SafeTRANS Roadmap on Foundations for Safety and Explainability of AI based Safety-Critical Applications

December 1, 2023

Executive Summary

Artificial Intelligence has grown into a front-runner technology of digital transformation, disrupting economy, society, and our life, and soliciting massive investment and funding worldwide. We are, however, running into the limits of controllability of large, highly connected AI based systems, due to lack of understanding their complexity and of methods and processes to establish their safety, reliability, and transparency. These limitations are novel in kind and severe, and will lead to dwindling public and consumer acceptance, and hence to drastic losses of business opportunities and markets, as currently to be witnessed in the automotive sector's broad retreat from highly autonomous driving. Given this key industrial relevance of assuring safety and explainability for AI based system, we plan to launch the *Joint German Research Center on Foundations for Safety and Explainability of AI based Safety Critical Applications*, providing the foundational research for solving the underlying research challenges in cooperation with research labs of relevant industries and organization. This center is coordinated by a core team leading researchers from industry, AI, and formal methods whose research demonstrates clear relevance for addressing safety and explainability of AI based safety-critical systems. This strategy paper outlines the key approach of a planned Lighthouse Project focusing on methods and tools required to establish safety and explainability for industry selected classes of safety critical applications with high business potential. It represents the joint position of the consortium led by a core team comprising key German industrial sectors wanting to deploy AI based components in safety critical applications and highly renown researchers in the areas of AI, Formal Methods, Control Theory, and Human Cognition pushing the fundamental barriers which currently block certifiability of such applications.

Contributors

The development of the industrial priorities of this roadmap are based on contributions and discussion with representatives from the following companies.

Industrial Core Team

• Bosch	Dr. Michael Pfeiffer
• SAP	Dr. Fei Yu Xu
• Siemens CT	Dr. Cornel Klein
• Siemens Mobility	Prof. Dr. Jens Braband, Dr. Claus Bahlmann
• Trumpf	Klaus Bauer, Dr. Zaigham Faraz Siddiqui
• VW	Dr. Fabian Hueger, Dr. Peter Schlicht,
• ZF	Dr. Manuel Götz

The development of the academic challenges of this roadmap are based on contributions and discussion with the following researchers.

Academic core team

- Prof. Dr. Christel Baier, Full Member of Cluster of Excellence CETI on Tactile Internet with Humans in the loop and Collaborative Research Center on Foundations of Perspicuous Software Systems
- Prof. Dr. Werner Damm, previously coordinator Collaborative Research Center on Analysis and Verification of Complex Systems, Chairman SafeTRANS, Moderator
- Prof. Dr. Martin Fränzle, previously coordinator Research Areas Hybrid Systems in AVACS and Vice-President Research Carl von Ossietzky Universität Oldenburg,
- Prof. Dr. Mathias Hein, Full Member Cluster of Excellence on Machine Learning for Science
- Prof. Dr. Johannes Helbig, Trusted AI Core-Team, Industrial Observer
- Prof. Dr. Holger Hermanns, Coordinator Collaborative Research Center on Foundations of Perspicuous Software Systems
- Prof. Dr. Peter Liggesmeyer, Scientific Director FhG IESE and Scientific Director of the Forschungsbeirat Plattform Industrie 4.0
- Prof. Dr. rer. nat. Wojciech Samek, Head of AI Department at Fraunhofer HHI, Fellow at Berlin Institute for the Foundation of Learning and Data.
- Prof. Dr. Philipp Slusallek, Founding Director Cluster of Excellence on Multimodal Computing and Interaction und Co-Chair Claire Research Network

Note

This roadmap is a slightly modified version of a document submitted to the German Ministry of Research and Education in December 2021. Based on the high relevance of the topic of this document, SafeTRANS decided at its Steering Board Meeting November 9 to make this document available as a public roadmap. All contributors of this document have given permission to its publication.

I. Motivation and Approach

The need to push research on safety and explainability of AI based safety critical systems is increasingly understood in industry. In the US, the Consortium on the Landscape of AI Safety - Bridging perspectives on trustworthy AI - is bringing together relevant disciplines, initiatives, and organizations interested in collaborating on developing a map of knowledge on the safety, assurance, robustness, and trustworthiness of AI and autonomous systems (see https://www.clais.org/home). The German Competence Cluster SafeTRANS has coordinated project incubation and roadmapping in the area of safety of transportation applications, and its most recent roadmap addressed challenges in ensuring safety, security and certifiability of future man-machine systems (see https://www.safetrans-de.org/en/activities/Roadmapping.php). The German Platform for Artificial Intelligence has in its white papers not only addressed the potentials of AI in multiple application domains such as for transportation, health, production, and critical infrastructures, but also pointed out the need to discuss criticality and certifiability of AI based systems (see https://www.plattform-lernende-systeme.de/publications.html). The need for quality assurance of AI based systems has recently also explicated within the Deutsche Normungsroadmap Künstliche Intelligenz (https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf). Multiple projects of the Leitinitiative of the German Association of the Automotive Industry (VDA) funded by the BMBF are addressing Methods, Tools and Processes for supporting safety cases and type homologation for highly autonomous driving, including AI based systems. Why thus awareness of the need for such research is high, solving the underlying deep research challenges calls for a concerted effort of industry and top scientists. No instruments for the required tight integration of industrial understanding of highly promising application domains and leading academics is available on the federal level. In launching this *Joint German Research Center on Foundations for Safety and Explainability of AI based Safety-Critical Applications* we kick-start a major encompassing strategy – The Trusted AI Initiative – for establishing an internationally leading position of European systems industry in the trustworthiness and safety of AI based safety critical applications.

Unfolding this approach from the academic perspective, this calls for a concerted foundational research effort, capitalizing and combining previous and existing large scale foundational DFG funded research initiatives on formal analysis of complex systems such as the Collaborative Research Center AVACS on Analysis and Verification of Complex Systems (participating Universities Oldenburg, Freiburg, Saarbrücken and the Max Planck Institute for Informatics), and the Collaborative Research Center Foundations of Perspicuous Software Systems (participating Universities Saarbrücken and Dresden, and the Max Planck Institutes for Informatics and for Software Systems), and research on analyzability and explainability for AI based systems such as within the Cluster of Excellence on Machine Learning for Science (Universität Tübingen and Max Planck Institute for Intelligent Systems). Pushing the limits of controllability by uniting top researchers to solve the challenges for safe AI will be a key to our digital sovereignty and our competitiveness in the digital economy of the future.

We have integrated leading industrial experts in AI and Safety to complement this academic perspective with research identifying highly promising application areas consciously covering multiple application domains, which all share the need for rigid quality standards and/or certification, notably automotive, rail, energy, industrial automation, and health products. The application scenarios developed by industry will be used as key drivers for academic research, providing yardsticks for measuring the industrial applicability of the proposed foundation research, and also guide prioritization of foundational research.

This roadmap presents academic fundamental research challenges addressing safety and explainability of AI based applications as identified by the academic core team. It then contrasts this with the application scenarios and their characteristics identified by the industrial core team. The last section determines the relevance of the academic challenges in addressing the industrial priorities.

II. Key research challenges

Though there is clearly increasing scientific interaction between the formal methods community and the community pushing foundations of AI, we feel that a concerted effort tightly integrating experts from both communities is required to address the challenge of assuring safety and explainability for safety critical AI based complex systems. We have identified a cluster of 8 groups of research challenges, each group containing some 5-10 challenges, as shown in Figure 1.

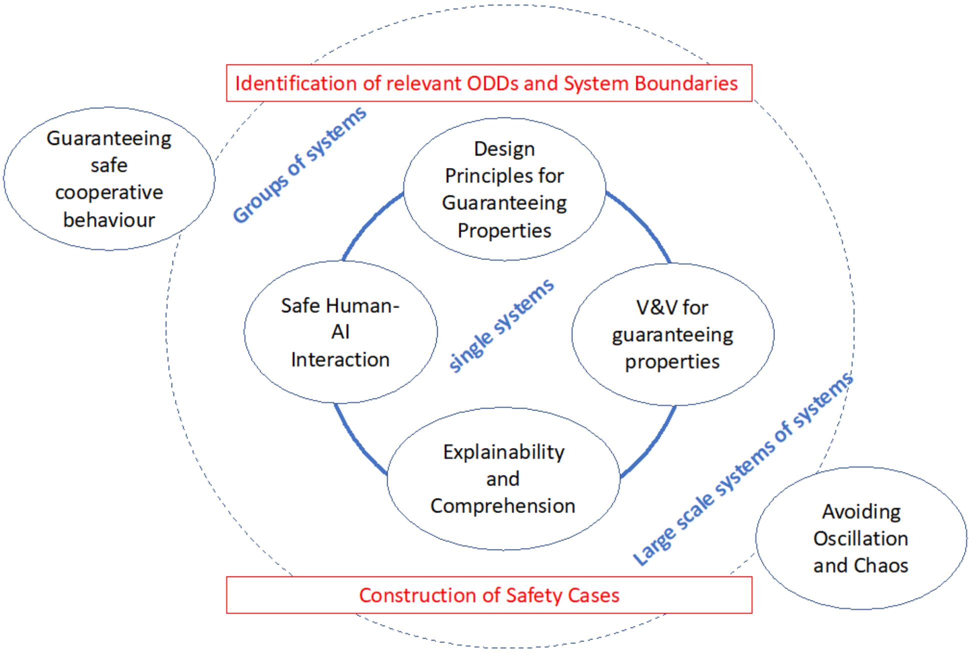

Identification of relevant ODDs and System Boundaries
Guaranteeing safe cooperative behaviour
Groups of systems
Design Principles for Guaranteeing Properties
Safe Human-AI Interaction
single systems
V&V for guaranteeing properties
Explainability and Comprehension
Large scale systems of systems
Avoiding Oscillation and Chaos
Construction of Safety Cases

Figure 1: Classes of Research Challenges to be addressed by JRC-G-SCA, addressing single systems, groups of systems, and systems of systems relying in AI based components. Cross-cutting topics are shown in red. Industrial participants will determine driving applications and their ODDs – research focus here is on deriving key properties of ODDs which can be exploited by methods and processes in all other classes of research challenges, and in identifying the required degree of resolution and confidence of systems in the environment of the ego-systems. Moving clockwise through the inner circle starting at 12 noon, the first class of challenges focusses on constructive methods of assuring safety, complemented by a class of challenges for finding analytical methods to establish safety. Explainability assumes system safety and complements this by methods for explaining actions or advices of AI based systems; comprehension takes this one step further in assuring human understandability of such explanations. Safe Human AI interaction goes beyond understanding of explanations viewing the operator and the AI based system as a team jointly pursuing shared objectives. The challenges arising from extending such cooperation to groups of AI-based systems and/or human operators is addressed on the group level. How to combine such systems to large scale systems of systems meeting system-of-system objectives while avoiding oscillation and instability possibly leading to chaotic behavior leads to additional challenges. The methods, verification evidences, explanations, cooperation and interaction principles investigated in these classes of challenges all contribute to the construction of safety cases for the pilot applications picked by industry, as long as systems are only deployed in their ODD.

The following subsections provide a compact characterization of the research challenges for each of the eight categories shown in Figure 1.

