computerized systems, most of the bugs hide in those situations where there is one failure already, which are not well tested.

Q14 Do you see the possible use of digital twins — abstractions models fed with real-time data — for the runtime evaluation of the AI and vehicle behavior?

✔ **Answer by Yoav Hollander, Iain Whiteside, David Hand**

One can certainly see a role for this sort of thing. For instance, NASA did with its moonshots and so on. I think systems like this are already embedded in larger systems so that you can monitor and predict performance and I think it is an excellent idea. Some companies, like Tesla, who have been public about this, have the next version of the software running within the deployed vehicle and are still tested on all the situations while the previous version is running. Moreover, that is not a direct answer to the question, but it is one way to have a new version being tested on a massive scale in actual road situations.

4 | Focus II: Assuring Safety in AI Components

4.1 Talk 4: Assuring the Safety of AI-based Autonomous Driving - Technical, Management and Governance perspectives

Simon Burton

Director Vehicle Systems Safety, Bosch & Honorary Visiting Professor, University of York

Summary

Assuring the safety of autonomous vehicles is a complex endeavor. By this, I do not only mean that it is technically difficult or involves many resource-intensive tasks that must somehow be managed within feasible economic constraints. Both are true. However, autonomous vehicles and their wider socio-technical context demonstrate complex systems' characteristics in the stricter sense of the term. That is, they exhibit emergent behavior, coupled with feedback, non-linearity, and semi-permeable system boundaries. These factors severely limit our ability to apply traditional control measures both at design and operation-time. System complexity also increases the risk of "systemic" failures of the system, which could not have been predicted based on an understanding of failure modes of individual parts of the system. These drivers of complexity are further exacerbated by the introduction of AI and machine learning techniques. The net result is a high level of uneasiness in the traditional safety engineering community regarding whether AI-based autonomous vehicles can ever be argued to be "safe enough".

In this presentation, I present how considering AI-based autonomous vehicles as a complex system could lead us towards better arguments for their overall safety, and ultimately their overall acceptance by society. Understanding the unique challenges that this level of complexity introduces is an important first step towards developing convincing arguments for reducing and managing complexity that increases the risk of systemic failures. Residual inadequacies of individual machine learning components are an inevitable side-effect of the technology. On the other hand, the potential overall safety benefit of autonomous driving is also evident. Therefore, I discuss how mitigations at the technical system layer and the safety management and governance layers, including effective standardization and regulation, are required to ensure an acceptably safe introduction of autonomous driving despite the many uncertainties involved. The presentation builds on the results of an ongoing study, under the auspices of the Royal Academy of Engineering's Safer Complex Systems program which I am currently performing together with a team of colleagues from the University of York.

Addressed questions

Q15 Are the proposed methods for the assurance of safety not also applicable to other system properties, and could we increase the interest and demand in using these methods by “marketing” them under a bigger umbrella?

✔ **Answer by Simon Burton**

The perception of safety can differ depending on the context. In the automotive industry and customer experience, the perception of safety can be generalized as “*the vehicle not operating in an unsafe manner: i.e., operating in a manner that puts the driver into a dangerous situation*”. For classic automotive systems, ensuring that a system’s unsafe operation will not occur has typically required a single view: i.e., Functional Safety (ISO 26262 [1]). However, complex systems such as ADS require a broader view of how safety is achieved: i.e., considering the availability, security, reliability, and others. Therefore, it is more desirable to refer to a broader definition: i.e., dependability.

Q16 There is a common opinion that the safety of autonomous systems should be at least an order of magnitude higher than human drivers. How could that baseline for safety be defined? i.e., are there any established ways to assess the safety of human driving?

✔ **Answer by Simon Burton**

Often the comparison is made to an average human driver. The difficulty is in the common understanding/defining what an average human driver characteristic is and, therefore, what the benchmarks are. For example, is it the average, stressed, distracted, and tired human driver with two kids in the back of the car? Or, is it the average professionally trained police driver, having a really good day? What we are promoting in standardization activities is to try to take a balanced approach. Independent of how good humans are, we set qualitative targets, how good an automated driving system should be. We then use statistical measures to measure how well we have an appropriate argument for that. To conclude, instead of comparing apples with pears, we should be aware of what are chosen statistical targets are actually measuring and how these can contribute to the overall safety argument.

4.2 Talk 5: Safety Cases and Safety Performance Indicators for AI Driven Vehicles

Mike Wagner

Co-founder & CEO, Edge Case Research

Summary

Autonomous mobility relies on cutting edge artificial intelligence to drive safely. Artificial intelligence has yielded tremendous progress: less than twenty years ago, the most advanced autonomous vehicles traveled a little faster than a walking pace. Thanks to

4.3 Talk 6: AI Safety from the Perspective of the DARPA Assured Autonomy Program

tens of billions of dollars of investment, cars can comfortably navigate highways without human intervention. However, are they safe enough? Edge Case Research helped answer this question in April 2020 with the release of UL 4600, the world’s first standard for evaluating autonomous products. Our presentation for the Autonomous AI + Safety Workshop will overview how UL 4600 handles machine learning and AI techniques for autonomy pipelines, focusing particularly on safety performance indicators (SPIs) relevant to autonomous perception, fusion, prediction, and planning. As a trusted third party with a global customer base, Edge Case has the visibility across industry segments to develop common SPIs and safety case templates, which we believe will accelerate the arrival of safe autonomous mobility.

Addressed questions

Q17 How can we reliably detect the “edges” of ODDs, especially where they include weather and events, not in the ODD?

✔ **Answer by Michael Wagner**

UL 4600 attempts to define how you will evaluate whether you are reliably detecting the edge cases. The standard does not define a way to do that explicitly. It leaves that up to developers. We do expect that a number of these detectors themselves are going to require, perhaps, machine learning and AI.

Q18 What does UL4600 say about what a process should be when Safety Performance Indicators (SPIs) fail when deployed? Should it always be a fleet grounding? How and when should a regulator then get involved?

✔ **Answer by Michael Wagner**

There is a difference between Safety Performance Indicators (SPIs) and trigger for a Minimum Risk Condition (MRC). When an SPI occurs, you are not necessarily unsafe, but you are also not sure that you are safe anymore. Eventually, it comes down to the context and design decisions of the developer. Certainly, there are some SPIs that are so fundamental that when you start to see an uptake in violations, you probably should reassess your operations.

4.3 Talk 6: AI Safety from the Perspective of the DARPA Assured Autonomy Program

Sandeep Neema
Program Manager, DARPA

Summary

The DARPA Assured Autonomy program aims to advance how computing systems can learn and evolve with machine learning to better manage variations in the environment and enhance the predictability of autonomous systems like driverless vehicles. In this talk, we present some of the key results of this DARPA program. Specifically, the talk

will discuss approaches for assessing the competence of neural networks in classification and regression settings.

