ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Fraunhofer
IKS

Closing the gaps: Complexity and uncertainty in the safety assurance and regulation of automated driving

S. Burton & J. A. McDermid

January 2023

Closing the gaps: Complexity and uncertainty in the safety assurance and regulation of automated driving

S. Burton,
J. A. McDermid OBE FREng

January 11, 2023

Executive summary

The increasing level of automation within an open context and use of artificial intelligence in cognitive cyber-physical Systems (CPS) is leading to emergent complexity and subsequently to uncertainties within the system assurance process. For example, in automated driving this is particularly true for the class of risks associated with the safety of the intended functionality (SOTIF) as described by the standard ISO 21448. In this report, we provide a definition of how complexity and uncertainty impacts the safety assurance of cognitive CPS. Based on this structured understanding of the problem, we propose an approach to managing the safety and regulating the deployment and operation of such systems in order to maintain an acceptable level of residual risk despite of, and with the intent of reducing, residual uncertainties. The approach includes criteria to guide decisions regarding the deployment and continuous assurance of the systems. The model used to structure these proposals includes a causal analysis of the factors impacting the complexity and resulting uncertainty (and, by extension, risk) that span the three layers of technical & human factors, management & operations and governance and regulation. These principles are generally applicable to a broad class of cognitive cyber-physical systems. However, in this report we focus on their application to automated driving systems.

Contents

1 Introduction

Cyber-physical systems (CPS) comprise physical components interacting with other technical systems, human operators and users. We define *cognitive* cyber-physical systems as systems that achieve higher levels of automation by exploiting capabilities such as perception, reasoning, learning both pre- and post-deployment, and adaptation. Recent advances in machine learning (ML) and increasing system inter-connectivity are enabling cognitive CPS to achieve higher levels of automation than previously possible. These capabilities allow not only for an ever-increasing transfer of decision-making responsibilities from human operators to the technical system, but also allow the system to operate in an increasingly open and dynamic environment.

Examples of such advances in cognitive CPS can be found in medical devices, intelligent traffic management systems, automated vehicles (passenger cars, delivery drones, trains), and robotics. These systems are intended to increase safety and the well-being of society and in some cases are also intended to have beneficial environmental effects. Indeed, we see a need to ensure that such systems are demonstrably safe, ethical and sustainable [1]. The risks involved in malfunctions of these systems can be severe in their consequences both in terms of direct harm caused by the systems as well as in undermining public trust in the technology, thus slowing the deployment of technologies with a huge potential for society.

We recognise that for such complex socio-technical systems, tolerable levels of residual risk cannot be achieved through technical measures alone. Human factors, safety management processes, operating procedures as well as regulatory constraints all contribute to achieving a socially acceptable and continuing level of safety. An analysis of fatal accidents involving automated driving systems, both during development as well as in series production confirms the emergent risks involved in developing and operating safe automated driving systems [2]. Specifically, the interaction between the technical & human factors, management & operations and governance & regulation perspectives directly contribute to or can significantly reduce residual risk [2]. These perspectives, or layers, are defined as follows [3]:

- **Technical & Human Factors** – This layer covers the technical design and safety analysis process that allows systems to be deployed at an acceptable level of risk, then actively monitored to ensure deviations (or "gaps") between what was predicted and what is actually happening can be identified and rectified. This layer includes not only the technological components but also the tasks performed by the users, operators and other stakeholders within a socio-technical context.
- **Management & Operations** – This layer coordinates tasks involved in the design, operation and maintenance of the systems, across the supply chain, enabling risk management and informed design trade-offs across corporate boundaries, control over intellectual property and liability, management of supply chain dynamics and long-term institutional knowledge for long-lived and evolving systems.
- **Regulation & Governance** – This layer consists of incentives and requirements for organisations to adhere to best practice through direct regulation, so-called soft law approaches or a consensus in the form of national and international standards. In formulating these standards and regulations, governments and authorities represent the societal expectations on the acceptable level of residual risk that is to be associated with the systems.