A Coping with open systems – identification of relevant ODDs and System Boundaries

No safety critical system can be designed without a precise specification of

- its allowed deployment contexts, often referred to as operation design domain (ODD)
- its interface to the system environment, including assumptions about the system environment valid in the given ODD, including, if applicable, their behavior

We propose research challenges both using AI and addressing AI-based implementation of perception of artefacts in the system environment to deal with the extreme complexity of environments in the targeted application domains automotive/rail/process control/ medical. These approaches require the availability of large representative data bases DB(ODD,APP) for the targeted ODDs and application classes APP of ground truth percepts of all relevant artefacts of the system environment and the controlled system, including those characterizing ODDs.

RC A

1. ***Relevance***: can we determine, for a given ODD and a given functionality F of the identified application, which artefacts observed in DB(ODD,APP) are **relevant** for performing F? Can we give an operational definition of **relevance** based on causality analysis and causal models?
2. ***Relative artefact completeness****:* can we assess if DB(ODD,APP) contains for all functions F of APP all relevant artefacts?
3. ***Achieving artefact completeness****:* can we develop application domain specific heuristics guiding the identification of all relevant artefacts (akin to the code of practice used in ADAS[1] design)? Can we find methods (e.g. augmentation) to complete the database by generating artefacts on multiple scales from one specific scale?
4. ***Achieving robustness****:* do "small" perturbations of physical attributes of artefacts in DB(ODD,APP) cause irrelevant artefacts to become relevant, or relevant artefacts to become irrelevant?
5. ***Validating ODD characterizations****:* can we assess, for all functions F of APP, and all observations in DB(ODD,APP) meeting the ODD characterization, that F is performed correctly in this observation?
6. ***Monitoring ODD compliance***: are all artefacts required to monitor ODD compliance identified as relevant? Are all observations in DB(ODD,APP) violating the requirements of a function F of APP violating ODD characterizations?

Justification of chosen research challenges:

1. Without providing answers to research challenge RC_A_1, it is impossible to assess the system safety and performance: take as an artificial example that a certain medical diagnosis method is only valid for Caucasian patients, but the database provides no information about the type of patient, then a medical information system could infer an incorrect diagnosis. Or assume that the database contains only the information that this traffic participant is a pedestrian, but provides no further classification such as being an adult or being a child: then a driver assistance system might incorrectly belief, that the default behavior of the pedestrian is compliant with traffic rules and in particular observing ongoing traffic before stepping into a street, while a child will just spontaneously follow a ball it lost to pick it up on the road.
2. If the database is not complete, then there are functions in the applications for which the database does not contain all relevant artefacts. Multiple application domains have developed different heuristics to approximate artefact completeness, based on learning from deployed systems. E.g. for highly autonomous driving, regulations currently negotiated on an international level foresee exchange of information between OEMs and national authorities to learn from incidents or accidents in the field, a process

[1] Advanced Driver Assistance System

established and supervised for civil aviation since many years by the FAA[2]. Methods such as why-because analysis or Hazop Analysis have been applied successfully as heuristic approaches towards achieving completeness.

3. ODDs define an operational umbrella for behaviors of the environment of the system which the system will take for granted when taking actions. Thus, safety cases must demand, that systems are only deployed within environments meeting ODDs constraints. As an example, an SAE level 4 function for driving on highways might specify, that the system is only allowed to be activated on highway segments without construction work; this being a level 4 function, it is the responsibility of the ego-system to detect sufficiently early signs for ongoing construction work in time for a safe takeover of driving activities by the driver while passing the construction site. Or an SAE level 4 function for managing traffic in roundabouts might be specified to be used only in countries, which statistically comply to observering lane boundaries, and mechanisms such as outlier detection must be invoked to disable the system in cities like Rome, where such statistical behavior models are violated.
4. If databases have non-robust characterizations of relevance, then most likely it will be impossible to detect relevant artefacts at runtime, due to the inherent imprecision in sensory systems, and delays encountered in propagating information from sensory data along the perception chain. Each application domain will have different stringency requirements on robustness, and it must be able to assess whether the data base is meeting such robustness criteria.

For AI based systems, which typically rely on artificial intelligence in identifying and classifying artefacts of the environment, a representative data-base meeting the above quality criteria is indispensable for training AI based components, for verification and validation activities, and for certification.

B Safety by Design - Design Principles for Guaranteeing Properties

In modern AI based systems machine learning (ML) plays a key role. While ML enables to solve a large class of prediction problems with unprecedented performance, it is much harder to provide guarantees that a specified behavior of the overall system is maintained under all – including hitherto unseen – circumstances or that, at least, the frequency of deviant behavior is confined to a societally acceptable level. In particular, this is true for applications, where one high quantitative safety targets meet limited or no control over the environment, as in critical cyber-physical infrastructures (with autonomous driving, smart supply grids, or smart health being typical examples). Thus, the following research challenges need to be addressed to ensure a safe behavior of ML based systems:

RC B

1. ***Robustness of decisions*****:** in many applications, it is reasonable to demand that decisions of an ML based classifier should not change under small perturbations, even if they are adversarially chosen, and larger perturbations which are known to preserve the class membership of the instance. Current approaches to adversarial robustness can be separated into certified approaches which guarantee that the decision does not change and empirical approaches like adversarial training which lead to empirically robust networks. Both approaches suffer from that i) current threat models for the perturbations like l_p-balls are detached from the requirements in specific safety-critical applications and are mostly agnostic to the wildly varying safety impact of individual misclassifications, ii) there is no generalization across different threat models which implies that each of them has to be separately enforced, and iii) are neither integrated into an overarching safety process nor linked to run-time safety mechanisms. Consequently, adversarial robustness might be a too strong notion and leads typically to losses in prediction performance so alternatives to the worst-case approach taking into account which instances are both likely to occur in nature and feature a potential safety impact along the causal chain induced by a functional architecture should be explored. Finally, current methods concentrate on achieving robustness for existing neural network architectures. However, a promising

[2] Federal Aviation Authority

direction is the construction of novel architectures of neural networks or other classifiers which yield certified (adversarial) robustness by design or which are at least more amenable to certification, as well as their embedding into an overarching safety-oriented system architecture.

2. ***Reliability of decisions***: a reliable quantification of the uncertainty of the predictions of a classifier or regression model is crucial in safety-critical systems, as is – beyond that – a reliable quantification of the distribution of possible alternative classification. The demand for reliability hereby can take various forms, starting from traditional bounds on statistical moments to safety-oriented polarities of inequational characterizations[3] to reliability figures directly derived from a detailed safety case of an overarching system architecture. Such reliability quantification has to be solid enough to serve the double purpose of supporting both a static quantitative safety case and run-time safety mechanisms, the latter e.g. transitioning to a safe state, or flagging the unreliability of the ML component in a larger system, or handing over the decision to a human e.g. in an automated diagnosis system in healthcare.
3. ***Active detection of scope of validity***: neural networks are known to be overconfident on the in-distribution but in particular for non-task related inputs (out-of-distribution). Out-of-distribution inputs are in particular an issue when the ML system is operating in an open world setting with little control over the incoming inputs (autonomous driving and automated diagnosis systems in healthcare). Thus novel (certified) techniques are required which ensure that either the high uncertainty is correctly flagged for <u>any</u> out-of-distribution input or which detect and flag these anomalous inputs. While calibration on the in-distribution has been studied for some time, rigorous guarantees are largely missing and achieving simultaneous calibration on in- and out-of-distribution inputs is an unsolved problem.
4. ***Certified Integration of Prior Knowledge: in*** several applications, a lot of prior knowledge is available e.g. in the form of physical laws governing the underlying dynamics of the phenomenon to be predicted or in terms of simple rules which are known to ensure the safe behavior of the overall system. While specific approaches exist for incorporating specific differential equations into neural networks, it is an open problem how to guarantee that the resulting system follows the dynamics in case where the full description of the physical system is unknown. Quantifying the trade-off between fitting the data and following the known dynamics can be a stepping stone towards more comprehensive system discovery (in the sense of automatically mining comprehensive system descriptions, an ML approach currently only working in the small) here.

   A related problem is immediate generalization from individual training data points to (uncountably infinite) sets of problem instances based on structural invariants of the problem domain, like, e.g., physical conservation laws. Such form of generalization in learning is routinely pursued by biological organisms (where, e.g., a single situation found to be uncontrollable by the human immediately signifies the same for a vast, in fact uncountable, set of unseen further instances characterized by higher speeds, higher masses, etc.), yet not currently well-understood for algorithmic learning.

5. ***Cyclic learning and active self-learning***: future applications, like long-term autonomous systems, call for adaptation to instationary, slowly changing operational contexts, posing new problems for guaranteeing system properties. From a system design perspective, providing such systems with means of cyclic learning (systems collect data during operation and are sent back to a fresh ML cycle pursued offline, then deploying the updated ML components to the system in the field), passive self-learning (systems collect data during operation and pursue fresh ML cycles online in the field), or even active self-learning (systems in the field pursue active experimentation to detect and understand the changes in the operational context) seem attractive means to provide the necessary adaptivity to changing operational contexts. Means for providing guarantees of safety, reliability, availability, and related properties for such system concepts are, however, even less understood than those for single-shot ML pursued in advance.

[3] essentially, that the likelihood of safety-critical classifications may never be underestimated, while the one of considered-safe, and thus triggering potentially risky actions, classifications may never be overestimated

6. ***Demand-driven design based on safety architectures and safety cases***: in cyber-physical systems, ML components become embedded into complex functional architectures inducing intricate inter-component dependencies and countless and varied fault propagation paths. For traditionally engineered system components, it is customary to decompose the overall quantitative safety, reliability, and availability targets through a stringent safety argument into detailed requirements for components. As this (mostly) top-down decomposition is based on (and requires) a detailed understanding of the expected failure modes of component types, its generalization to ML-based components currently is an open issue. Solving this issue requires both a deeper understanding of the possible failure modes of ML-based components (in particular, when these components come equipped with the extra safety mechanisms detailed above, like uncertainty quantification, out-of-distribution detection, etc.) and means of actively controlling failure modes during ML training. The latter comprises both means of actively suppressing possible failure types as well as actively controlling the ways remaining failures manifest (e.g., as erratic failures vs. fail silent).

Finally, all the above points have to be addressed potentially simultaneously in a safety-critical system. While the ML community has done research in some of the individual points, there is almost no research how to guarantee all of these properties simultaneously. While there is a huge focus on prediction performance, it is ignored that in safety-critical systems we have a multi-objective problem. Thus, it is crucial that the trade-offs in the development of the ML system have to be made transparent to the user/engineer, which should be a part of the certification process of such systems.

C Validation and Verification for Guaranteeing Properties

Verification and Validation (V&V) are crucial activities across system design to demarcate ODDs. Typical activities are formal verification to assure system-level properties, as well as validation techniques, including simulation and testing, that empirically evaluate the conformance of a system design with respect to a specification of the intended behavior. Validation techniques are the mainstream approach to building up trust and ironing out errors during the system design phase. For mission- or safety-critical systems, the corresponding standards additionally enforce the use of state-of-the-art verification. While the foundations and base technologies of V&V are well-established and understood, modern AI based systems involving ML pose entirely new research challenges that need to be addressed to enable V&V of ML based systems:

RC C

1. ***Hardening the role of the specification in AI.*** A common characteristic across all instances of V&V is that their frame of reference is provided by an explicit specification of the intended functionality (e.g., “Never two trains on the same track segment!” or “The risk of more than one power line dropping is below 10-9.”) that needs to be validated or verified. Machine learning generally treats this specification in an implicit and incomplete manner. Grosso modo, in supervised learning the specification is in the labelling of the training data, for unsupervised learning the specification is built into the reward function used for training. This discrepancy is at the root of many problems when aiming for V&V of ML-based systems, and needs to be put in focus.
2. ***Robustness of ML classifiers.*** (see first item under B).
3. ***Robustness of ML in the control loop.*** State-of-the-art program-analysis techniques are not yet able to effectively verify safety of systems comprising ML-based components. Existing work has either focused on the ML-component itself (e.g., with respect to robustness) or on models that invoke the ML-component. Solving this issue requires deep investigations and integration of program analysis techniques based on abstract interpretation for program code and for ML components, especially neural networks.
4. ***Physics-aware and neuro-symbolic AI.*** The stunning success of ML-based methods has limits in contexts where massive amount of training is needed to gain knowledge that actually would be available upfront, be it due to physical laws (e.g., gravity, thermodynamics) or logical reasoning. The

latter may especially be provided by classical, symbolic AI. There is a vast amount of possibilities that are worthwhile to explore with respect to the cross-fertilization of ML-based AI with symbolic AI or contextual physical knowledge, bearing the promise to bring the effectiveness of AI to entirely levels.