Addressed questions

Q19 How scalable is the proposed Assurance Architecture for large Convolutional Neural Networks (CNNs)?

✔ **Answer by Sandeep Neema**

Indeed, the scalability of the Assurance Architecture for large CNNs (and Neural Networks (NNs) in general) is a challenge and an active topic of research. The Assurance Architecture consists of techniques for formal verification as well as runtime monitoring of specific properties. Moreover, there is significant variability concerning the specific properties (and specific approaches) that one is trying to verify or monitor. For example, verification of reachability properties tend to be hard, and the collection of approaches that we pursue work on NNs with hundreds of thousands of nodes, which may be adequate for NNs used in control, but nowhere close to NNs used in perception. On the other hand, for robustness properties, there are already approaches that scale to CNNs, with millions of nodes representative of NNs used in perception. For distribution-shift monitoring and confidence estimation problems, the scalability is determined by the method used to summarize the training set and estimate the manifolds, rather than NN architecture itself. Researchers in the program have experimented with VAEs and SVDDs for summarizing the training set. They have been able to perform distribution shift monitoring in real-time for large scale CNNs.

Q20 How specific are the manifolds to the intended class of objects? Are there chances that an object A is in or close to the manifold around object B, but A and B are of different classes?

✔ **Answer by Sandeep Neema**

For classification tasks, manifolds do not always do the perfect separation. There certainly are cases when an object belonging to a class A, lies closer to the manifold of objects of class B. The classification task will result in a misclassification in this situation. For confidence estimation, we are learning and approximating the manifolds for the intended class of objects, and use the distance from the centroid of the manifold as an approximation of the confidence estimate. In one sense, the confidence estimation has the same problem as the classifier concerning these confounding cases. However, usually, such misclassifications occur closer to the boundaries of manifolds. The distance measure naturally provides low confidence estimates for objects on the boundaries of manifolds, reflecting the uncertainty inherent in such confounding cases.

4.4 Panel Discussion on Assuring Safety in AI components

Addressed questions

Q21 Does UL4600 make a statement about the completeness of triggering conditions? I think a systematic method to find triggering conditions is required to build necessary confidence.

 **Answer by Mike Wagner, Simon Burton**

UL4600 is clear about the necessity of building a safety case to provide evidence that a certain level of safety completeness has been reached before deploying an ADS on public roads. However, despite the augmentations brought in the safety case, it is clear that there will always be a specific residual risk. One of the reasons for this is due to the inequality of the exposure your system has to the real world at the development phase in comparison to when the system is deployed on public roads - i.e., the amount of testing on public roads at development time is much smaller in comparison to once the system is deployed. Hence, there is an obligation to continue monitoring new kinds of triggering conditions that have not been anticipated (i.e., unknown triggering events) in the pre-deployment phase.

Furthermore, the differentiation between known and unknown triggering conditions is important. For known triggering conditions, it is useful to establish a common database that appropriately summarizes them. Over time the database of known triggering conditions is expected to grow: i.e., the more public road driving is done, the more new triggering conditions will be discovered. This dataset can then be used to perform certain coverage tests at the development phase: i.e., by applying the known triggering cases to the system under development. When it comes to the unknown triggering conditions, it is much harder to argue that all unknown triggering conditions have been covered. The questions scientists and practitioners are challenged with are:

- *When do we actually start going into public roads?*
- *How many of these unknowns can be left out before we actually start driving on public roads*

For that reason, a step-by-step approach is needed to expand the complexity of the domain and expand in the authority of the systems (in terms of how much responsibility is given to the system).

Q22 AI-based algorithms are complex and opaque. Are there any approaches to a better understanding of how the AI-based algorithm work?

✔ **Answer by Simon Burton, Sandeep Neema, Mike Wagner, Iain Whiteside**

Based on the experience from system safety assurance cases, a lesson learned is that there is no “single silver bullet” that gives all the answers about the system’s performance. It is a comparable situation when it comes to a better understanding of AI-based algorithms. Several approaches are currently being developed. One of them explores how easy it is to force the DNN into a wrong classification: e.g., by changing a few pixels around a particular frame. Other approaches are sensitivity analysis, standard testing, analysis of the training data, and similar. Eventually, a combination of different approaches will need to be taken into account to get a better picture of the behavior of the AI-based algorithm and its characteristics. Furthermore, there is a research domain that primarily focuses on explainable AI [2], [3].

Q23 What component of an AD can/could AI replace, can the control be given to an AI component?

✔ **Answer by Simon Burton, Sandeep Neema, Mike Wagner, Iain Whiteside**

A couple of years ago, there was an interest in looking at end-to-end learning, also known as “common AI” - i.e., using AI from sensing to actuation tasks. However, the interest has changed from “Common-AI” toward using AI only in specific parts of the system, where AI is better than classic algorithms. The reasons for such a shift is (i) the opaqueness problem of AI-based components, (ii) the difficulty in defining how good AI-based components should be, as well as (iii) how to validate them. The smaller the functionality of the AI component is, and the more specific its functions are defined, the easier it is to tackle the three challenges mentioned above. Today, AI-based algorithms are commonly used when it comes to recognizing an object in the environment, predicting the future states of the environment, and planning.

Q24 If a product is certified according to UL 4600, is the intent that the product has now been determined sufficiently safe for commercial deployment? In other words, what does UL 4600 certification mean?

✔ **Answer by Mike Wagner**

Generally speaking, the idea of a standard is to lay out a set of guidelines for evaluating the safety of a system under development. The guidelines include performing risk analysis, identifying what it means to be safe, identifying the required safety integrity levels, and so on. Once the system is developed following the applicable standards, all safety-relevant activities are summarized in a Safety Case: i.e., a structured argument supported by evidence, intended to justify that a system is acceptably safe for a specific application in a specific operating environment. UL4600 standard aims at guiding safety experts in creating a safety case, so the evidence is self-sufficient, and the argument is sound.

✔ **Comment by Simon Burton**

Standards, whether they are from ANSI/UL, ISO typically, document state of the art and best practice, so that they can be referred to in guiding to ensure some consistency in approaches. However, also when the worst comes to the worst, it gives some way of measuring what could have been expected by companies developing and operating such vehicles.

Q25 How do we develop standards like UL in a space where there is so much agility and change at the moment?

✔ **Answer by Simon Burton**

Today, there are several standards under development (i.e., by ISO, IEEE, ANSI/UL) that define the safety requirements and processes for safe automated vehicles. Moreover, UNECE is developing regulations that refer to said standards. For complex systems like ADS, we can work on an industry and society consensus and publish rules and laws on an international level. Once the consensus is reached, we can then decide what type of standards are necessary to apply for making systems good enough. Furthermore, one should be careful about being too absolute about saying “*if specific standards are fulfilled, the system is legally safe*”. There is a lot more to do than fulfill a standard: i.e., type approval, which is covered by UNECE.