Within this report we refer to increasing levels of automation rather than using the term autonomy which could, depending on interpretation, imply that the system is capable of independently adapting its goals. We believe that such systems, even if feasible with current technology, should not be considered for safety-related tasks in open and dynamic contexts[1] due to the additional layer of uncertainty in assuring the appropriateness of evolving system goals. Further, we only consider the use of ML for specific tasks within a well-defined context (also referred to as "Narrow AI"),

[1] We recognise that the term "fully autonomous" can be used to describe systems that already exist in some domains, e.g. rail and factories, but here the environments are relatively controlled or constrained. In reality there is a spectrum, and the level of complexity and openness that can safely be managed will change as system design and assurance capabilities advance.

rather than end-to-end applications of ML or even artificial general intelligence (AGI) which again would lead to a raft of open issues which require significantly more research and public debate. These restrictions allow us to formulate a proposal for an iterative approach to the introduction of cognitive CPS for increasing levels of automation in safety-relevant contexts. Throughout the report we will refer to examples from the automated driving domain to illustrate the concepts and provide specific recommendations for assurance and regulatory approaches, but it is intended that the ideas presented here could be adapted and applied in other complex, open and dynamic domains.

We elaborate on the challenges involved in regulating the safety of such systems and in particular focus on the following questions: Which criteria can be used to justify the safe deployment of such systems? How can the continuing safety of such systems be assured? Which indicators can be used, from the perspective of regulators, to decide when corrective action must be taken to maintain safety? We address these questions by combining the following perspectives. First, we examine various manifestations of uncertainty caused by the increasing complexity of systems and the environments in which they operate. Second, we provide recommendations for how this uncertainty can be addressed within the system at the three layers introduced above: technical & human factors, management & operations and regulation & governance.

The rest of this report is structured as follows: in the following section we summarise related work and provide a set of relevant definitions. In section 3, based on these definitions we formulate the problem of safety assurance under uncertainty. In section 4 we propose an initial set of conditions for the safe deployment of cognitive CPS for safety-relevant, highly-automated functions. In section 5 we describe a continuous approach to ensuring the ongoing safety of the systems despite residual uncertainties and an evolving operational domain. We close the report with a discussion on open research questions and conclusions.

2 Definitions and state of the art

In this section we define concepts that can be used to characterise the challenges related to the safety assurance of complex, highly automated systems.

2.1 System complexity, the semantic gap and uncertainty

In [2] we described how emergent complexity within highly automated driving systems leads to specific safety assurance challenges, requiring a holistic approach across the technical & human factors, management & operations and regulation & governance layers[2] to reach a suitable level of residual risk. The generic framework discussed within the paper was originally defined by the same authors as part of a study [3] intended to provide conceptual clarity around the factors that lead to systemic (as opposed to mechanistic) failures in complex systems which have a safety impact. We define a complex system as follows:

- **Complex system:** A system that exhibits behaviours that are *emergent* properties of the interactions between the parts of the system, where the behaviours would not be predicted based on knowledge of the parts and their interactions alone.

From the perspective of complexity science [4], complex systems share a number of characteristics including: semi-permeable system boundaries, non-linear behaviour, mode transitions & tipping points, and self-organisation. Such properties undermine typical approaches to safety assurance which are based on models of well-defined (i.e. known) system behaviour and causes of failures due to individual component faults. The lack of knowledge about the causes of emergent behaviour that is at the core of the definition of complexity used above is strongly related to the concept of uncertainty as illustrated in the following definition.

- **Uncertainty:** Any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system [5].

[2] Hereafter we refer to these as the TMG layers for brevity.

Uncertainty is often categorised as either aleatoric or epistemic uncertainties, both of which are relevant for the concepts described in this report. Aleatoric or stochastic uncertainty refers to the inherent randomness of properties of the physical world that are observed by the system. Epistemic or systematic uncertainty refers to inadequacies of the model(s) used to interpret the physical world. Uncertainty can impact the safety of the system in various ways:

- **Specification uncertainty** includes uncertainty in the definition and completeness of appropriate safety acceptance criteria and the definition of acceptably safe behaviour in all situations that can reasonably be anticipated to arise within the target operational design domain. This also includes the ability to postulate a sufficiently complete model of the operational design domain, which can be used to reason about the completeness of training and test data. Specification uncertainty can also be rooted in competing objectives and stakeholder-specific definitions of tolerable residual risk at the management & operations and regulation & governance layers. Furthermore, specification uncertainty can lead to unresolved questions related to ethical/socially acceptable behaviour of the system. The inability to provide a complete specification of the (safe) behaviour of the system is inherently linked to both the semantic gap [6] and emergent properties of complex systems.
- **Technical uncertainty** relates to a lack of predictability in the performance of the technical components of a system. An example of technical uncertainty is the unpredictable reaction of the system to previously unseen events, or differences in the system behaviour despite similar input conditions. This can include a mismatch between the system's model of its environment or own internal state and the ground truth, limitations in sensing capabilities including degradation under adverse weather conditions, lack of robustness in ML-based functions, variations in actuator performance and the impact of security vulnerabilities. Technical uncertainty can manifest itself within different components within a “sense, understand, decide, act” functional effect chain. Cumulative effects of uncertainty can propagate throughout the system and may lead to seemingly non-deterministic behaviour or sudden mode changes (tipping points). Technical uncertainty can also include imprecision in the modelling or measurement of the performance of the technical system itself. This, in turn, can lead to a lack of confidence in assurance arguments for the system (assurance uncertainty).
- **Assurance uncertainty:** Assurance uncertainty is related to a lack of knowledge regarding the completeness and/or validity of an assurance argument for critical (safety) properties of the system. This can include a lack of confidence in the validity (including statistical confidence) of evidence supporting the assurance argument as well as the chain of reasoning itself. Assurance uncertainty can also include a lack of confidence in the validity and appropriateness of the overall claim of the assurance argument (e.g. due to specification uncertainty) as well as the continued validity of the argument over time, as the system, its environment and societal expectations on tolerable residual risk evolve.

In addition, uncertainty can be defined in terms of the following quantitative and qualitative properties:

- **Statistical uncertainty:** can be expressed in statistical terms, such as with probability distributions or using belief theory (Quantitative).
- **Scenario uncertainty:** can only be described using scenarios, which are plausible states of the system and/or its environment in which knowledge is lacking without any statistical support (Qualitative).
- **Ontological uncertainty:** is defined as complete ignorance of a model about relevant aspects of the system [7]. This means the system, including its assurance activities is not aware that its knowledge is subject to uncertainty leading to so-called *unknown unknowns*. Resolving this level of uncertainty therefore requires measures external to the system.

In [6], the authors discussed the notion of the *semantic gap* and its impact on safety assurance related to systems with increasing automation and the use of ML. The semantic gap is defined as follows:

- **Semantic gap:** The gap between intended and specified functionality — when implicit and ambiguous intentions on the system are more diverse than the system's explicit and concrete specification [8].

The semantic gap is a direct consequence of system complexity and can be caused by:

- the *complexity and unpredictability of the environment* in which the system operates,
- the *complexity and unpredictability of the system itself* as well as the systems interactions with other technical systems and human actors (including operators, users, and bystanders), and
- the increasing *transfer of the decision-making responsibility* from a human actor to the system, as the system will not have the semantic and contextual understanding of the decision-making that the human does.

The semantic gap can lead to uncertainty within the specification of the system and subsequently also to *liability*, *moral responsibility* and *assurance gaps*.

System complexity, uncertainty and the semantic gap

Complexity of the system, its environment and the task to be achieved leads to emergent behaviour that cannot be predicted during system development. In order to achieve a tolerable level of residual risk, despite these underlying uncertainties, their sources and our ability to measure and bound them must be acknowledged. Limitations in our abilities to adequately consider these factors lead to uncertainties in the assurance process. This in turn leads to semantic gaps between our expectations on the system and our ability to specify, and therefore ultimately ensure and assure, safe behaviour in sufficient detail necessary to achieve a tolerable level of residual risk.

2.2 Robustness, resilience and systemic failures

Traditional approaches to achieving the functional safety of electrical & electronic systems, as defined by standards such as ISO 26262 Road Vehicles - functional safety [9] have addressed the issue of risk associated with component faults in the system. Typically, a model of the system and an analysis of the impact of component faults are used to assess the potential for safety-related system failures such that potentially critical malfunctions can be eliminated or mitigated through other system measures. This approach to safety can be seen as improving system *robustness*, i.e. the ability of a system to cope with foreseen events. However, despite improvements in the modelling of faults and the analysis of fault propagation within the system this approach has its limits [10].

The automotive safety standard ISO 21448 [11] addresses the *safety of the intended functionality* (SOTIF), defined as the absence of unreasonable risk due to hazards resulting from functional insufficiencies of the intended functionality or its implementation. This model of safety addresses a wider range of causes of safety hazards than individual component failures and includes performance limitations of sensors and ML components and foreseeable misuse by a human operator. In particular, the standard explicitly addresses hazards caused by specification and technical uncertainties as described above. ISO 21448 uses the term triggering conditions to describe scenarios or input conditions that reveal insufficiencies in the system. It distinguishes between *known triggering conditions* which relate to well understood limitations of the system which can be mitigated against and *unknown triggering conditions* which characterise as yet unknown limitations of the system that could lead to hazards (see the definition of ontological uncertainty above). The objective of the safety assurance methodology outlined in the standard is to iteratively discover known triggering events and to handle these in the system design or by restricting the operating conditions of the system, whilst providing an argument for the acceptably low likelihood of there being residual unknown triggering conditions (which are encountered in operation).