5. ***Dealing with uncertainty and rare events.*** ML-based decisions are known to work well for the implicitly defined average case as determined at training time. But this is not enough for safe deployment in critical contexts, where rare events can have drastic impact on system safety.
6. ***Compositional validation and verification.*** One of the conceptually most appealing approaches to master complexity is the compositionality principle. This means that components are studied in isolation in order to derive guarantees regarding their functioning, and that these component-level guarantees are combined along the structure of the system, so as to derive system-level guarantees. This approach is routinely applied in V&V practice, but it is far from trivial to extend to systems with ML-based components, simply because it is not sufficiently understood how their properties will compose.
7. ***Laboratory conditions for verifiable ML.*** A basic principle across many sciences is to make scientific progress first in a well-defined laboratory setting, and then to gradually transfer the results to the real-world, while maintaining a good understanding of the additional complexity added during each transfer step. This makes it possible to understand limits (if it does not work in lab, it will never work in real life) and to control the risks induced by a new technology. This principle is fully compatible with V&V techniques. It needs to be developed and addressed explicitly as a basis for the many upcoming approaches to perform V&V of ML-based systems or components.

D Explainability and Comprehension

A key blocking factor for the use of AI-based components in safety critical applications is the lack of introspection into how decisions are made by AI-based components using non-symbolic methods: how can we trust such systems to take decisions impacting safety, if the very reasons why such decisions are taken are not transparent? How can we understand the causes of mal-functioning, when decisions taken by such components endanger safety or actually violate safety? For safety critical applications involving humans-in-the-loop, how could the human possibly challenge decisions of an AI based component, if the reasons for taking such decisions are not transparent? What kind of explanation would the human need to perform this task? Even if mechanisms can be found which provide such explanations, how can we assure that such explanations are understandable by humans, within the short time frames for safety critical human-in-the-loops applications?

To overcome these blocking factors, a key thrust in solving the challenge of explainability and understandability is required, which should address the following research challenges:

RC D

1. ***Mimicking human explanations:*** what principles of human reasoning and human explanations such as counterfactual reasoning can be mechanized so as to present explanations? Which forms of explanation, ranging from mere causal “happens because” explanations over elicitation of behavioral alternatives circumventing the issue at hand (“happens unless”) to prioritized hints constructively resolving the issue at hand (“best avoided by”), exist and are appreciated by humans depending on the situation? What are the mechanisms needed for generating this variety of explanations? What kind of information should an explanation contain to be actionable, e.g., for verifying or challenging decisions of the system? Does it suffice to only explain in terms of the input (e.g., attribution maps measuring the importance of a pixels or pixel interactions), or do we need to explain the reasoning of the model (but is such an explanation not as complex as the model itself, so what do we gain). Or do we need additional data to contextualize the explanation (e.g., explaining by example). Or should we

target interactive explanations, i.e., explanation which the user can interact with and which allow to explain different aspects of the model behavior?

2. ***Generating sufficient evidences for explanations:*** For a given application class and given ODD, what types of artefacts about the system itself and its environment must be monitored at run-time (see also the discussion on relevance in Research Challenge A)? What measures can be invoked to argue that a given explanation is deemed sufficiently detailed, what levels of abstractions can be accepted to avoid information overflow? How do we know that the explanation really reflects the model behavior, how can we measure it? Is the explanation process itself robust for the given DB(ODD, APP)?
3. ***Integrating Explainability into AI-based components:*** how can we use insights from research addressing the first two challenges to integrate explainability into AI-based components, e.g. by hybrid approaches integrating logic reasoning components mimicking human reasoning, and/or by explicitly enforcing learning of all artefacts considered relevant in human explanations? Where do we integrate it, at the model architecture level (e.g., attention mechanisms) or in the training process (e.g., by augmenting the training loss). How do we find the right trade-off between imposing structure (by integrating explanations) and flexibility (i.e., letting the model learn) and ensure that we do not lose the advantage of data-driven learning?
4. ***Composability of explanations:*** can we provide mechanisms systematically aggregating explanations from sub-systems to explanations for composed systems, e.g. using variants of justification logic? Can we better study the behavior of the composed system by integrating explanations over multiple samples (the whole dataset)? Can we identify inconsistencies in the explanations derived from different sub-systems? ´Can we derive guarantees on the model behavior from the explanations? Or assuming we have guarantees for the systems, can we use them to better interpret the explanation.
5. ***From explanations to comprehension:*** what visual, haptic, acoustic metaphors are able to translate machine readable explanations to information chunks understandable by humans within the constraints imposed by the application context, such as real-time constraints, e.g. through concepts such as augmented reality. How can we establish a feedback channel in order to interact with the system through the explanations?

E Safe Human-AI Interaction

As increasingly activities previously handled by humans are delegated to AI-based components, the level of discourse in Human-Machine interaction moves to increasingly complex and semantically rich areas of discourse, which demand completely new concepts for safe human machine interaction. In particular, an assessment of the degree of comprehension achieved in guiding human perception towards artefacts of the supervised system which demand immediate attention in current decision making, or in assessing the understandability of explanations, is no longer possible without using models of human state, attention, perception, introspection, and decision making for the type of supervision and interaction by human operators for the targeted classes of applications. In particular, finding the essential essence of an explanation which provides exactly the missing pieces to assure a sufficient coherency of the situation awareness of the human operator and the AI-based system is impossible without integrating such human models into the cyber-physical system, thus guiding the process of finding the right level of abstraction to be passed to this operator in this particular situation, taking into account his/her level of training, his/her mental state, level of attention, etc. To be able to argue, on top of this, that all safety relevant aspects of the interaction between operators and systems have been understood by the human operator with high level of confidence scales the challenge in designing Safe Human-AI interaction to yet another dimension of complexity. These challenges have been noted in particular application domains, and even deemed unsolvable with current techniques by some industrial players, such as realizing a safe-handover to the driver in case of automation failures, as foreseen in SAE Level 3: the issue of guaranteeing a sufficient situational awareness for handling control back-to the driver when he/she was previously allowed to be out the loop has been discussed controversially, since the time-frames required to achieve this were deemed to be so long, that one might as well implement full autonomy, i.e. SAE Level 4. This leads to the following research challenges:

RC E

1. ***Human Modelling for Safe Human AI Interaction:*** which minimal set of facets of human modelling of professionals or semi-professional interacting with AI-based systems is required to provide a safe basis for reducing critical information to be communicated to the operator to a level of abstraction which can be transported and comprehended by the operator under the given real-time constraints? Facets include: human state, background knowledge, attention level, situational awareness, refutable default assumptions, level of training.
2. ***Quality Assurance for human models:*** how can we establish sufficient levels of confidence in the validity of such models? How can we effectively trace fluctuating levels of confidence and how can we robustify our human-AI interaction against epistemic uncertainty in the model?
3. ***Implementability and adaptability of human models:*** how can we derive from these such human models which can be integrated into the AI-based system? How can we assess the identified key facets of human models through the system, so as to individualize the model to fit the particular operator in this particular application?
4. ***Guaranteeing safety for Human-AI Interaction:*** how can we assure that a sufficient degree of situational awareness for safe operation of the controlled process can be established withing the given timing constraints, and that explanations of the system for possible operator actions are understood?
5. ***Challenging the AI System***: what mechanisms can be provided to allow the operator to refute or challenge the AI´s systems situational awareness or explanation? How can we measure the degree of discrepancy of situational assessments of the operator and the AI based system? How can we support formation of a sufficient degree of consensus within a given available time interval? How can we support gradual refinement of explanations, focusing on key priorities first, and providing for operator driven elaboration as needed?
6. ***Handover of Control:*** how can we assess, whether the degree of situational awareness achieved through such interactions is sufficient to return certain levels of control to the operator?
7. ***Safety and Stability of Human-AI based systems control loop***: how can we assure, that the interactions of the operator with the controlled system achieve the systems objectives, notably including safety and stability of the controlled system?
8. ***Accountability and Traceability:*** how can we determine in an a posteriori analysis of accountability for and traceability of actions endangering or violating safety and/or stability of the controlled system? In particular, how can we reliably distinguish between human responsibility and system responsibility?

F Guaranteeing Safe Cooperative Behavior

Whenever groups of humans and CPS cooperate to achieve shared objectives in a given role and/or for an agreed time period, such cooperation can only be successful, if all participants in this group share sufficiently similar beliefs about the system to be controlled. As an example, inconsistent views of a sudden loss of CO2 exchange and loss of O2 saturation, such as the surgeon believing that the patient has a heart attack, while the anesthesiologist excludes this cause and beliefs that this is a surgery induced problem such as strong inner bleeding due to inadvertent puncturing of a blood vessel during the surgery, hinder true cooperation to stabilize the patient. Inconsistent assessments of prevailing whether conditions by rescue teams and fire fighters in a large wildfire might cause inhabitants to be evacuated into an area, where particle concentrations are causing dangerous levels of air contamination. Inconsistent beliefs about vehicle speeds can cause traffic accidents. Multiple levels of cooperation must be addressed in order to be able to guarantee safe cooperative behavior, ranging from achieving shared situational awareness, to sharing intentions and plans, to resolving conflict situations, to sharing strategies to achieve goals: increasing the degree of disclosing such information, to systems agreeing to cooperate, will increase the likelihood that their joint resources will allow safe cooperation in reaching shared objectives.

These challenges require as prerequisites, that challenges for explainability and understandability have been solved, and extends the challenges addressed under Safe Human-AI interaction, where one operator and a CPS jointly control some process, to scenarios involving multiple humans teaming up with multiple CPS to control some process. The additional challenges arising in this constellation are many:

RC F

1. ***Shared situational awareness***: how can we assure under real-time constraints, that all team members have sufficiently consistent beliefs about the state of the controlled process?
2. ***Shared mutual introspection***: how can we assure that team members have sufficiently consistent beliefs about the intentions and plans of other team members to achieve shared objectives in controlling the process?
3. ***Achieving safe cooperation***: How can we assure, that individually followed strategies do not lead to blocking moves of other team members but instead merge individual behaviors seamlessly to joint cooperative behavior achieving the shared objectives?
4. ***Achieving safe abortion***: How can we assure that the occurrence of rare events making it impossible to pursue the planned cooperation lead to a safe fallback strategy?
5. ***Negotiating cooperation***: how can we resolve conflicting interests of agents during team-build up negotiations under real-time constraints? Which level of goal sharing is required for successful cooperation? What incentive mechanisms can be used to motivate buy-in, if aggreging to cooperation temporarily impacts my own goals? What level of agreement must be reached: does it suffice to agree on establishing shared situational awareness, or to agree to exchange intentions, plans and/or strategies?