✔ **Comment by Mike Wagner**

I can completely agree with the need for agility and one of the reasons why we encourage everyone in this forum to contribute and help develop the next version of the 4600 standards, which is currently worked on.

5 | Survey Results

5.1 Contributors

In total, 23 contributions were made to the post-event survey. A summary of the contributors' workplace, their role, company/institution, and experience is summarized in Figure 5, under Survey Question (SQ1-SQ4). Contributors' workplace was from 10 different countries. Concerning their current role in the company, the distribution is as follows: 29% research and teaching-oriented (e.g., Ph.D. student, Professor), 28% have managing roles, 24% general technical/engineering (e.g., system architect, project engineer), and 19% general safety (e.g., functional safety, safety experts). Furthermore, 30% work in a Tier 1 company, 26% are working in a research institution or university, 9% in a Tier 2 company, 9% for OEM, 9% for a government organization, 4% for a semiconductor company, and 4% for other. Finally, 34% of the contributors are involved in the development or research of SAE L2 AD, whereas 33% in SAE L4 AD, 11% in SAE L5 AD, and 22% are not involved currently.

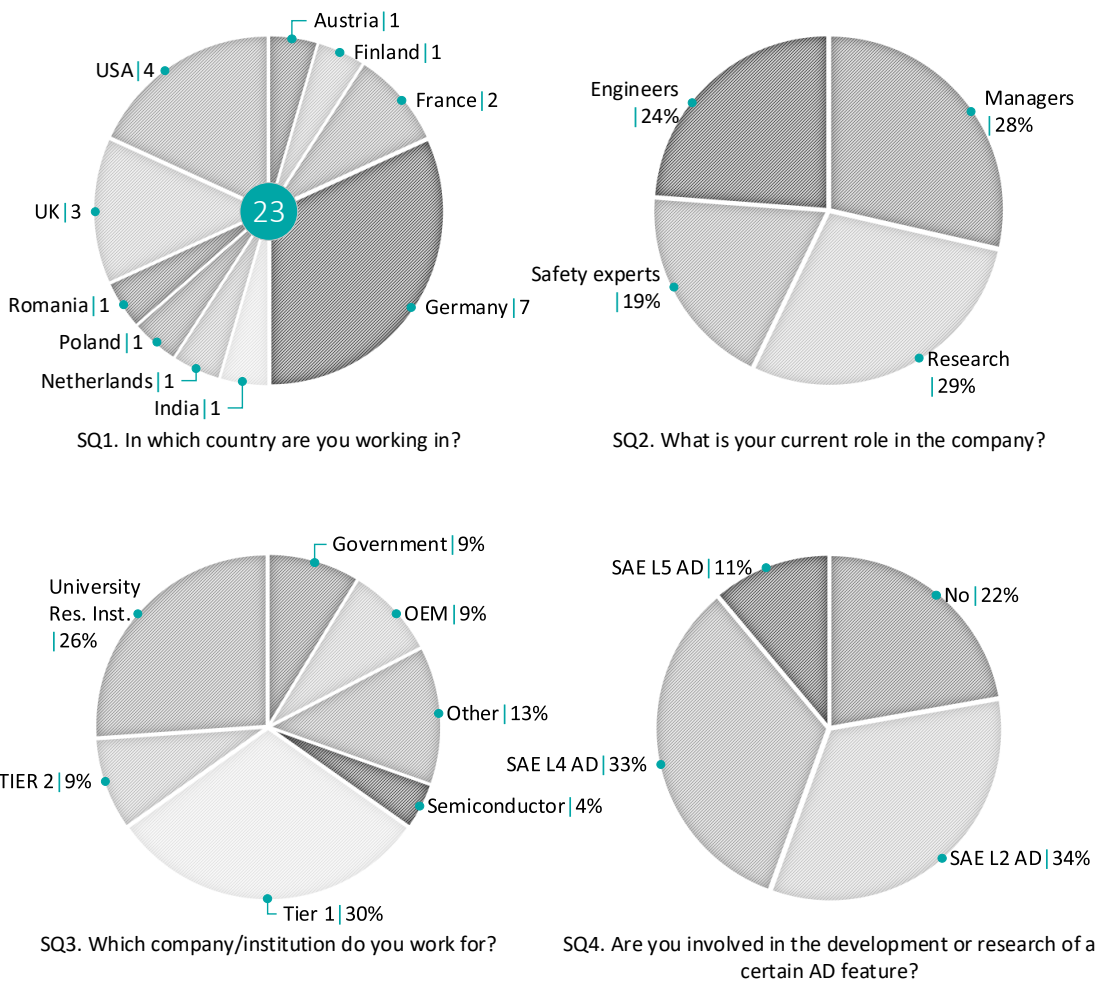


Figure 5: Information about the contributors of the survey.

5.2 Subject: General AD

The topics discussed during the Chapter Event often focused on challenges that are faced during the development of future SAE L4 ADS¹. Therefore it is important to have a common understanding of when SAE L4 ADS are expected to be on public roads and which ODD SAE L4 AD is bringing the most value. The results of the questions are summarized in survey questions SQ5-SQ7.

SQ5 When do you expect SAE L4 AD features to be available in the Highway operational design domain?

Results

Figure 6, left side presents the results. The majority expect SAE L4 AD to be on public highway roads between 2026-2028 (44%). Others expect them to be available in 2020-2022 (17%), 2023-2025 (13%), or later than 2028 (9%). Last, 17% did not know.

SQ6 When do you expect SAE L4 AD features to be available in the Urban operational design domain?

Results

Figure 6, right side depicts the results. Most contributors (52%) expect SAE L4 AD to be available on public urban roads later than 2028. Others expect them to be available in 2023-2025 (18%), 2026-2028 (17%). None of the survey participants believe that SEL L4 AD will be available on public urban roads in 2020-2022. Last, 13% do not know.