However, the standard does not provide guidance in how to set quantitative risk acceptance criteria for determining whether a sufficiently small number of unknown triggering conditions remain within in the deployment environment and for complex cognitive CPSs residual unknown triggering conditions will inevitably remain. In particular, the remaining portion of triggering conditions that consist of complex interactions of environmental and system states are difficult if not impossible to predict (as opposed to more obvious triggering conditions such as lighting impacting camera sensors). Building systems that are robust against a known model of system behaviour (and known triggering conditions), including bounded statistical uncertainties may therefore not be sufficient. In order to address the issues of emergent complexity and semantic gaps, we need to engineer systems that are *resilient*, i.e. systems with the ability to absorb the unforeseeable or, at least, unforeseen (due to the ontological uncertainty as defined above).

Despite providing a significant extension of the traditional model of functional safety, the SOTIF approach does not address the *resilience* of the system to unknown triggering conditions and is limited in its perspective to technical performance limitations of the system. We therefore propose the following extension to the model of failures considered by the safety standards by defining the concept of systemic failures as follows:

- **Systemic Failure:** Failure at a system level caused by interactions between behaviours of the system's components and interactions with, or dependencies on, its environment.

Furthermore, we do not limit the causes or effects of systemic failures to technical issues but consider failures across all three of the TMG layers. As an example at the Technical & Human Factors layer, either the technical system or the human driver take inappropriate or conflicting decisions to the automated driving system due to them having an incomplete model of the environment or the system state, or their respective models are inconsistent. At the level of regulation & governance, systemic failures could lead to inadequate deployment decisions or inadequate regulatory control resulting in an increased risk to society due to new technologies.

Robustness, resilience and systemic failures

Current iterations of safety standards for road vehicles focus on reducing risk by increasing robustness of the system against component faults and known insufficiencies of the system to remain safe in the presence of known triggering conditions. ISO 21448 addresses the issue of unknown triggering conditions through the perspective of arguing that the probability of their occurrence or their causing hazardous behaviour is sufficiently low. This argument itself will be subject to significant assurance uncertainty, which increases with the system complexity. However, in order to remain safe despite the potential for systemic failures, systems must be engineered to be resilient against unknown triggering conditions and emergent behaviour.

2.3 System-theoretic safety analyses

As described above, it appears that mechanistic approaches to safety analysis will be insufficient to predict hazards caused by systemic failures of the system due to emergent complexity. Traditional approaches to safety analysis, both inductive such as Failure Modes and Effects Analysis (FMEA) [12] and deductive such as Fault Tree Analysis (FTA) [13] which are recommended by existing safety standards such as ISO 26262 [9] are no longer sufficient. Leveson's System Theoretic Accident Methods and Processes (STAMP) and Systems Theoretic Process Analysis (STPA) [14] approach builds upon sociotechnical systems theory. In this approach, the system is seen as having a number of hierarchical levels each with its own control structure with controls and constraints operating vertically between the levels. Hazardous incidents are therefore seen as control failures. The hierarchical system-theoretic approach appears well suited to model the interactions between the layers in the TMG model, e.g. inadequate control actions at the governance level leading to inadequate safety management practices.

STPA has gathered increasing attention in recent years in the automotive industry and has been applied to the safety analysis of automated driving systems [15, 16]. An example of the application of the technique to a highway pilot application can be found within ISO 21448. However, to the best of our knowledge these analyses have so far focused on the technical perspective and its interaction with human operators of the system rather than the interactions between the TMG layers. Monkhouse et al [17] recently proposed an enhanced vehicle control model that models the shared cognitive load between the driver and the automated driving system. This could provide a valuable basis for evaluating potential emergent behaviour between the technical system and human operator (for example in critical handover conditions where the Automated Driving System (ADS) reaches its technical limitations).