G Avoiding System Oscillation and Instability

A desirable property of machine-learning and the system components employing ML-based functionality is their high level of adaptability, formally expressed by various universal approximation theorems and technically being one of if not the main driver of their application. This flexibility, however, comes at the price of inducing the potential for oscillatory and instable system behavior at a variety of behavioral scales and the associated time scales, ranging from instable feedback dynamics in control applications due to the highly non-linear transfer functions implemented by, e.g., neural networks to long-term oscillations and drift in self-adaptive systems. These behaviors are hard to control, and guarantees for their absence consequently hard to obtain, due to both the complex transfer functions implemented by machine-trained components (e.g., a recurrent neural network with its interaction between – potentially even gated – recurrence and a topologically complex, at smallest scales alternating between convex and concave as well as non-smooth gradient changes, state evaluation function mediated by a neural network) and the imprecise relation between the mathematical formulations of the optimization goals underlying algorithmic learning and their actual realizations via heuristic optimization methods (like the gradient descent driving backpropagation). Overcoming these problems and thus being able to provide the technically and societally required guarantees on system stability across the aforementioned time scales poses a number of research challenges, the prominent ones being:

RC G

1. ***Stability verification of ML-in-the-loop control systems*:** While the automatic safety verification of merely feedforward neural networks against robustness criteria or behavioral specifications is in its infancy and

conception of their generalization to recurrent structures just commences, automatic or interactive stability verification asks for an even significantly broader scope of verification. Verification of stability properties of embedded ML-in-the-loop applications inherently has to cover the joint feedback dynamics of a recurrent neural network structure and its environment, the latter comprising further software and hardware artifacts as well as controlled physical processes. The intricacies of such an endeavor obviously go well beyond the aforementioned ML verification approaches addressing a neural network in isolation.

2. ***Optimization-theoretic characterization of ML-based self-adaptation***: Control theory has over recent years seen a shift from algorithmic and analytic characterizations of automatic control to optimization-theoretic formulations. The attraction of the latter perspectives is that it provides a move from the "how" to the "what", that comes especially handy when guarantees on the various dynamic aspects of control, like stability, tracking accuracy, cost, etc., have to be given. In such settings, the optimization-theoretic formulation permits a separation of concerns between the construction of controllers ("how do I synthesize a controller implementing the required optimization?"), the properties of the construction process ("how close does the optimization come to optimally solve the optimization problem?"), and the dynamic properties of the constructed controller ("what dynamic properties of its feedback behavior follow in the given environment from the controller providing an (almost) optimal solution?"), rather than blending the three into a single-shot, end-to-end algorithm verification problem ("given this control algorithm, what dynamic properties does it exhibit in the given environment?"). In that respect, the optimization-theoretic formulation provides an independent specification of the construction. Similar optimization-theoretic characterizations for long-term self-adaptation by self-learning components are, however, currently lacking, yet could provide a similar move from the "how" (i.e., foreseeing all the self-adaptations that could ever occur) to the "what" (i.e., a situation-independent closed-form characterization of the goals of self-adaptation), thus providing an advance in both our mathematical understanding of self-adaptation and in development of a corresponding tool set for analysis.
3. ***Proving stability properties of ML-based self-adaptive systems***: The aforementioned optimization-theoretic characterization is a stepping stone towards decomposing the overall problem of stability verification of ML-based self-adaptive systems, yet does not in itself solve this problem. The development of practical and scalable methods for analyzing dynamic properties of self-adaptive systems, especially in an open-world context not strictly delineating the range of possible worlds to be adapted to, remains another challenge.
4. ***Ensuring stability of interacting systems of individually learning systems:*** Systems learning from observations in the field (either in the form of cyclic offline learning or directly through online learning in-situ) are prone to oscillations due to the instationarity of the individual component's environment, which alters behavior due to concurrent learning processes. The number of scenarios where this effect may emerge is vast, covering dedicated adversarial learning schemes just alike systems of adaptive multi-agent systems. Means of controlling oscillations in learning are, however, not well-understood and would have to borrow from a broad set of scientific disciplines, ranging from cognitive theories of learning in zoological herds (like temporal and hierarchical symmetric breaking mechanisms inherent to such herds/flocks) to dynamical systems theory (e.g., generalizing notions of diffusive dynamics to learnt facts).

H Construction of Safety Cases

Systems certification aims to provide assurance to stakeholders that the underlying system meets a specific degree of quality, with respect to a set of properties and corresponding metrics. The typical approach to certification involves applying standardized holistic methods for providing such assurance e.g., via verification and validation. Over the course of development, assurance activities contribute towards certification by building models and collecting analysis results and other artifacts that support claims of safety, reliability and similar properties. Regarding safety, functional safety standards, such as IEC 61508, were developed. Safety standards widely share the common concept of a Safety Integrity Level (SIL), which offers a combined means of describing the rigor of required assurance activities, commensurate to the level of risk needed to be addressed. However, applying standard guidelines is not enough to guarantee success, as appropriate tailoring and management of the development process is still necessary. A safety case comprising a safety argument is a means to structure all these safety assurance activities and to thus demonstrate that an acceptable level of safety is achieved by the system. This particularly implies a corresponding level of confidence with respect to achieving safety, which is reached by the systematic structuring and alignment of all assurance related aspects as part of the argument.

The conceptually appealing concept of safety cases comes with a number of important challenges:

RC H

1. ***Mastering complexity of the safety argument.*** The safety argument is particularly important when safety standards cannot describe a one-size-fits-all solution. This is the case for AI (and machine learning in particular) as safety assurance will need to be very specific to the application domain, the system under development, the function that is realized with AI and the context in which the function used, i.e., the ODD. Safety case and safety argumentation provide a basis to structure the complexity of safety assurance and to communicate across all the involved stakeholders including certification bodies. This chiefly needs to tie in with validation and verification technology to back the structuring of the arguments. If correctly applied, a safety case can also be used to show compliance with relevant guidelines, standardization and legislation, and make transparent cases where the argument contains certain trade-off decisions against societal agreed criteria or regulation. Several modelling languages like the Goal Structuring Notation (GSN) or the OMG Structured Assurance Case Meta-Model (SACM) are being proposed structure safety arguments. These merely syntactic structures however need semantic underpinnings in order to faithfully apply divide and conquer strategies.
2. ***Quantifying and minimizing out-of-distribution risk.*** ML introduces additional challenges, in terms of the lack of explicit specification (now found in the training data and learning process), and of its lack of predictability with respect to input perturbation. This lack of predictability and transparency is challenging for assuring and certifying safety-relevant properties of ML components. At the core of the problem is the need to establish quantifiable guarantees for risk-aware automated decisions. Methods tackling this problem are being developed in the research community but are thus far confined to clean-room lab conditions. The gap to more realistic, more relevant and more precise real-world scenarios is moving in focus. Due to the inherent complexity, mechanisms which monitor the operational conditions of an ML-powered system and adapt its operation according to the estimated uncertainty are also needed. In doing so, the ML uncertainty must be estimated with a sufficient level of confidence. This results in complex probabilistic calculations for which proper guidelines are missing. Moreover, for many methods and techniques it is hardly possible to evaluate them with respect to their quantitative impact on risks. They are applied to assure that risks satisfy the 'as low as reasonably practicable' (aka ALARP) risk criterion. However, the 'best effort' implication of this criterion is not sufficient as guideline for applying these methods in a particular case and for considering complex dependencies between these methods. Needless to say, the application of the methods and techniques does not only require guidelines but also certified tool support.
3. ***Environmental factors.*** With respect to the complex environments, a safety case has to exploit the modeling and analysis results produced from Operational Design Domain (ODD) activities as the ODD defines the part of the environment where safety claims about the autonomous behavior are possible. ODD analysis can be used in combination with risk analysis to identify critical situations required to be explicitly addressed by the autonomous system's development. As such, they can establish detailed goals which the safety case should explicitly address e.g. by linking such goals to results of corresponding validation testing for the critical situations, as well as the acceptable completeness and correctness of the ODD. A particular challenge is to properly integrate the huge range of environmental factors of the ODD in the safety case.
4. ***Recertification.*** As the safety case is limited to the ODD and the considered environmental factors, it becomes invalid if the environment changes over time and introduces new safety-relevant environmental factors. Further, unknown epistemic gaps about the environment and security issues may require continuous engineering of the safety case and the autonomous behavior. And even the system itself might adapt or change over time (e.g., due to critical security updates or continuous learning) which also implies the need to reconsider and manage safety assurance and certification. A key challenge is to establish corresponding continuous engineering frameworks as well as to investigate means and capabilities for systems to conduct automated online re-certification after relevant changes in the ODD or in the system itself. Construction, representation and management of safety cases will be a key ingredient in this regard.

5. ***Ethical dimension*.** As autonomous behavior automates complex situation-specific decision making, it is a notorious challenge to define which decision or action is morally right or wrong in which situation. The consequences of an action are typically not certain, and different consequences can occur with unknown likelihoods. This makes it hard to argue in the safety case that an autonomous system will always choose an action where the related risks are acceptable. This comes with a perceived dilution of the moral responsibility of developers for failures. Safety cases can be adapted to address such concerns, extending upon the work related to ML explainability and comprehension (see above). In doing so it has to comply with social agreed criteria / regularia and make trade-offs decisions against these criteria transparent.
6. ***Privacy vs. Accountability.*** Another specific concern that needs to be addressed with respect to ML-powered applications is assuring that the means of data collection for training or logging purposes comply with the individuals' right to privacy. For example, when training autonomous commercial vehicles, large datasets from driving on real roads need to be compiled, including for the purpose of identifying pedestrians. An extended safety case could also argue that privacy-preserving ML techniques such as federated learning, differential privacy, and image obfuscation are adequate for the protection of pedestrian individual's identity while still enabling the reconstruction of the individual accountabilities, in case an accident has unfolded. Similar arguments would have to be included for assuring the security and appropriateness of the transmission and storage of such data, accounting for legal and regulatory requirements e.g. GDPR.

III. Industrial Challenges

At a Trusted AI workshop on Safety and Explainability of AI based Safety-Critical Applications, industrial representatives from Bosch, SAP, Siemens Mobility, Siemens Corporate, Trumpf, VW, and ZF presented their companies priorities in establishing safe, explainable, trustworthy systems incorporating AI components in key system functions.

This chapter summarizes the key findings of this workshop, through integrating the proposed challenges along four major categories:

I Overarching High Level Challenges
II Process related Challenges
III Challenges related to Design Principles, Methods, and Tools
IV System Challenges

I Overarching High Level Challenges

Artificial Intelligence (AI) techniques and algorithms, and especially Machine Learning (ML) techniques, which are currently in the focus of attention, promise huge benefits for mankind, both on a society level as well as on individual level. At the same time, they enable machines to take decisions that (a) directly impact human's lives and well-being and (b) which were previously taken by humans only. In addition, training, testing and using ML techniques requires and produces huge amounts of data, where this data may be directly mapped to individual human beings, thus facing privacy concerns.

Using AI in safety-critical applications and systems, i.e., in systems, whose failures may endanger human's health and lives, or cause economic or environmental damage, places even higher requirements on the quality and 'correctness' of these algorithms, because the potential damage caused by 'bad decisions' is so much higher than in non-safety-critical applications.