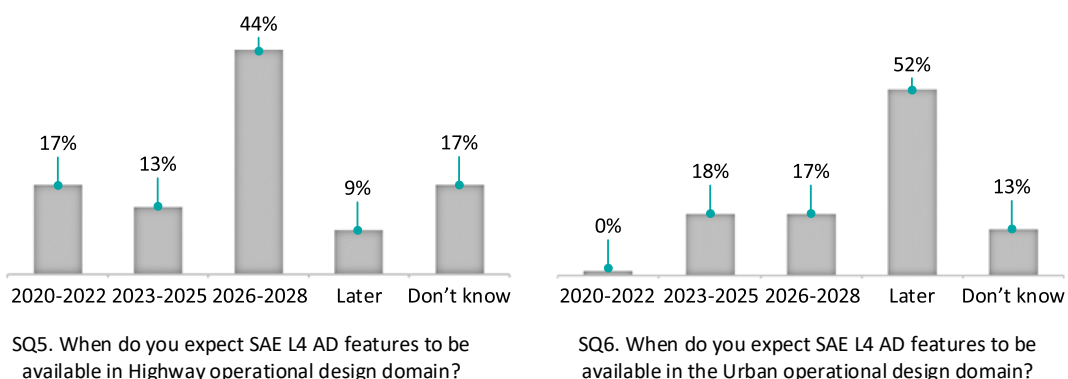


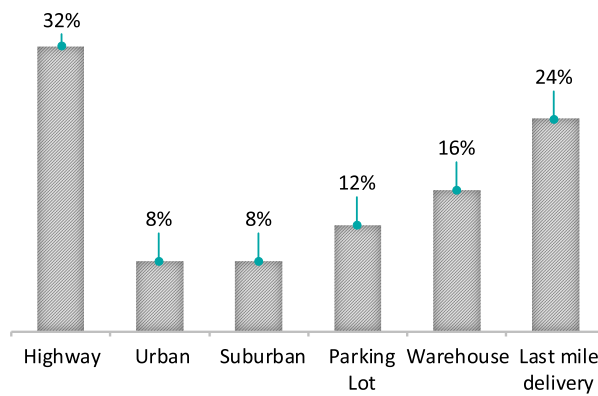
Figure 6: General AD questions - part 1.

¹SAE L4 ADS performs the complete dynamic driving tasks (DDT) and DDT Fallback (i.e., no fallback-ready driver needed) within a limited ODD.

SQ7 In which of the operational design domains do you think SAE L4 AD brings the most value?

Results

Figure 7 outlines the results. Here, a significant portion (32%) of the contributors believe that SAE L4 AD will bring the most value in the highway ODD. Last-mile delivery ODD comes second with 24%, Warehouse ODD third with 16%, and parking lot ODD with 12% and suburban and urban both with 8%.



SQ7. In which of the operational design domains do you think SAE L4 AD brings the most value?

Figure 7: General AD questions - part 2.

5.3 Subject: The Autonomous

It is essential for an initiative to continuously receive feedback from contributors on the selected approaches and vision. Hence, we asked the following questions SQ8 and SQ9.

SQ8 Do you think the approach proposed by The Autonomous is feasible?

Results

Figure 8 depicts the results. The majority (78%) of the survey participants believe that The Autonomous approach is feasible, whereas 9% do not. 13% have provided no answer.

For the sake of transparency, opinions (positive and negative) from the survey contributors are summarized below.²

Participants - justifying their answers

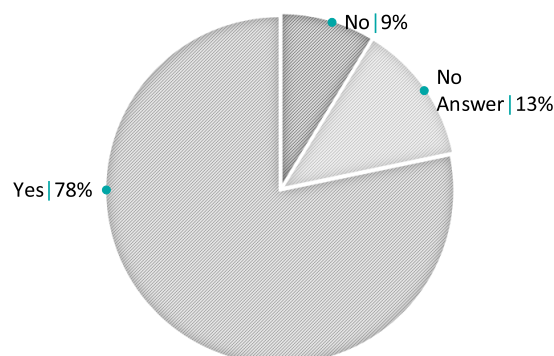
No: It is a good goal to have. However, commercial pressures will prevent organizations from sharing information to the level required.

Yes: Development effort is too big for one entity to handle that. The collective collaboration is the key. A common consensus approach also enables decision/legislation makers to settle and propose policies to support this venture. After all, we all make or break the fundamental structure of our societal interactions.

Yes: The problem is too complex for single organizations to work alone.

Yes: Yes, but you need stability on the approaches to solve the problem. It is too early right now.


Yes: Research and development activities need to converge to solve a problem efficiently. Bringing safe autonomous mobility is a truly challenging task that can be achieved faster if we converge efforts.



SQ8. Do you think the approach proposed by The Autonomous is feasible?

Figure 8: Results from survey question SQ8.

²Only spelling and grammar changes have been made. The out-of-context text has been removed.

 **Participants - justifying their answers**

- Yes:** Only if an environment is enabled where industry-wide players can confidently share (without being identified) boundary case scenarios to improve reliability and safety of the AI systems
- Yes:** Bringing together worldwide collaborating expertise to develop reference architectures and scenarios is essential because it is much too complex for any single company or organization.
- Yes:** The ecosystem being fostered is an integral aspect of safe AI
- Yes:** Feasible but difficult, given the many other organizations attempting a similar role (e.g., HEADSTART, VVMethoden, UNECE, WEF Safety Pool, etc.)
- Yes:** Main stakeholders understand the need to join forces to tackle such a complex problem and share experiences as more and more AV hit the road.

SQ9 In your opinion, what do you think the main challenges will be for forming The Autonomous ecosystem?

 **Results**

The list below summarizes the main challenges indicated by the participants.

- Building a strong and sustainable community that creates a win-win scenario for all participants.
- Cohesion between manufacturers (i.e., supportive and sharing)/
- Fair compensation for the contributions with added value in accordance with the business interests of laggards. Stakeholder interests may be too heterogeneous.
- Making it clear to stakeholders that you have the backing and technical expertise to be worthy of significant collaboration efforts from them.
- Delivering results that are readily usable.
- Emphasize the Return of Investment of each stakeholder involved.
- Commercial competitiveness and intellectual property.
- Policy and regulation
- Public acceptance.

5.4 Subject: Neural Network Verification and Validation

SQ10 In your opinion, what is the most promising technique for verification of AI-based components?

Results

The list below summarizes the main verification techniques of AI-based components indicated by the participants.

- Combination of production testing, simulation, formal verification, runtime verification.
- AI decomposition of the behavior and systematic verification of parts of it according to well-stated specifications. It is extremely time-consuming and very rigorous, but for verification (not validation or evaluation), this is the most promising way I see.
- On-road data collection.
- Automated validation of data from central global database/scenarios from precision/recalls across the ODD.
- Standard scenario set that is simulated and then validated where possible by real driving in designated ODD.
- Scenario-based verification with scenarios obtained from vehicles operating in the field.
- Virtual testing and sampled physical testing
- Coverage Based Pseudo-Random Generation
- Adaptive scenario-based testing

Opinion from a participant

- Why does AI need different V&V approaches than complementary methods? It is not a matter of AI or non-AI but a matter of system complexity in general.

SQ11 Concerning V&V: what gaps do you think exist in the emerging standards e.g., Appendix B of ISO/CD TR 4804 [4] (derived from the Safety First for Automated Driving White Paper) and UL 4600 [5]?