Nevertheless, the ability to understand or predict systemic failures is inherently restricted by the model chosen to represent the system and its interactions with the environment. Rasmussen [18] introduced a risk management framework based on the analysis of variations in behaviour (rather than fault models) across six levels of the sociotechnical systems: government, regulators/associations, company, management, staff, work. This approach inspired the layered framework described in [3], where we extended the layer of work with technical considerations of complex systems and provided a causal structure for analysing how factors contributing to system complexity can lead to systemic failures. The Functional Resonance Analysis Method (FRAM) [19] introduces a notion of functional resonance to complement causal models or theories. Functional resonance can be seen as (not necessarily intended) interactions or dependencies between functions. The approach is based upon the premise that normal variability in performance of tasks can lead to unexpected and undesired consequences. The analysis is based on the dynamic nature of interactions within the system in combination with performance variations of tasks that are coupled via input/output, control, constraint, resource and temporal relationships. This type of analysis may inspire approaches that integrate the notions of uncertainty described above to detect and analyse dependencies between components of the system, including across the TMG layers.

System-theoretic safety analysis

Approaches to safety analysis currently proposed by safety standards (e.g. FMEA, FTA, STPA) do not appear sufficient to capture and analyse the subtle interactions that occur within the wider context of automated driving systems. In particular, the impact of specification, technical and assurance uncertainty as well as vertical causalities across the TMG layers are not currently modelled. However, some of these approaches including STPA and FRAM appear to have the potential for being extended to cover these aspects (see discussion of future work in Section 6)

2.4 Self-adaptive systems

The introduction of self-management mechanisms into software-based systems capable of self-adaptation (also known as dynamic safety management) [20, 21, 22] has been proposed as a means to increase resilience. A common goal of research in this area is to maximise utility of the system whilst maintaining an adequate level of safety by enabling the system to adapt to changes in the operation environment or capabilities of the system itself based on a model of risk at runtime. These approaches typically make use of an abstract system architecture as described in Figure 1[3]. The main control task is performed by the *managed system*. A *managing system* has the task of monitoring properties of the managed system and the environment which can be used to determine a notion of current risk and based on some decision model enact adaptations of the system necessary to ensure that the risk remains below a tolerable level. More recent work has begun to acknowledge the impact of various form of complexity in the design of self-adaptive, autonomous systems [23].

These approaches typically address technical uncertainty and not issues related to specification or assurance uncertainty and result in a number of challenges related to the assurance of the dynamically adaptive behaviour [24, 25]. These challenges include:

- Perpetual assurance: continuous generation of evidence that system requirements are met, despite adaptation of the system and its environment,

[3] Note that this shows interactions at the Task & Technical layer in the TMG model, although it has implications for the Management and Governance layers.

- Compositional assurance: avoiding re-validation of the entire system (or emergent systems-of-systems) for each change and
- Feedback and monitoring: defining effective observation points for determining when the assurance process has not been effective[4].

In Section 5 we extend this model of self-adaptive systems to examine approaches to continuous assurance and regulation.

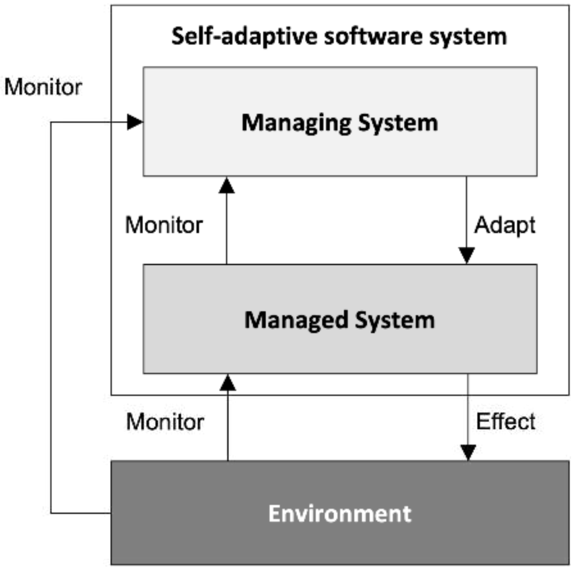

Figure 1: Model of a self-adaptive system

Self-adaptive systems

Self-adaptive systems have the potential for increasing system resilience against changes in operating context, system capabilities or unknown triggering events. The key challenges to applying this architecture paradigm to complex highly automated driving systems is related to the difficulty in determining a model for the *managing system* that is capable of determining potentially hazardous conditions which the managed system itself is unaware of, and is capable of determining appropriate adaptation actions. The definition of such a model is exacerbated due to the issue of semantic gaps [6]. Furthermore, the application of the self-adaptive systems paradigm requires continuous assurance activities in order to avoid increasing assurance uncertainties over time and, ideally, to reduce those uncertainties giving a progressive increase in safety and assurance of safety.