Overarching challenges that need to be overcome to make AI/ML algorithms amendable in and for safety-critical applications thus fall into three categories:

How to **effectively implement** overarching AI ethics requirements such as

I.1.1 **Transparency**: Humans must be enabled to understand why and how the AI algorithm is deriving its decisions.
Related qualities:
explainability, explicability, understandability, interpretability, communicability, disclosure (see also III.3.4)

I.1.2 **Fairness**: Decisions taken by the AI system must be fair

Related qualities:

Justice, consistency, inclusion, equality, equity, (non-)bias, (non)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution

I.1.3 **Safety and Security**: Decisions of the AI-system must not cause harm
Related qualities:
non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion

I.1.4 **Responsibility**: Responsibility for effects based on AI-system's decisions must be clearly defined.
Related qualities:
Responsibility, accountability, liability, acting with integrity

I.1.5 **Privacy**: Privacy considerations for data needed to train and test an AI system and data produced by the AI system must be observed
Related qualities: Privacy, personal or private information
more general: data protection (also including non-personal data, like IP,...)

Additional requirements on the usage of AI algorithms, which might have different weight depending on the concrete system to be developed:

I.2.1 **Supporting Well-being of users**
Qualities: Benefits, beneficence, well-being, peace, social good, common good

I.2.2 **Supporting Autonomy of users**
Qualities: Freedom, autonomy, consent, choice, self-determination, liberty, empowerment

I.2.3 **Trust (of users into the system)**
We will address the issue of narrowing down the description of the desired properties of the AI applications to a list of domain independent and measurable key trustworthiness indicators (KTIs), that can operationalize the application of key trustworthiness principles (KTPs) such as accuracy, interpretability, robustness, fairness and data efficiency. The objective is to build measurable KTIs that can assess the trustworthiness of AI-based systems, essential to scale-up deployment, making them: more predictive, more explainable, more stable, more inclusive and more efficient. The KTIs can be used to assess AI learning models, providing a statistical test that can indicate, for example, whether a specific model has a robustness metric higher than a set threshold (or not), and similarly for interpretability, fairness and data efficiency.
(all challenges addressed in challenge groups II, III, IV are directed to support this overarching requirement)

I.2.4 **Sustainability**
Qualities: environment (nature), energy, resources (energy)

I.2.5 **Dignity**

I.2.6 **Solidarity**
Qualities: Solidarity, social security, cohesion

I.3 How do we efficiently and effectively build AI-based systems (a.k.a. intelligent systems) that satisfy industrial-grade qualities?

I.3.1 How do we verify and validate, that such intelligent systems are reliable, available, and safe? (Refined in all subsequent challenge groups)

I.3.2 How do we keep industrial-grade quality when systems become intelligent and improve trustworthiness for users? (Refined in all subsequent challenge groups)

II Process related Challenges

Established development and lifecycle processes have been conceptualized and designed for the needs and specifics of “traditional” non-AI/ML systems. The same is true for relevant standards and guidelines such as the key functional safety standard IEC 61508. Adequate guidance regarding lifecycle processes for systems with AI/ML components is required, correspondingly there are different ongoing activities e.g. in the field of safety standardization.

II.1 One particular aspect that needs to be taken into equation for systems with AI/ML components is how **data is addressed within the lifecycle**. Both, data used for training (and testing) ML components during the design phase as well as data collected during operation needs to be attained (it might be recorded, generated, tailored and/or augmented), qualified (e.g. with respect to completeness, non-biasness, fitness), documented, labeled, stored, maintained, secured - etc.

II.2 A further aspect where lifecycle processes for systems with AI-components will deviate from established ones is **ensuring that an AI-based function operates as intended**.

II.3 Releases should be required to pass **maturity gates** and they should meet **meaningful key performance indicators**.

II.4 The importance of an **explicit operational design domain (ODD)** is increasing and **ODD compliance** should be monitored.

II.5 **Explainability approaches** might be used during development to detect flaws within a learned model and to inform a directed retrain for mitigating these flaws.

II.6 Consequently, an appropriate development process shall **support agile and frequent iterations in AI development** and validation, including save and secure updates of systems already operating ‘in the field’, to cope with defects and failures detected during operation.

II.7 Regarding validation, a **systematic and thorough validation approach** needs to be enabled by, for instance, utilizing the ODD concept and/or information gained through the application of explainable AI approaches. Validation must ensure key system qualities, especially safety, availability, reliability (i.e., dependability), privacy, and security, but also additional qualities especially associated with AI components, like resilience, etc..

II.8 Any information relevant across the system lifecycle shall be captured, stored and maintained in an appropriate way. **Digital dependability identities (DDI)** are a concept to achieve this by means of a model-based approach, thus the DDI concept should be extended towards the specifics of ML-based systems.

II.9 For ML-based functions the acquisition of training (and validation) data is of utmost importance. **Continuous engineering and DevOps type lifecycles** therefore gain attractiveness, a system might collect data during operation which is fed back into development to continuously improve the system. In this regard, scenario-based evolution, e.g., by step-wise increasing capabilities of intelligent systems, can be an interesting way to go. At the same time, the respective ODDs should be sampled systematically.

II.10 On a general note, future development and lifecycle processes shall be designed to be **auditable**. They shall be **traceable, reproducible and measurable**. Ideally this holds true for both the processes as well as the resulting products.

III Challenges related to Design Principles, Methods, and Tools

We have structured the presentation of these challenges into the following subcategories:

III.0 Scoping of ODDs
III.1 How to ensure representativity of data used for training
III.2 How to derive specifications for ML based systems
III.3 How to test, validate, verify ML based systems
III.4 How to ensure interpretability of ML based components
III.5 Safe techniques for AI compression and deployment
III.6 Safe techniques for system optimization
III.7 Compositional safety analysis of systems including AI components

III.0 Scoping of ODDs

According to the SAE: J3016 standard, an ODD (Operational Domain Definition) comprises the "operating conditions under which a given [system] or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain [environmental] characteristics." The SafeTRANS Roadmap on Safety, Security, and Certifiability of Future Man-Machine Systems specifically describes how, by restricting ODDs of these systems appropriately, i.e. placing constraints on operational conditions, environments, cooperation modes, and similar, highly automated (AI-based) systems can be build and validated to operate safely in such restricted environments.

Restricted or so-called **narrow ODDs** can be specified for meaningful applications -- one of many examples is the Metro line 1 in Paris which runs fully automatically, and where it is guaranteed by construction of the railway and the stations, that no humans nor other trains can enter the driveway of the metro train, such greatly simplifying the ODD. Application of narrow ODDs is limited, however, and guaranteeing that the real-world operational domain is indeed as narrow as specified by the ODD -- ie., guaranteeing that indeed no human nor other train can enter the driveway of Paris Metro Line 1 -- is expensive. Some of the methods and algorithms used to validate systems designed for narrow ODDs can also be applied to validating systems for wider, i.e., less restricted, ODDs, but most of the do not scale well to be applicable for wider ODDs. For **wide ODDs**, however, it is often not even known how to validate systems at all, thus rendering a multitude of business cases for these ODDs unrealizable.

For example, for Low-Speed Automated Driving (LSAD) systems such as pods and shuttles, the ODD may include urban areas with predefined routes that include pedestrians and cyclists. On the other hand, for a motorway chauffer system, an ODD may include a four-lane divided motorway and dry conditions only [1]. The types of scenarios a vehicle may encounter will be a function of its defined ODD, making this fundamental to any safety evaluation and scenario identification.

Many “Narrow ODD” business cases exist (e.g. a freeway pilot system for cars or an automated tram in protected areas). They can often be specified and solved with (comparably) simple technology, allowing for (comparably) straightforward homologation and safety guarantees. They often rely on simple but effective infrastructure rooted sensors, measures and close the system to protected traffic, eliminating interaction with vulnerable road users. “Wide ODD” business cases exist but are not solved yet. Wide ODDs often cannot be specified by logic and rules. Instead, ODD boundaries and state space decision boundaries are learned from data. Nevertheless, ODDs must be constrained as much as possible to allow for safe operation and combined with traditional traffic safety technology where possible. In wide ODDs, the challenges are related to safe and trustworthy technology and homologation ecosystems for AI and ML. R&D in this area includes confidence modeling, which aims at detecting out-of-distribution cases and ODD boundary violation.

AI-algorithms governing system behavior need to cover the complete ODD, be it wide or narrow. However, since any ODD contains an infinite set of scenarios and scenario variants, **confidence in AI decisions** has to be determined. A possible approach is to determine the ‘distance’ of the current input data for the AI algorithm to the data that has been used to train and/or test the AI algorithm, where ‘closeness’ would typically determine ‘confidence’. These kinds of measurements are directly related to the robustness of the AI algorithm under consideration: Slight perturbances in input data should only produce slight differences in output data; this should hold for any pair of input data from the ODD, but especially at the ODD border. The subsequent challenges elaborate these observations.

III.1 How to ensure representativity of data used for training

The quality of machine learning results and the trustworthiness of any application based on these results rely on the data preparation process and on the quality of the datasets used for training. The machine learning processes require various forms of data preparation, correction, and consolidation, combining complex data transformation processes and cleaning techniques. The data input to the ML algorithms is assumed to conform to "well-behaved" data distributions, containing not significant biases or inconsistencies. Data quality is a multidimensional, complex, and morphing concept[4]. Since the beginning of the Century, there has been a significant amount of work in data quality management initiated by several research communities: database, statistics, knowledge engineering, and AI. As a result, many data quality definitions, metrics, models, and methodologies have been proposed. Our goal must be to consolidate these approaches and to extend them in order to meet the full range of requirements for machine learning (in particular deep learning) and hybrid AI. Based on best practices, standards for the quality of ML data need to be developed and adopted across domains.

The Fraunhofer Guidelines for High Quality Data and Metadata (NQDM) of 2019 lists the following dimensions of data quality:

1. Currency: Data describes the current reality. Therefore, it is recommended to pay attention to a timestamp and, if necessary, a version number when recording and naming the data.
2. Accuracy: The data should contain correct values and be as error-free as possible. Here a datum is faulty if it does not correspond to its classification.
3. Precision: Depending on the application, the precision of the data is of high relevance, so that, for example, rounding of values should be avoided. The content descriptions of the data should also be as precise as possible in order to quickly assess the relevance of data.
4. Conformity: When providing data, attention must be paid to the expectation conformity of the contained information in a certain usage context and format, for example when naming attributes and vocabulary. For a universal use of the data, appropriate standards should be used where possible, e.g. ISO 8601 [91] for dates.
5. Consistency: Data should be free of contradictions, both in itself and across data sets. This dimension may already be covered by accuracy.
6. Transparency: and trustworthiness The origin, originality and changes to the data should be made traceable, so that the transparency and credibility of the data can be strengthened, thereby gaining the trust of the users and also meeting ethical requirements.
7. Reliability: In order to assess the reliability, or the degree of maturity, of a piece of information, it can be assigned a status (see also DCAT-AP.de).
8. Understandability: The data structure, the naming of the data, as well as data interfaces should be easy to understand.

[4] Berti-Equille, L. (2007). Measuring and modelling data quality for quality-awareness in data mining. In *Quality measures in data mining* (pp. 101-126). Springer, Berlin, Heidelberg.

9. Completeness: A data set should be complete; Attributes, which are mandatory for the further use of the data set, must therefore contain a value.