Results

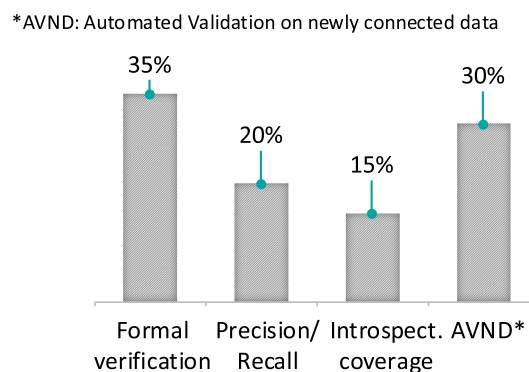
The list below summarizes the answers of the participants.

- Gap manifests mainly in "How to achieve that?".
- As with all standards, they can only ever specify generically "what" you have to do, but not the specifics of "how" you do that for your application.
- Measurable performance indicators.
- Method to approach, in a systematic manner, an incremental coverage of the ODD.
- UL4600 is goal-oriented, which makes it generally applicable and capable of keeping pace with technological developments. However, by being too generic, it does not address the characteristics of technologies (such as simulation, for example) used in V&V of AI behavior or any other SW behavior.

SQ12 Rank the techniques for V&V by their importance.

Results

Figure 9 outlines the results. The participants have voted as follows: 35% for formal verification, 30% for Automated Validation on a newly connected data (AVND), 20% for precision/recall, 15% for introspective coverage.



SQ12. Rank the techniques for V&V of AI-based components by their importance?

Figure 9: Results from survey question SQ12.

SQ13 Adversarial attacks have been shown to exist for all Deep Neural Networks. Do you believe them to be important to worry about in practice?

*It's the human nature to
play, hack and cheat. 😊*

— François E. Guichard

Results

Figure 10 depicts the results. The majority (82%) of the survey participants do think adversarial attacks are important and should be taken into account in practice. Whereas 9% do not, and 9% do not know.

Participants - justifying their answers

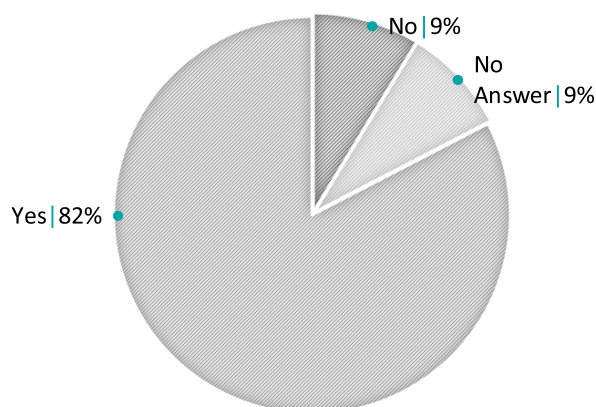
No: There is still much work to do in nominal operational conditions.

Yes: Deep Neural Networks (DNN) will be updated from remote and does not sit physically in a closed well protected, and end-to-end protection.

Yes: Fabricated sample data can lead to various misbehavior of the DNN and can lead to the wrong prediction. Besides the detection of obfuscated code, adversarial attacks are a major research aspect in AI security.

Yes: Any vulnerability on a fleet of vehicles will have a huge impact if anyone chooses to exploit it.

Yes: At times, adversarial attacks (assuming affecting the ODD) are not different from a perception limitation in certain environmental conditions.



SQ13. Adversarial attacks have been shown to exist for all Deep Neural Networks. Do you believe them to be important to worry about in practice?

Figure 10: Results from survey question SQ13.

 **Participants - justifying their answers**

Yes: Adversarial robustness is a very narrow property. Many many more critical properties are at least as important, but that said it is important.

5.5 Subject: Assuring safety in AI components

SQ14 In your opinion, what is the key challenge in assuring the safety of AI components?

 **Results**

The list below summarizes the main challenges indicated by the participants. These are:

- **Integrity:** Measuring their integrity in real-time.
- **Determinism:** Because AI components evolve over time, for the same set of inputs, you get different outputs. Furthermore, AI behavior is not predictable in special “corner” cases.
- **Explainability:** AI-based algorithms lack sufficient explainability because of the “black box” manner they are used.
- **Validation and Verification:** The lack of understanding (data science) makes it difficult to validate and verify.
- **Requirements definition:** Specifying what the system should and should not do. If you cannot specify a property, you can neither verify nor validate it.
- **Others:**
 - Statistical/distributional paradigms in learning. Multiple outputs governed by assertive architectures.
 - Limiting AI to functional areas where it is absolutely needed. Using traditional safety and monitoring elsewhere.
 - Adopting appropriate testing techniques.
 - Show that follows the design intent in all relevant cases.

SQ15 One of the main challenges for high fidelity simulation for AD is generating realistic synthetic scenes. Deep Neural Networks are known to be very sensitive to their input distribution: a topic known as domain adaptation. Do you believe this presents a problem for simulation-based verification for AD?

Results

Figure 11 depicts the results. To a large extent (65%), the participants have answered with "Yes", whereas 22% with "No" and 13% have provided no answer.

Participants - justifying their answers

No: While it is difficult, it is the only way to explore the likely events that can happen in the operational space.

No: Synthetic scenes can help create scenes beyond what is typically found in the real world, thereby being able to test for unexpected or infrequent scenarios.

No: Is your goal Validation or Training? If validation, noisy input is reality. High fidelity is a non-starter.

No: I think it helps the virtual evaluation of AD behavior. It can, of course, be complemented with other techniques, but it still helps.

Yes: It is difficult to know what "criteria" a DNN is using to make its decision. For example, if a DNN identifies that a car is a car because it has door mirrors, what happens when the DNN encounters a car without door mirrors?

Yes: The models in all simulations must be validated if this is not possible, the sim should not be used for any purpose relating to safety.

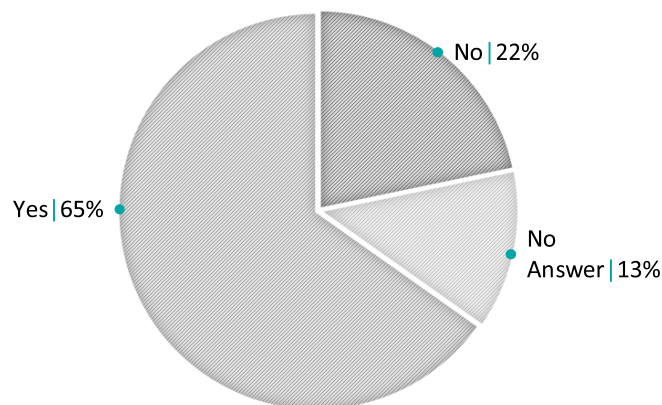


Figure 11: Results from survey question SQ15.

Participants - justifying their answers

Yes: There is a math problem related to “continuity”. When you train your AI agent with your data, it only makes sense if your algorithm (function) is continuous on the range of your use case.