3 Safety assurance under uncertainty

The definitions presented in Section 2 illustrate the relationships between emergent system complexity, uncertainty and the resulting risk of systemic failures. The increase in automation within an open context, coupled with the introduction of technologies such as ML can amplify the underlying causes of this complexity and thus the resulting uncertainties. We therefore assert that, without addressing these issues, a tolerable level of residual risk for systems such as fully automated driving will not be achievable. The challenge of ensuring the safety of complex, highly automated cognitive CPSs can thus be formulated as follows:

[4] In the sense that unsafe behaviours have arisen in operation; it seems unlikely that monitoring would detect cases where beneficial, safe behaviour had been excluded from a deployed system because it could not be adequately assured.

Ensuring the safety of complex, highly automated cognitive CPSs

Continuously ensure an acceptable level of residual risk, *despite* the emergent *complexity* of the system, *changes* to both the deployment context and the system itself over time and the resulting *specification, technical and assurance uncertainties*.

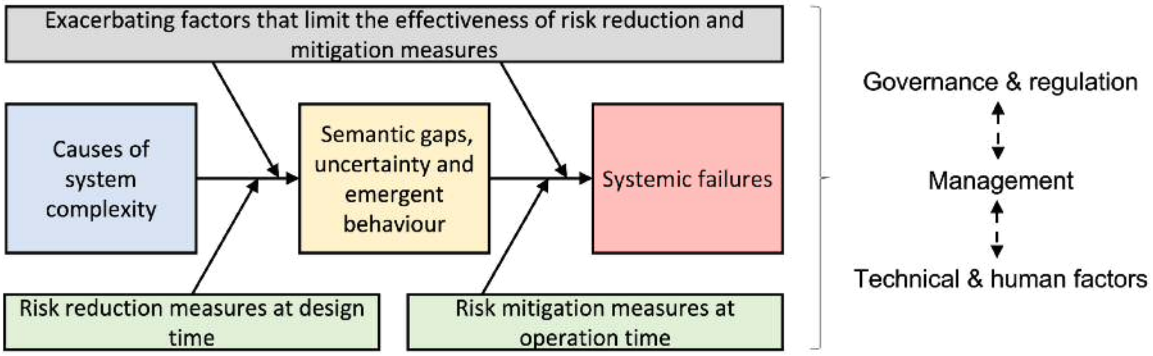
Exacerbating factors that limit the effectiveness of risk reduction and mitigation measures
Causes of system complexity
Semantic gaps, uncertainty and emergent behaviour
Systemic failures
Risk reduction measures at design time
Risk mitigation measures at operation time
Governance & regulation
Management
Technical & human factors

Figure 2: Safer complex systems framework

Meeting this challenge requires measures to not only ensure the *robustness* of the system, but also to ensure the *resilience* to unforeseen and unforeseeable causes of hazards. These measures must be coordinated across the three TMG layers both during design time and throughout operation, with the objective of continuously reducing the residual level of uncertainty and hence reducing the risk associated with systemic failures.

Measures at design-time can reduce the risk of systemic failures by eliminating or limiting the impact of causes of complexity and thus the resulting uncertainty within the system. Measures during operation can mitigate residual risk within the system by reducing the impact of uncertainty and thus the the probability of systemic failures. Figure 2 summarises the causal relationships between the different factors described above.

These observations lead to the following principles for assuring the safety of highly automated, complex cognitive CPSs:

1. A systematic analysis of the potential causes of complexity and resulting uncertainty across all TMG layers shall be performed in order to estimate the risk associated with the system and to identify appropriate mitigation measures during design and operation.
2. It shall be argued that effective safety measures were applied at design-time across all TMG layers to address the risk associated with the complexity of the system. These measures shall address both robustness and resilience of the system.
   - The system shall be robust against known sources of technical uncertainty within the system. This shall include both statistical uncertainty and scenario uncertainty.
   - The system shall be designed to be resilient to both previously unknown triggering conditions as well as the effects of emergent behaviour arising from interactions with other systems and bystanders.
3. It shall be argued that effective safety measures are applied during operation across all TMG layer to address the residual risk associated with the complexity of the system. These measures shall include:
   - Measures during operation to uncover previously unknown triggering events as well as to better estimate the probability of residual unknown triggering conditions. This will include refining the model of the operational design domain as well as an understanding of the (emergent) behaviour of the system.

- Procedures to be able to continuously adapt the system, its assurance methods and operating procedures should previously unknown sources of complexity, uncertainty and risk be discovered during deployment.