The very nature of ML based system design, where classical system implementation is replaced by training machine learning component from data, makes it mandatory to apply the highest quality assurance measures to gathering and maintaining such data. In particular, key activities related to specification development, must be redirected and modified in order to ensure, that the implementation, whose specification is now only given implicitly through data and the training phase, is meeting the requirements for the system component to be implemented with machine learning techniques. This leads to the following seven subclasses of challenges:

III.1.1 Ensuring completeness of observations
III.1.2 Ensuring representativity of distribution functions for relevant data
III.1.3 Coping with rare events
III.1.4 Coping with concept shifts
III.1.5 Ensuring robustness of training data
III.1.6 On-line Monitoring deployment contexts
III.1.7 Data efficiency

III.1.1 Ensuring completeness of observations

For (safety-) critical system development, a precise **characterization of the Operational Domain** (typically referred to as ODD for Operational Domain Definition) is indispensable. Classical system development processes require the **characterization of the system interface**; this subsumes an identification of all relevant artefacts in the environment of the system, which must be observable in order to meet system requirements. Translated into the ML context, this calls for the derivation of a taxonomy of all artefacts which must appear in the characterization of the operational domain: what type of ground truth labels must be associated with the date used to characterize the operational domain? Often, such taxonomies are organized hierarchically; a challenge in system design then lies in finding the right level of detail: **how many labeling classes do we really need to implement the system? How can we avoid, that too detailed labelling classes increase the risk of misinterpretation? What is the right balance between the risk of confusion and the needs in environment perception required for proper function implementation?**

III.1.2 Ensuring representativity of distribution functions for relevant data

While these challenges focus on **what** has to be observed, representative of data requires a second key quality gate to be achieve: the distribution function of the occurrence of such artefacts must match their distribution in real-life ODD-compliant application scenarios. E.g. in highway chauffeur applications for autonomously driving vehicles, pedestrian are very unlikely to occur, but still relevant, in contrast to robo-taxi applications, which must have representative distributions functions for pedestrian on crossings, on the pavement, around a bus-stop, etc. Note that such distribution functions will be culture dependent – e.g. be different for data gathered in Bombay and in Berlin. **Assuring the representativity of distribution functions for all relevant artefacts is thus a major challenge**. It must avoid having undesired **biases,** and it must guarantee that rare but possible **edge cases are covered.** Such distribution functions may be time dependent (such as rush-our traffic vs night traffic).

III.1.3 Coping with rare events

A related problem is reliability in rare but dangerous situations. In Highly Automated Driving (HAD), for example, cars must respond appropriately to events such as a child running across the road. Obtaining sufficient training data for such rare events can be very costly, ethically unacceptable, or even impossible. Lack of robustness can impair models also when specific data are input to the model, possibly leading to biased decisions.

III.1.4 Coping with concept shifts

A further challenge is caused by the fact, that the system´s environment is itself evolving our time: what might be representative in the 2000 for urban environment will no longer be representative in 2020, since e-scooters appeared in the meantime. This entails the challenge, that above quality criteria – which artefacts have to be represented with which distribution functions – becomes a **life-cycle assessment of data quality assurance**, enforcing re-training of ML based components as **concepts shift** (thus requiring safe over-the-air updates of such components).

III.1.5 Ensuring robustness of training data

Robustness can be defined in terms of limited variation between the predictions generated by a machine learning model based on non-perturbed data and the predictions generated by a machine learning model involving perturbed data. Neural network models excel at processing inputs that are similar to training data, but their understanding of "similarity" can be seriously different from that of a human domain expert. For example, in computer vision, visual recognition of human motion and gesture (in traffic in other domains) often fails for poses, occlusions, and motion classes that are very different from training poses and scene backgrounds, and even the slightest image changes can lead to drastic errors in visual recognition.

It is well known that small perturbations in data for object perception can lead to safety-relevant misclassifications. Such perturbations can be caused purposely by attacks on the system, or could arise from noise or in general sensor imperfections. **Ensuring robustness of training data** thus is a key challenge, leading to research questions such as “how can we determine the safety impact of adversarial vulnerability”, “which perturbation classes have high significance”, “what is a safety metric for determining the required degree of robustness against such perturbations”.

III.1.6 on-line Monitoring deployment contexts

A safety critical ML based component must only be activated in application context which match the quality requirements on data and distribution used during the training phase. To detect out-of-limit and in particular ouf-of-ODD usage as well as concept shifts, any such component must be safeguarded by on-line monitoring of the compliance of actually observed data and distributions to those established at the data-quality assurance phase. We thus face the challenge to **generate on-line monitors reliably detecting out of scope environments** for ML based components.

III.1.7 Data efficiency

Data-efficiency is defined as the increase in the accuracy of a machine learning technique as a function of the size of the training set. Data efficiency is important in domains where training data is only scarcely available (e.g., diagnosing rare diseases) or only available at substantial costs (e.g., the automotive domain). Unfortunately, most modern machine learning techniques, and deep learning in particular, are characterized by a low data-efficiency. For example, AlphaGo played 4.8 million training games, and NVidia's face generating GAN was trained on 10M samples. Theory tells us that for purely data-driven learning, computing needs to scale with at least the fourth power of the improvement in performance. In practice, the actual requirements have scaled with at least the ninth power (Thompson et al, IEEE Spectrum, Sep. 2021), which means that a system would need more than 500 times the computational resources in order to half its error rate. All this is clearly unsustainable and is in stark contrast with human learning where only a few training samples are often sufficient to learn accurate and robustly generalizable classifications or predictions. Human learning is not purely data-driven but benefits from rich sources of background knowledge to guide the learning process. We will define new measure of data-efficiency that include the costs and benefits of background knowledge, as encoded for example in strong symbolic priors and semantic loss functions, allowing us to measure the data-efficiency of neuro-symbolic systems that exploit such background knowledge as part of their learning.

III.2 How to derive specifications for AI based systems

As machine learning often is used as a constructive means of mining and operationalizing implicit knowledge from observations, a fundamental complication of validation and verification approaches, be they formal or semi-formal, for AI is **lacking availability of explicit specifications**. For the time being, it neither is clear how specifications or candidates for such could also be mined from case/observation-based reasoning akin to, but independently from the training of the AI components, nor what other principles could be used to establish concise and unambiguous specifications of desired system properties. This problem is exacerbated further (a) in settings where **(semi-)formal specification languages for the properties at hand are hitherto unknown** - and sometimes hardly conceivable -, as in image classification tasks or for ethical principles, or (b) when epistemic or aleatoric

uncertainties, presence of outliers, or other sources of **data quality issues**, like problematic coverage of rare or borderline situations, provide additional noise that would have to be eliminated in any specification elicitation process, be it manual or mechanized.

III.3 How to test, validate, verify AI based systems

Presently there is a tendency to put special emphasis on validation activities for ensuring key properties of ML-based systems. This is due to the fact that there is often no complete requirements specification for ML-Systems and the learned model is not human understandable and hard to analyze. Established approaches of verification along the left side of the V-Model are thus not applicable and the hope is to mitigate this problem by means of a strong validation.

However, a proper validation is very challenging as well for mostly the same reasons. Without a sound and complete requirements specification it is hard to systematically derive test cases. In addition, ML components might exhibit significant changes in their outputs based on seemingly insignificant changes of the inputs. We are therefore often left with a black box ML component, a huge input data space and no means to systematically create equivalence classes to tackle the combinatorial complexity of testing.

To overcome these challenges, we require a better understanding with respect to **systematic V&V methods for ML-based systems**. Integrated Systems- Dependability- and ML-Engineering methods shall be developed that provide optimal traceability across the engineering processes and products and specifically enable a **systematic validation (including objective pass and fail criteria based on e.g. quantified performance)** as well as utilization of generated evidence as part of **sound dependability argumentations (i.e. assurance cases)**.

Due to the complexity of ODDs, means are required to efficiently and effectively conduct a large number of tests. **Test automation and the utilization of simulation-based approaches** can be of great help here. In addition, methods are required to **enable thorough V&V of developed AI / ML also at the edge of the ODD**. This is particularly challenging if the ODD is large and complex, and there is the corresponding risk of unknown edge cases. Systematic identification of unknown cases by bootstrapping market introduction and taking advantage of function-monitoring by using human-in-the -loop decisions in operation is required. The runtime aspect (continuous engineering, monitoring, runtime assurance, e.g.) will generally gain importance, **continuous quality and dependability monitoring and assurance of AI in operation** might remediate some of the challenges encountered at development time.

In addition, the **possibility of unified function release standards** for formal domain-wide product releases shall be investigated. Such a function-oriented V&V (and maybe certification) approach should also be complemented with respect to V&V activities concerning data. **Standardized data quality methods** to enable formal assessment for release shall thus be investigated as well.

III.4 How to ensure interpretability of ML based components

Interpretability can be defined as the capability of a machine learning model to explain predictions with a (limited) set of predictors. The challenge rests in defining KTIs that can measure the interpretability of the predictions made by AI, improving transparency. Machine learning models such as neural networks, random forests and gradient boosting often predict the desired outputs very accurately, but it is generally difficult to tell what the model has learned and why it produces the output it does ("black box").

Interpretability must support all of the following use cases:

III.4.1 How to generate **justifications** for decisions/actions of ML based systems allowing causality analysis in case of system failures

III.4.2 How can such justifications be translated to **explanations** interpretable by human operators for à postiori failure analysis?

III.4.3 How can such explanations be tailored to background knowledge of human operators so as to support **real-time interactions** between AI based systems and human operators

III.5 Safe techniques for AI compression and deployment

Embedding AI into cyber-physical environments will often induce rigorous **constraints on its resource consumption** concerning runtime, computing power, memory (for programs, constants, data), energy, precision of arithmetical operations, permissible data types, etc. This in turn induces a need for **optimized implementation of AI software that reflects these constraints** by various forms of compression, in particular due to trivialization and elimination of parameters (like non-zero waits in neural networks), consequential elimination of connections and computations, as well as due to compression of bit-widths and numerical data-types. Such compressions are essential to deployability of AI components into the massively resource-constrained, hard real-time application domains, yet obviously modify the behaviour of the components not only w.r.t. these resources, but also w.r.t. their functionality. This induces a quest for either **automated approximate-equivalence-checking**, rigorously ensuring functional equivalence up to a user-defined tolerance margin across the whole input range, or for **constructive means providing maximal compression** while guaranteeing such user-defined approximate equivalence.

A related question is how **updates to already deployed AI components** can be compressed, communicated in a bandwidth-efficient manner, and safely patched into the AI system in settings of cyclic offline learning or distributed in-situ learning.

III.6 Safe techniques for system optimization

Safety is only one out of many properties and a safety architecture has typically not only impact on safety but also on other functional and non-functional properties. Therefore, analysis and comparison of suitable safety architectures is required to come up with viable and optimized products and services. Exploring different architectural designs is not novel in safety engineering but the existing approaches need to be extended with respect to typical safety patterns for AI and autonomous systems like Doer-Checker, Simplex architecture, runtime monitoring, Failover, and related extensions to deal with inherent uncertainties in the environment as well as uncertainties of AI/ML components.

In addition to optimization before deployment, continuous engineering processes shall enable to optimization after deployment by collecting field data in order to replace worst-case assumptions holding in any ODD with valid assumptions in the considered ODD.