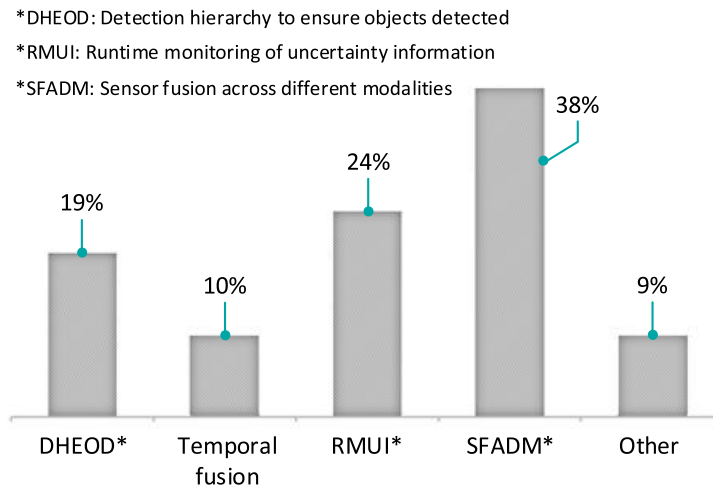
Yes: Ideally, simulation-based verification should cover end-to-end systems, so the generation of raw sensor data is important. Such sensor data needs to match real-world data sufficiently to reproduce the same performance.

Yes: Ignoring domain adaptation creates uncertainty. Being able to understand, quantify, and argue how the system may fail. However, it is OK because the input is not representative of the real-world or extremely unlikely, will contribute to a more compelling safety case.

SQ16 In your opinion, what are the most important system-level mitigation you can think of for a perception system with AI components?

Results

Figure 12 presents the results. 38% of the participants have voted for sensor fusion across different modalities (SFADM), 24% for runtime monitoring of uncertainty information (RMUI), 19% for detection hierarchy to ensure objects detected (DHEOD), 10% for temporal fusion. 9% have given other answers, such as using redundant, diverse implementation and proper design and development of the AI-components in the first place.



SQ16. In your opinion, what are the most important system-level mitigations you can think of for a perception system with AI components?

Figure 12: Results from survey question SQ16.

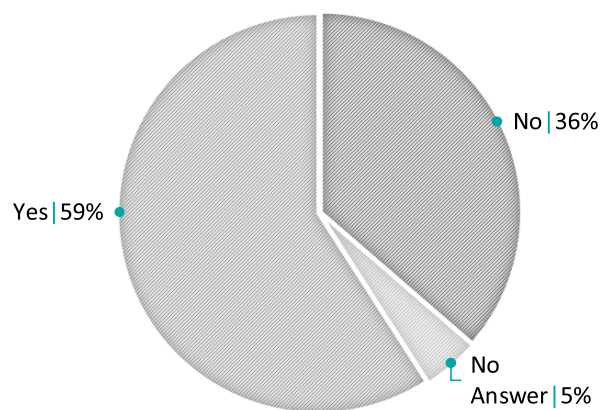
SQ17 Do you believe that scenario-based testing should be the dominant method for testing an AD system?

Results

Figure 13 depicts the results. To a large extent (59%), the participants believe that scenario-based testing should be the dominant method for testing an AD system. Whereas 36% do not think so, and 5% have provided no answer.

Participants - justifying their answers

- No:** Scenario has its limitation in the context of the complexity of AI-based systems.
- No:** This quantitative approach can only test what you specify. Can you specify it all?
- No:** Scenario-based testing is critical, but not at the expense of other techniques.
- No:** Completeness of scenarios need to be shown.
- Yes:** There is a lot of accidentology statistics, these mapped to the ODD shall provide us with expected scenarios, else the space of the possibles is too large.
- Yes:** If it is done not only at design time but also at runtime. At design time, testing scenarios can be creative and complemented by real-time runtime scenarios that reflect concrete technical situations where the system needs to make decisions.
- Yes:** It is hard to see how to "frame" the problem if you do not use scenarios as the way to focus the testing.
- Yes:** Carefully thought and collected scenarios (ones triggering disengagements) are key to efficient testing.



Q17. Do you believe that scenario-based testing should be the dominant method for testing an AD system?

Figure 13: Results from survey question SQ13.

Participants - justifying their answers

Yes: Scenario-based testing, via real-world testing or simulation, is a foundation to build public and authorities confidence. I expect pure statistical analyses and reasoning to complement what cannot be tested or account for the long tail of corner and edges case. I cannot suggest what should be an acceptable ratio of the weight of those two approaches in the safety case, but it seems that the former will play a heavier role to the strength of the safety case.

Yes: Scenario-based testing facilitates risk-based testing and can be related to safety hazard analysis, etc.

SQ18 How do you believe, can an AI component be incorporated into a Safety Case argument?

Results

The list below summarizes the answers of the participants:

- By limiting the use to essential functional areas only.
- By a combination of monitoring, diverse implementation, and intensive testing.
- By measuring the functional integrity. This can be monitored as part of a safety system.
- By combining evidence, chosen architecture and rationales, chosen dataset (training, validating, testing) and rationale, chosen training method, and rationale, an argument on AI-component uncertainty estimate and influence in the system behavior.
- By producing reproducible evidence that the component does its intended function.
- Through arguing and managing risks introduced by the uncertainty of known and unknown.
- The safety case needed to explicitly state the AI performance and the level of uncertainty present.
- An AI component is like any other component. One has to define its expected behavior and test against that specification.

Opinion from a participant

- The problem with AI seems to be the “probabilistic” nature and the data to train the AI agent. It seems to be a solvable problem that requires further serious research and honest work (away from any populist marketing disruption)

Appendices

A | List of Abbreviations

| | |
|----------------|--|
| AD | Automated Driving |
| ADAS | Advanced Driving Assistance Systems |
| ADS | Automated Driving System |
| AI | Artificial Intelligence |
| ANSI | American National Standards Institute |
| ASIL | Automotive Safety Integrity Level |
| AUTOSAR | Automotive Open System Architecture |
| AV | Automated Vehicle |
| CD | Commission Draft |
| CNN | Convolutional Neural Network |
| CPS | Cyber-Physical System |
| DNN | Deep Neural Network |
| ECU | Electronic Control Unit |
| FO/FD | Fail-Operational/Fail-Degraded |
| FuSa | Functional Safety |
| ISO | International Standardization Organization |
| L1 | SAE Level 1 |
| L2 | SAE Level 2 |
| L3 | SAE Level 3 |
| L4 | SAE Level 4 |
| L5 | SAE Level 5 |
| NN | Neural Network |
| ODD | Operational Design Domain |
| OEM | Original Equipment Manufacturer |
| PAS | Publicly Available Specification |
| SAE | Society of Automotive Engineers |
| SaFAD | Safety First for Automated Driving |
| SOTIF | Safety of The Intended Functionality |
| TR | Technical Report |
| UL | Underwriters Laboratories |
| V&V | Verification and Validation |

B | Compliance Guidelines

Ensuring safety is the key to gaining acceptance of autonomous mobility on a broad scale. The Autonomous will start this critical discussion by gathering together the complete autonomous mobility ecosystem and facilitate a mutual exchange of ideas by offering various workshops on key topics (Safety & Security, Safety & AI, Safety & Architecture, Safety & Regulation), panel discussions, and keynote speeches.