In the following sections we will first define a set of criteria by which the safe deployment of such systems can be argued, thus focusing primarily on *design-time measures*. We then focus on *continuous* approaches to safety assurance and regulation with the focus on observations that are required at the three layers of the TMG to ensure safe operation of the system within a dynamic environment (*operation-time measures*).

4 Conditions for safe deployment

Government policy makers and regulators should consider the growth in complexity and the trends in the scope and capability of systems when determining the conditions under which new classes of technologies and systems are deployed. In doing so, the regulatory structures and safety management systems themselves should be considered from the perspective of emergent complexity and resulting uncertainty which could lead to systemic failures in the governance and management of the systems. This section focuses on regulatory decisions to allow the deployment of safety-relevant highly automated cognitive CPSs that exhibit the properties of complexity and uncertainty outlined earlier.

4.1 Understanding the systematic task complexity

In order to manage the emergent risk related to cognitive CPS, it is imperative that manufacturers, operators and regulators consider the inherent **systematic task complexity** of the functions to be fulfilled by the system and the context in which the system operates. Systematic task complexity can be defined as the difficulty of achieving completeness in modelling all of the necessary aspects and parameters for the target function and operating scenarios of the system. It can also be interpreted as a measure of the expected level of residual specification, technical and assurance uncertainty in the system. There are therefore several dimensions in which the systematic task complexity should be considered in the deployment decision, including:

- The confidence in which an appropriate specification of safe behaviour and tolerable residual risk for the system can be defined.
- The level of inherent technical uncertainty within the system.
- The ability to understand the system behaviour, despite specification and technical uncertainty, especially in critical situations or during incident analysis.
- The appropriateness of available methods and technologies for assessing the capabilities of the system including an evaluation of residual uncertainties.
- The ability to monitor the behaviour of the system during operation, including the ability to predict increasing levels of risk due to emergent behaviour through the observation of lead indicators.

The definition of systematic task complexity implies that some automation tasks are inherently simpler to assure to a high level of confidence than others. For example a traffic sign recognition and speed limit warning function leads to lower levels of specification, technical and assurance uncertainty than pedestrian recognition for urban automated driving.

Based on an analysis of the systematic task complexity and expected manifestations of uncertainty, appropriate measures should be defined in the systems engineering process including continuous assurance to avoid the presence of systemic failures (see Figure 3). Within such a process, the analysis of systematic task complexity would be used to inform and refine system-theoretic safety analyses in order to achieve greater coverage of potential root causes of systemic system failures.

Regulators could require an evaluation of the impact of systematic task complexity on the safety of the system as part of the approval process. This evaluation could be performed according to a fixed set of criteria and should also include an analysis of the applicability of existing standards (see below). The establishment and acceptance of a set of definitive criteria for assessing the systematic task complexity is a topic of future research, however it is expected that such measures will inevitably be relative rather than absolute in nature, allowing for a comparison with well established and understood classes of systems. Furthermore, a standardisation of the definition of common use cases and factors impacting their systematic task complexity would support a consistent evaluation across manufacturers and systems.

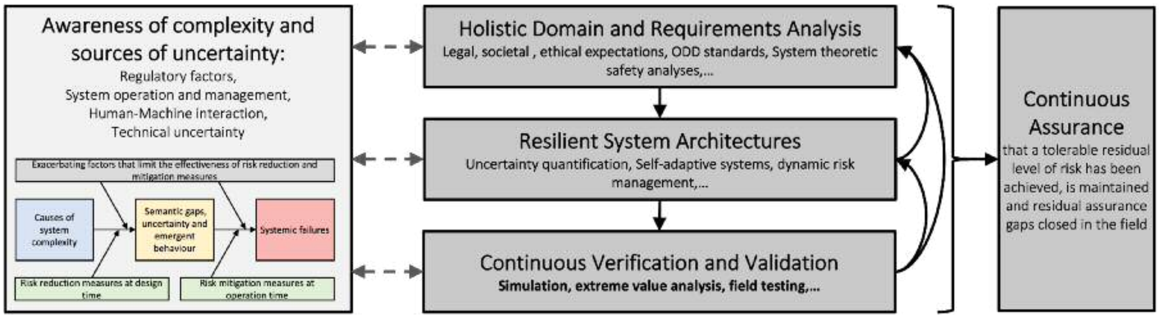
Awareness of complexity and sources of uncertainty:
Regulatory factors,
System operation and management,
Human-Machine interaction,
Technical uncertainty
Exacerbating factors that limit the effectiveness of risk reduction and mitigation measures
Causes of system complexity
Semantic gaps, uncertainty and emergent behaviour
Systemic failures
Risk reduction measures at design time
Risk mitigation measures at operation time
Holistic Domain and Requirements Analysis
Legal, societal , ethical expectations, ODD standards, System theoretic safety analyses,...
Resilient System Architectures
Uncertainty quantification, Self-adaptive systems, dynamic risk management,...
Continuous Verification and Validation
Simulation, extreme value analysis, field testing,...
Continuous Assurance
that a tolerable residual level of risk has been achieved, is maintained and residual assurance gaps closed in the field