III.7 Compositional safety analysis of systems including AI components

The **heterogeneity of system architectures** including AI components will generally necessitate **heterogeneous means of specification and verification** also, as e.g. apparent from the frequent combination of computer vision components (no closed-form specification, statistical verification based on samples being the method of choice) with engineered safety mechanisms (mathematical specification available, formal verification possible) within the same overall system architecture. An obvious, but hardly understood consequence is the **need to integrate such heterogeneous forms of requirements and arguments into an overall safety case**, especially when quantitative arguments cannot easily be decomposed due to complex stochastic dependencies, uncontrolled confounders, or even risk of undetected correlations and confounders. All these effects prohibit a straightforward decomposition of the overall safety case into component-local cases, **necessitating novel forms of compositional safety analysis** for large and heterogeneous system architectures.

IV System Challenges

We have grouped the challenges regarding building ML based systems in the following categories

IV.1 Safety Architectures encapsulating ML based subcomponents
IV.2 Adaptable and evolvable architectures for ML based systems
IV.3 Stability of AI based control
IV.4 System release challenges

IV.1 Safety Architectures encapsulating ML based subcomponents

The safety impact of AI components in critical systems is plurivalent: while the pertinent problems of ensuring trustworthiness of AI components indicate potential safety risks due to the use of AI, **AI-based defense mechanisms** in security frameworks demonstrate the technology's potential for flexibly and adaptively counteracting diverse kinds of threats. Leveraging the latter potential also in safety architectures by exploiting **AI-based components to establish redundancy** to, e.g., engineered components is deemed attractive. An open and complex question is, given that such AI components would have to be trained (and may even learn on in situ) based on observations obtained from the residual system or prototypes thereof, how independent such an AI component actually is and what level of redundancy it therefore can guarantee. This induces a provoked quest for **statistical means for reliably measuring the relevant statistical independence and the consequential level of redundancy**.

A related question concerns the **mitigation strategies and the selection of fallback functionality** should the redundancy collapse, or more generally the provision of overall **robust architecture blueprints for AI-based systems**, including flexible deployment, update, and replacement of AI-based sub-systems.

IV.2 Adaptable and evolvable architectures for ML based systems

The addition of – often computation-intense – AI to cyber-physical infrastructures with their mostly ECU-based hardware architecture will require new system architectures, including means of allocation of such computation-heavy functionality by deferring it to **cloud or edge services**. AI-as-a-service does, however, induce additional problems for the **safe release of updates**, as the reactions of all possible clients in all possible situations to changes in the service's functionality, be they intended or consequences of attacks, have to be anticipated and controlled. This is particularly true when the operational domain of the service changes, e.g., due to **scenario-based evolution** providing a step-wise increase of system's capabilities either in the service or its callers.

IV.3 Stability of AI based control

A desirable property of machine-learning and the system components employing ML-based functionality is their high level of adaptability, formally expressed by various universal approximation theorems and technically being one of if not the main driver of their application. This flexibility, however, comes at the price of inducing the **potential for oscillatory and instable system behavior at a variety of behavioral scales** and the associated time scales, ranging from **instable feedback dynamics in control applications** due to the highly non-linear transfer functions implemented by, e.g., neural networks to **long-term oscillations and drift in self-adaptive systems**. These behaviors are hard to control, and guarantees for their absence consequently hard to obtain, due to both the complex transfer functions implemented by machine-trained components and the imprecise relation between the mathematical formulations of the optimization goals underlying algorithmic learning and their actual realizations via heuristic optimization methods (like the gradient descent driving backpropagation). Overcoming these problems and thus being able to provide the technically and societally required guarantees on system stability across the aforementioned time scales poses a number of research challenges.

IV.4 System release challenges

Many organizations face challenges in moving ML and Hybrid AI models into production environments. Therefore, deployment must be made flexible in continuous testing and changing algorithms. On average, between 60% and 80% of models created with the intent to deploy are never deployed. Plus, it typically takes six to eight months to deploy a model. However, if models are deployed that were created six to eight months ago, those models may already be obsolete due to to the fast innovation cycles in the field. Also, organizations struggle to integrate ML applications with existing production applications; they waste time and money on projects that are never put into production. MLOps can greatly reduce the risk of such failures and get models into production more quickly where they will ultimately provide the most value to a business. One of the main benefits of MLOps is that it enables ML or Hybrid AI models to deliver high technology readiness value quickly. MLOps does this by ensuring that models can be repeatedly deployed and continuously monitored. The MLOps process allows for: deploying more models faster with automated processes; accelerating time-to-value with rapid delivery of models; optimizing productivity via collaboration and reusing models; reducing risk of wasting time and resources on models that are never put into the final set of results; continuous monitoring and updating of models as data/knowledge changes over time.

IV.5 System Safety challenges

Demonstrating safety of systems with ML components is a key challenge. Assurance cases are one building block to tackle this challenge, as they help to **formalize, structure and relate all assurance (safety, dependability) related artefacts as a comprehensive argumentation**.

An assurance case should accompany a system **during the complete lifecycle: build, release, monitoring and managing during operation, updates and re-releases and even decommissioning**. At development time the assurance case helps to **ensure and demonstrate key system properties** and to provide a strong basis for **human plausibilisation** as well as for **communication between all stakeholders** including certification bodies. During release and deployment (as well as regarding updates and re-releases) it is a basis to **ensure the operational** context is appropriate and important assumptions are not violated. During operation the assurance case supports **continuous monitoring of system properties and the ODD** as well as **runtime assurance approaches**.

Even in the phase of decommissioning the assurance case might provide guidance to prevent physical harm or other unwanted events.

Moreover, **assurance case fragments or patterns** can conserve important knowledge and be a guide for developers. Arguments addressing the usage of ML in a safety-critical system are similar across domains but the **variety of safety arguments is high** and challenging to agree on the **strength of arguments**.

IV.6 System Liability challenges

Product safety and liability are complementary legal frameworks aiming to provide trust and safety to consumers. EU product safety legislation as well as corresponding national initiatives aim at ensuring that only safe products can be placed on the market. EU product liability legislation provides for liability of producers of defective products that cause damage to natural persons or their property. In addition, various national liability regimes may apply if damage occurs[5]. The adequacy and completeness of liability regulations in the face of technological challenges are indeed crucial for society. If the system is inadequate or flawed or has shortcomings in dealing with the damages caused by emerging technologies, victims may end up partially compensated. On the other hand, an overprotective liability regime risks to stifle the development and use of such innovation. Montagnani and Cavallo (2021) identify four main categories where adjustments to existing liability regimes may be needed in order to cope with AI and emerging digital technologies:

1. cases where a (reinterpreted) product liability can still be applied;
2. cases in which strict liability should be extended also to other entities;
3. cases in which there is the need to further develop the notion of duty of care;
4. cases that can be addressed through vicarious liability, by equaling the device to a human auxiliary.

[5] MONTAGNANI, M. L., & Cavallo, M. (2021). Liability and emerging digital technologies: an EU perspective. *Notre Dame Journal of International & Comparative Law*, *11*(2), 208.

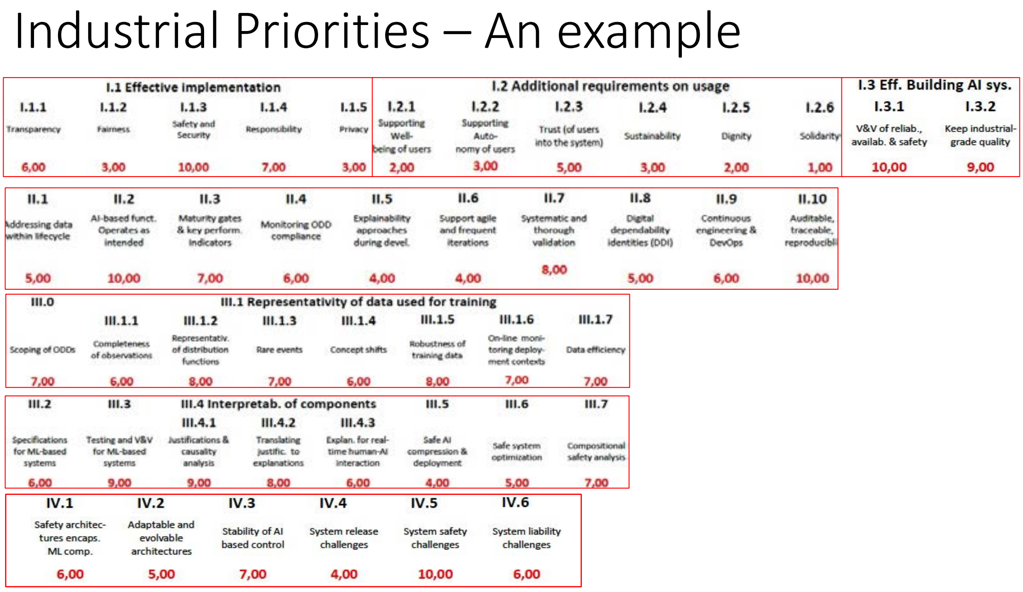
IV. Dependency Analysis and Industrial Priorities

We developed a range of analysis techniques allowing to relate the relevance of academic challenges in addressing the industrial priorities.

1. Industrial Partners are asked to give their companies priorities of the industrial challenges, with a range from 1 (lowest priority) to 10 (highest priority)
2. We analyzed the degree of contribution of industrial challenges in categories II, III, and IV to the overarching challenges in category I.
3. For each of the industrial challenges in one of the categories II, III, and IV we analyzed, which foundational challenges contribute with which strength to attack this industrial challenge. We measured the degree of contribution in a scale 0, 0,25, 0,5, 0,75 to 1, with 1 indicating a very strong contribution.
4. We analyzed causal dependencies between academic challenges: academic challenge AC is causally dependent on academic challenge AC´, if solving AC is necessary to solve AC´.
5. We developed a parametrizable measurement of *importance* of a particular academic challenge for a given industrial partner.