At The Autonomous, we are committed to ensuring that all discussions take place in full compliance with the rules of competition law. In order to allow for an open exchange of ideas within the limits of the law, this Guideline sets out practicable rules for The Autonomous. Compliance with this Guideline is obligatory for all organizers and participants.

1. **Permitted topics:** Topics which may be covered in discussions, workshops and meetings organized by The Autonomous include:
 - 1.1. General technical and scientific developments relevant to autonomous mobility;
 - 1.2. Legislative proposals and/or regulatory measures and their impact on the autonomous mobility ecosystem;
 - 1.3. The political environment;
 - 1.4. Current economic developments and general developments in the industry (if publicly available);
 - 1.5. Exchange of freely available information e.g. economic data available online or in annual reports.
2. **Non-permitted topics:** Participants may not discuss, agree, share information on, or in any other way coordinate their behavior regarding competitively sensitive issues, including:
 - 2.1. Current and future prices, including selling prices, purchase prices, price components, price calculation, rebates, and intended changes in prices;
 - 2.2. Terms and conditions of supply and payment for contracts with third parties;
 - 2.3. Market sharing, including discussions on the division of sales territories or customers (e.g., by size, product type, etc.);
 - 2.4. Co-ordination of bidding towards third parties, including information on customers' commercial expectations and the firm's proposed response, as well as information on proposed bids (whether a bid will be submitted, for which lots, etc.);
 - 2.5. Boycotts against certain companies, e.g., agreements not to work with certain customers or suppliers, or to exclude specific companies from discussions on the establishment of a technical standard;
 - 2.6. Information about business strategies and future market conduct, such as planned investments or the commercial launch of new technologies or products (if not publicly available). In particular, agreements to delay a new technology or to fix the commercial terms of its introduction are prohibited;

-
- 2.7. Detailed information on financial performance, such as recent information on profits and profit margins on a non-aggregated basis (if not publicly available);
 - 2.8. Information on internal research and development projects. This comprises estimations about the feasibility of specific technical solutions or the costs attached to the implementation of a specific solution.
3. **Measure to ensure compliance:** In order to ensure compliance and to contribute to an open discussion, The Autonomous will implement the following measures:
 - 3.1. Attendance by legal counsel: All discussions and workshops will be attended by in-house or external legal counsel. Legal counsel may break off or adjourn the discussion in case of doubts with regard to competition law compliance.
 4. **No Reliance:** The purpose of this Guideline is to briefly summarize the competition rules applying to discussions at The Autonomous. It, however, cannot address the full complexity of the applicable law and does not constitute legal advice to participants and their respective firms as to their obligations under competition law. At The Autonomous, we encourage participants to familiarize themselves with the rules of competition law. Should any participant have doubts as to the legality of any discussion in the course of The Autonomous, she/he may:
 - 4.1. raise such doubts to the legal counsel attending the discussion. The legal counsel shall record any such request in the minutes;
 - 4.2. leave the meeting if the discussion continues without the participant's doubts having been resolved. The legal counsel shall record the name of the participant as well as the exact time of the participant's departure in the minutes.

C | Standard Settings Guideline

Ensuring safety is the key to gaining acceptance of autonomous mobility on a broad scale. To address security concerns in connection with autonomous driving, safety proves to be the main concern and challenge for mass adoption. These current challenges and associated investment costs cannot be mastered by a single OEM, Tier 1, or Tech company. Just like in aviation, autonomous driving needs to set common technical and ethical standards, legislation, and a process to learn from past incidents and avoid future ones.

At The Autonomous, our mission is to establish a global safety reference, created by the global community, which facilitates the adoption of autonomous mobility on a grand scale. We are committed to ensuring that this process takes place in full compliance with the rules of competition law. To this end, this Guideline supplements The Autonomous' Compliance Guideline, by setting out practicable rules for standard-setting processes at The Autonomous. Compliance with this Guideline is obligatory for all organizers and participants.

1. **Openness and transparency:** The Autonomous follows an open and transparent approach to participation in its panels, workshops, and other working groups. The establishment of a global safety reference will follow the following principles:

- 1.1. Unrestricted participation: involvement is open to all industry stakeholders. Active involvement may only be limited if absolutely necessary (i.e., to prevent inefficiency) and based on objective and non-discriminatory criteria;
- 1.2. Transparency: all attendees of The Autonomous, as well as all other stakeholders concerned, will be informed of any announcement, progress, and outcome;
- 1.3. Review and comments: Stakeholders not participating in the process will be able to review and comment on the result of the standard-setting process. Any agenda referring to activities of The Autonomous will be disseminated to participants in due course prior to the execution of the activity. Participants shall have the right to comment or to contribute to such an agenda.

2. **Non-exclusivity, free access**

- 2.1. No obligation to comply: Participants are free to develop alternative standards or products that do not comply with the evolving standard;
- 2.2. Free access to standards: Any developed standards will be accessible for all interested stakeholders (whether or not they participated in The Autonomous) on fair, reasonable, and non-discriminatory terms.

3. **IPR Policy**

3.1. **Definitions:**

- 3.1.1. "Affiliate": any subsidiary or holding company of a participant, any subsidiary of any of its holding companies and any partnership, company, or undertaking (whether incorporated or unincorporated) in which a participant has the majority of the voting rights or economic interest.