The following snapshots of these stages represent snapshots of the views generated by this analysis for an industrial partner of the industrial core team of this initiative. These steps are currently used in the formation of a project involving the industrial core team and the academic core team

Step 1



Industrial Priorities – An example

I.1 Effective implementation					I.2 Additional requirements on usage						I.3 Eff. Building AI sys.	
I.1.1	I.1.2	I.1.3	I.1.4	I.1.5	I.2.1	I.2.2	I.2.3	I.2.4	I.2.5	I.2.6	I.3.1	I.3.2
Transparency	Fairness	Safety and Security	Responsibility	Privacy	Supporting Well-being of users	Supporting Auto-nomy of users	Trust (of users into the system)	Sustainability	Dignity	Solidarity	V&V of reliab., availab. & safety	Keep industrial-grade quality
6,00	3,00	10,00	7,00	3,00	2,00	3,00	5,00	3,00	2,00	1,00	10,00	9,00

II.1	II.2	II.3	II.4	II.5	II.6	II.7	II.8	II.9	II.10
Addressing data within lifecycle	AI-based funct. Operates as intended	Maturity gates & key perform. Indicators	Monitoring ODD compliance	Explainability approaches during devel.	Support agile and frequent iterations	Systematic and thorough validation	Digital dependability identities (DDI)	Continuous engineering & DevOps	Auditable, traceable, reproducibl
5,00	10,00	7,00	6,00	4,00	4,00	8,00	5,00	6,00	10,00

III.0	III.1 Representativity of data used for training						
	III.1.1	III.1.2	III.1.3	III.1.4	III.1.5	III.1.6	III.1.7
Scoping of ODDs	Completeness of observations	Representativ. of distribution functions	Rare events	Concept shifts	Robustness of training data	On-line moni-toring deploy-ment contexts	Data efficiency
7,00	6,00	8,00	7,00	6,00	8,00	7,00	7,00

III.2	III.3	III.4 Interpretab. of components			III.5	III.6	III.7
		III.4.1	III.4.2	III.4.3			
Specifications for ML-based systems	Testing and V&V for ML-based systems	Justifications & causality analysis	Translating justific. to explanations	Explan. for real-time human-AI interaction	Safe AI compression & deployment	Safe system optimization	Compositional safety analysis
6,00	9,00	9,00	8,00	6,00	4,00	5,00	7,00

IV.1	IV.2	IV.3	IV.4	IV.5	IV.6
Safety architec-tures encaps. ML comp.	Adaptable and evolvable architectures	Stability of AI based control	System release challenges	System safety challenges	System liability challenges
6,00	5,00	7,00	4,00	10,00	6,00

Step 2

		I.1.3	**I.3.1**
		Safety and Security	V&V of reliab., availab. & safety
		10,00	10,00
II.	**Process related challenges**		
II.1	Addressing data within the lifecycle	1,00	1,00
II.2	Ensuring that AI-based functions operate as intended	1,00	1,00
II.3	Maturity gates & meaningful key performance indicators	1,00	1,00
II.4	Monitoring ODD compliance	1,00	1,00
II.5	Using explainability approaches during development	1,00	1,00
II.6	Support agile and frequent iterations in AI development	0,50	1,00
II.7	Systematic and thorough validation approach	1,00	1,00
II.8	Digital dependability identities (DDI)	1,00	1,00
II.9	Continuous engineering and DevOps type lifecycles	0,50	1,00
II.10	Auditable, traceable, reproducible, and measurable lifecycle processes	1,00	1,00
III.	**Challenges related to design principles, methods, and tools**		
III.0	Scoping of ODDs	1,00	1,00
III.1	*How to ensure representativity of data used for training*		
III.1.1	Ensuring completeness of observations	1,00	1,00
III.1.2	Ensuring representativity of distribution functions for relevant data	1,00	1,00
III.1.3	Coping with rare events	1,00	1,00
III.1.4	Coping with concept shifts	1,00	1,00
III.1.5	Ensuring robustness of training data	1,00	1,00
III.1.6	On-line monitoring deployment contexts	1,00	1,00
III.1.7	Data efficiency	0,25	0,00
III.2	How to derive specifications for ML based systems	1,00	1,00
III.3	How to test, validate, verify ML based systems	1,00	1,00
III.4	*How to ensure interpretability of ML based components*		
III.4.1	Justifications allowing causality analysis	1,00	1,00
III.4.2	Translating justifications to explatations interpretable by humans	1,00	1,00
III.4.3	Explanations for real-time human-AI interactions	1,00	1,00
III.5	Safe techniques for AI compression and deployment	0,00	0,25
III.6	Safe techniques for system optimisation	0,00	0,25
III.7	Compositional safety analysis of systems including AI components	1,00	1,00
IV	**System challenges**		
IV.1	Safety architectures encapsulating ML based subcomponents	1,00	0,25
IV.2	Adaptable and evolvable architectures for ML based systems	0,25	0,25
IV.3	Stability of AI based control	1,00	1,00
IV.4	System release challenges	1,00	0,00
IV.5	System safety challenges	1,00	1,00
IV.6	System liability challenges	1,00	0,25

Step 3 and 4

		IV.5	
		System safety challenges	
	Weight by overarching high level challenges	3,96	
	Industrial weight of challenge	10,00	
A	**Identification of relevant ODDs and system boundaries**		
A1	Relevance	1,00	9,40
A2	Relative artefact completeness	1,00	9,40
A3	Achieving artefact completeness	1,00	9,40
A4	Achieving robustness	1,00	9,40
A5	Validating ODD characterizations	1,00	9,40
A6	Monitoring ODD compliance	1,00	9,40
B	**Safety by design - design principles for guaranteeing properties**		
B1	Robustness of decisions	1,00	9,40
B2	Reliability of decisions	1,00	9,40
B3	Active detection of scope of validity	1,00	9,40
B4	Certified Integration of Prior Knowledge	1,00	9,40
B5	Cyclic learning and active self-learning	1,00	9,40
B6	Demand-driven design based on safety architect. & safety cases	1,00	9,40
C	**Validation and verification for guaranteeing properties**		
C1	Hardening the role of the specification in AI	1,00	9,40
C2	Robustness of ML classifiers	1,00	9,40
C3	Robustness of ML in the control loop	1,00	9,40
C4	Physics-aware and neuro-symbolic AI	1,00	9,40
C5	Dealing with uncertainty and rare events	1,00	9,40
C6	Compositional validation and verification	1,00	9,40
C7	Laboratory conditions for verifiable ML	1,00	9,40
D	**Explainability and comprehension**		
D1	Mimicking human explanations	1,00	9,40
D2	Generating sufficient evidences for explanations	1,00	9,40
D3	Integrating explainability into AI-based components	1,00	9,40
D4	Composability of explanations	1,00	9,40
D5	From explanations to comprehension	1,00	9,40
E	**Safe human-AI interaction**		

E1	Human modelling for safe human AI interaction	0,50	4,70
E2	Quality assurance for human models	1,00	9,40
E3	Implementability and adaptability of human models	1,00	9,40
E4	Guaranteeing safety for human-AI interaction	1,00	9,40
E5	Challenging the AI system	1,00	9,40
E6	Handover of control	1,00	9,40
E7	Safety and stability of human-AI based systems control loop	1,00	9,40
E8	Accountability and traceability	1,00	9,40
F	**Guaranteeing safe cooperative behaviour**		
F1	Shared situational awareness	1,00	9,40
F2	Shared mutual introspection	1,00	9,40
F3	Achieving safe cooperation	1,00	9,40
F4	Achieving safe abortion	1,00	9,40
F5	Negotiating cooperation	0,50	4,70
G	**Avoiding system oscillation and instability**		
G1	Stability verification of ML-in-the-loop control systems	1,00	9,40
G2	Optimisation-theoretic characterization of ML-based self-adaptation	1,00	9,40
G3	Proving stability properties of ML-based self-adaptive systems	1,00	9,40
G4	Ensuring stability of interacting systems of indiv. learning systems	1,00	9,40
H	**Construction of safety cases**		
H1	Mastering complexity of the safety argument	1,00	9,40
H2	Quantifying and minimizing out-of-distribution risk	1,00	9,40
H3	Environmental factors	1,00	9,40
H4	Recertification	1,00	9,40
H5	Ethical dimension	0,50	4,70
H6	Privacy vs. Accountability	0,50	4,70

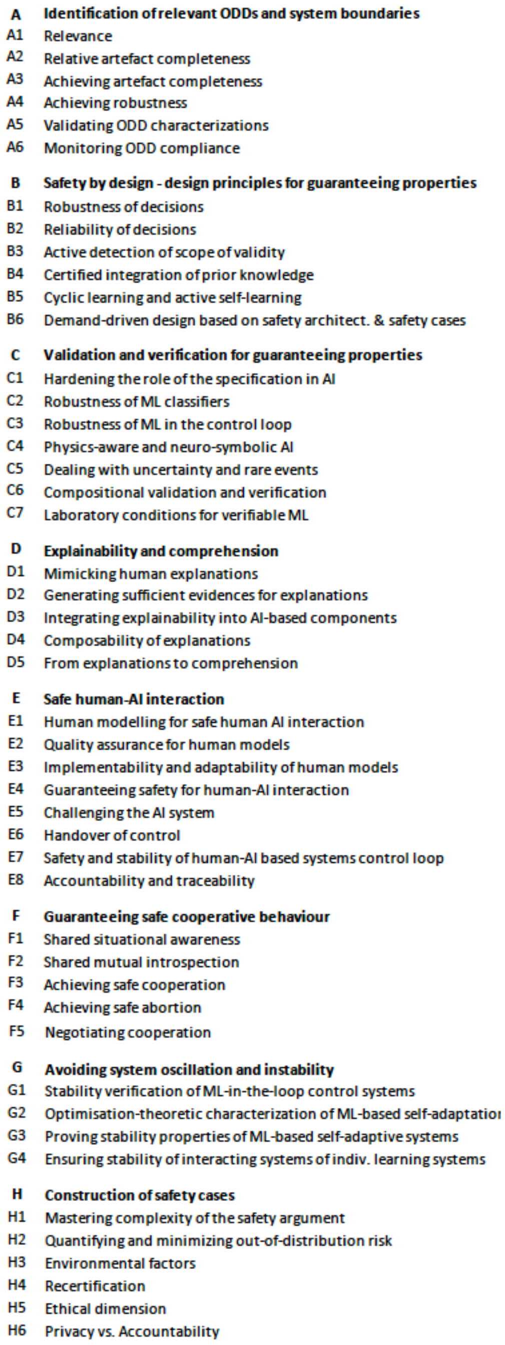
Step 5

mportance of challenge

1,55
1,99
2,48
2,18
2,33
2,17
2,54
2,82
2,09
1,40
1,14

I

2,09
2,08
2,31
1,93
2,41
2,55
1,55
1,79
2,39
1,83
2,09
2,00
2,78
1,84
1,51
1,84
2,47
1,69
2,69
1,37
1,69
1,83
1,97



A	**Identification of relevant ODDs and system boundaries**
A1	Relevance
A2	Relative artefact completeness
A3	Achieving artefact completeness
A4	Achieving robustness
A5	Validating ODD characterizations
A6	Monitoring ODD compliance
B	**Safety by design - design principles for guaranteeing properties**
B1	Robustness of decisions
B2	Reliability of decisions
B3	Active detection of scope of validity
B4	Certified integration of prior knowledge
B5	Cyclic learning and active self-learning
B6	Demand-driven design based on safety architect. & safety cases
C	**Validation and verification for guaranteeing properties**
C1	Hardening the role of the specification in AI
C2	Robustness of ML classifiers
C3	Robustness of ML in the control loop
C4	Physics-aware and neuro-symbolic AI
C5	Dealing with uncertainty and rare events
C6	Compositional validation and verification
C7	Laboratory conditions for verifiable ML
D	**Explainability and comprehension**
D1	Mimicking human explanations
D2	Generating sufficient evidences for explanations
D3	Integrating explainability into AI-based components
D4	Composability of explanations
D5	From explanations to comprehension
E	**Safe human-AI interaction**
E1	Human modelling for safe human AI interaction
E2	Quality assurance for human models
E3	Implementability and adaptability of human models
E4	Guaranteeing safety for human-AI interaction
E5	Challenging the AI system
E6	Handover of control
E7	Safety and stability of human-AI based systems control loop
E8	Accountability and traceability
F	**Guaranteeing safe cooperative behaviour**
F1	Shared situational awareness
F2	Shared mutual introspection
F3	Achieving safe cooperation
F4	Achieving safe abortion
F5	Negotiating cooperation
G	**Avoiding system oscillation and instability**
G1	Stability verification of ML-in-the-loop control systems
G2	Optimisation-theoretic characterization of ML-based self-adaptatio
G3	Proving stability properties of ML-based self-adaptive systems
G4	Ensuring stability of interacting systems of indiv. learning systems
H	**Construction of safety cases**
H1	Mastering complexity of the safety argument
H2	Quantifying and minimizing out-of-distribution risk
H3	Environmental factors
H4	Recertification
H5	Ethical dimension
H6	Privacy vs. Accountability