-
- 3.1.2. “Essential”: an intellectual property right is essential where it would be technically (but not necessarily commercially) impossible, taking into account normal technical practice and state of the art generally available at the time of adoption of the standard, to implement the respective standard without making use or infringing the IPR in question.
- 3.1.3. “FRAND terms”: fair, reasonable, and non-discriminatory terms.
- 3.1.4. “Implement/Implementation”: (i) to make, market, sell, license, lease, otherwise dispose or make use of equipment; (ii) repair, use or operate equipment; or (iii) use methods – as specified in the respective standard.
- 3.1.5. “Intellectual Property Rights” or “IPR”: any copyright, Patent, registered design, and any application thereof. IPR does not include trademarks, trade secrets, moral rights, right of know-how, and confidential information.
- 3.1.6. “Patent”: any patent, utility model, or any application for such.
- 3.2. **Scope of Application:** Participants owning any Essential IPR shall be free to exploit such IPR outside the scope of The Autonomous at their absolute discretion and any revenues or other benefits, which the participant may receive from such exploitation of such Essential IPR, shall be for the participant’s own account.
- 3.3. **FRAND commitment**
- 3.3.1. Save in the case of any Essential Patents identified in accordance with Section 3.4.4, a participant will give an undertaking that it is prepared to grant licences to anyone wishing to Implement the standard to which the Essential IPR relates:
- (i) on FRAND terms;
 - (ii) to all its Essential IPR relevant for the respective standard;
 - (iii) to the extent necessary to permit the Implementation of the respective standard.
- 3.3.2. The undertaking pursuant to Section 3.3.1 may be made subject to the condition that those who seek licenses agree to reciprocate.
- 3.3.3. Where a participant has elected not to declare or has failed to declare any Essential IPR for a given standard in accordance with Section 3.4.4, the participant shall be deemed to have given the undertaking in accordance with the terms of Section 3.3.1.
- 3.3.4. Both, the participant who has given an undertaking pursuant to Section 3.3.1 or who is deemed to have given an undertaking pursuant to Section 3.3.3, and any beneficiaries of such undertaking wishing to acquire a license in accordance with Section 3.3.1, acknowledge and agree that:
- (i) They will act in good faith, in order to negotiate a license agreement;
 - (ii) If both parties have not been able to agree on an Essential IPR license, each party has the right to pursue the matter before the national courts to resolve the matter.
- 3.3.5. Each participant will ensure that its Affiliates and its Affiliates’ successors in title will give an undertaking pursuant to Sections 3.3.1 to 3.3.4 above. If a participant or its Affiliate transfers ownership of Essential IPR that

is subject to an undertaking 3 pursuant to Sections 3.3.1 to 3.3.4 above, such undertaking shall include appropriate provisions in the relevant transfer documents to ensure that the undertaking is binding on the transferee and that the transferee will similarly include appropriate provisions in the event of future transfers with the goal of binding all successors-in-interest. The undertaking shall be interpreted as binding on successors-in-interest regardless of whether such provisions are included in the relevant transfer documents.

3.4. Declaration of Essential IPRs

3.4.1. Prior to any official adoption of any standard or part thereof, each participant shall provide a written declaration of the Essential IPR relevant to the subject matter. Such declaration shall list:

- (i) all potentially relevant Essential IPR held by the participant or its Affiliates;
- (ii) filing and registration number, application date and if published the title of the respective Essential IPR;
- (iii) terms (i.e., explicitly (non-FRAND terms as opposed to clause 3.3.1, but without specifying royalty rates on any other royalty terms)) on which the participant or its Affiliate is prepared to grant licenses to other participants or any third parties; and
- (iv) statement whether the declaration is made subject to the condition that those who seek licenses agree to reciprocate.

3.4.2. In the absence of a declaration of any Essential IPR, the participant will be deemed to have given the undertaking for that Essential IPR associated with the relevant standard or part thereof, in accordance with Section 3.3.3.

3.4.3. Any declaration may identify such Essential Patents, for which the participant or its Affiliate are unwilling or unable to enter into an undertaking to license on FRAND terms in accordance with Section 3.3.1. The declaration shall:

- (i) identify any such any Essential Patent, by way of filing number, date, and if published, optionally its title;
- (ii) describe in sufficient detail the reasons why the participant or its Affiliate are unwilling or unable to enter into an undertaking to license on FRAND terms in accordance with Section 3.3.1.

3.4.4. Where a participant, in accordance with Clause 3.4.3, has identified an Essential Patent, which the participant, or its Affiliates, is unwilling or unable to license in accordance with Clause 3.3.1, the participant will lose its right to participate and to receive undertakings pursuant to Clause 3.3.1 from other participants in relation to the respective standard or part thereof to which an Essential Patent relates, if:

- (i) any other participant informs the Chairman within a reasonable period, in writing, that it does not accept that the reasons in the relevant declaration (as required in accordance with Clause 3.4.3(ii)) are reasonable and justified; and

-
- (ii) based on its duly justified non-acceptance of these reasons pursuant to Clause 3.4.4.(i), wishes that the aforesaid participant shall not be able to rely on its right to participate and to receive undertakings pursuant to Clause 3.3.1 from other participants.

3.5. Disputes concerning ownership of Essential IPR: If two or more participants claim ownership of the same Essential IPR, the participants claiming ownership shall:

- (i) negotiate and resolve the question of ownership in good faith and
- (ii) if no solution is found pursuant to section s3.5.1, have the right to pursue the matter before the national courts to resolve the dispute.

D | Acknowledgments

First and foremost, sincere thanks to all keynote speakers, namely David Hand, Iain Whiteside, Michael Wagner, Sandeep Neema, Simon Burton, Yoav Hollander. Their constant support over the past months and in-depth knowledge in the field resulted in outstanding presentations and discussions.

Furthermore, profound gratitude to all the participants at the virtual Chapter Event as well. Their questions enriched and deepened the discussions throughout the workshop.

Special thanks also go to the contributors of the post-event survey who enhanced the quality of discussions and ultimately of this report. In this post-event survey, the contributors were given the option to select whether their names should be mentioned or not. The following is a list of a substantial number of contributors: Alwyn Earle Goodloe, Andrei Aksjonov, Berthold Puchta, Daniel Petrisor, Dr. Michel Parent, Dr. Rahul Razdan, Emilia Cioroai, Eric Barbier, Francois E. Guichaard, Helen Monkhouse, Hendrik Weppelmann, J. Ibanez-Guzman, Om Ranjan, Zeyn Saigol.

Likewise, warm thanks to all reviewers - for all your comments and ideas for enhancement you have proposed.

Sincere thanks to Five for co-hosting this event with The Autonomous. It has been a pleasure working with you on this project.

Many thanks to Georg Kopetz, Marc Lang, Ricky Hudi, and Stefan Poledna for initiating The Autonomous and believing in this cause.

Last but not least, warmest thanks to The Autonomous team - Iulia Alina Baidac, Luisa Griesmayer, Susanne Blum, and Philip Schreiner - for your excellent work and continuous support.

E | Feedback

In our continuous effort to develop The Autonomous as an open platform and space for dialogue among different stakeholders, we welcome all feedback and interest in making safe autonomous mobility a reality. We highly value any comments, ideas, or suggestions you may have to help improve the outcome of this report or contribute to the initiative. Please do not hesitate to contact us at: [contact@the-autonomous.com].

References

- [1] International Organization for Standardization (ISO). Road vehicles-functional safety standard, ISO 26262. *International standard*, 2018.
- [2] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2:2, 2017.
- [3] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction*, pages 295–303. Springer, 2018.
- [4] International Organization for Standardization (ISO). ISO/CD TR 4804, Road vehicles — safety and cybersecurity for automated driving systems — design, verification and validation methods. <https://www.iso.org/standard/80363.html>. Accessed: May-2020.
- [5] Underwriters Laboratories. ANSI/UL 4600 standard for safety for the evaluation of autonomous products. <https://ul.org/UL4600>. Accessed: May-2020.