Figure 3: Complexity-aware systems engineering

4.2 Definition of tolerable residual risk.

As systems become more complex, the concepts of risk and acceptable levels of safety become harder to define. Individual measures of risks such as failures over time will no longer be sufficient due to:

- Difficulty in demonstrating with sufficient confidence that the targets are met.
- The high level of heterogeneity in the operational design domain, making it more difficult to extrapolate observations made during test.
- Inevitably, this heterogeneity in the operating domain will lead to the need for scenario specific risk acceptance criteria which should be standardised across the industry alongside approaches to confirming that the criteria have been met.
- The evolving nature of the system environment and the system itself, meaning that historical data is an unreliable indicator of future risk.

As a result, individual quantitative measures of risk such as accidents over time lose relevance and must be complemented with additional quantitative and qualitative measures of risk that can be directly both accurately predicted during design and observed during operation.

At the time of deployment of the system, a set of risk indicators should be defined. These could be based on the performance of similar systems or previous system with lesser levels of automation (following the GAMAB principle[5]). Safe tolerance intervals and statistical confidence required of the evidence compared to the baseline should be defined and factored into target values. This baseline level of risk should be used as a reference during monitoring of the system in operation (see next section) and the statistical confidence in the values should be refined as more observations are made. The evaluation of these observations must also consider the heterogeneity of the scenarios in which the observations are made, in order to avoid false conclusions when performing statistical extrapolations. The targets will also need to be adjusted over time as the use of the system scales (e.g. failure per operational hour) and societal expectations adapt.

[5] The French GAMAB (Globalement au moins aussi bon), i.e. globally, or generally, at least as good as) is a relevant risk acceptance principle when replacing one system with another, and could be considered a basis for evaluating highly automated driving systems where it is possible to get the comparator (baseline) data.

A monitoring scheme for observing these risk indicators should be established and managed by the operators of the vehicle in coordination with manufacturers. A subset of these indicators should be reported and regularly assessed by the responsible regulatory authorities (see next section). This requires a clarification of responsibilities for the collection and analysis of the data (e.g. between regulatory authorities, operators and manufacturers).

An initial approach to addressing these issues for autonomous driving was produced by the Centre for Data Ethics and Innovation (CDEI) [26], reflecting insights from ethics and responsible innovation, as well as systems, safety and software engineering. This report outlined a regulatory framework, building on the work of the Law Commission, separating responsibilities for pre-operational acceptance from in-service monitoring and risk evaluation. It proposed the production of a safety case for initial deployment that would be submitted to the relevant regulatory authority for approval. The safety case would be based on the definition of a Safe and Ethical Operating Concept (SEOC) which, *inter alia*, includes compliance with relevant road rules, avoidance of at-fault collisions, mitigates the risk of 'non-fault' collisions, deals with entry to and exit from the Operational Design Domain (ODD), etc. In this way a manufacturer would propose *its* way of achieving safety (the SEOC, which must comply with relevant legislation), and provide evidence that it had done so. This part of the approach is essentially technical and does not directly address residual risk.

Various bodies, e.g. the EU, have mentioned targets such as $10^{-7}/h$ hazardous events (see the EU regulations on automated driving [27]). In the proposed CDEI framework, responsibility for evaluating performance against such targets would lie with an in-use regulator, who would monitor and analyse data from operations. Should the observed accident rate exceed the target, the regulator would take action. For example, if the issues related to one manufacturer, then the regulator might work with them to reduce *assurance uncertainty* and, thereby, to drive improvements to that individual vehicle or ADS. Alternatively, if the problems identified affected all vehicle makes equally, then the regulator might conclude that this was due to *specification uncertainty*, and seek to update technical specifications and regulations accordingly. This gives a way of identifying and managing residual risk, through activity at the Governance layer in the TMG model, which then drives appropriate action at the Management layer, leading to changes in the Task & Technical layer (likely technical changes to the vehicle), avoiding problems of evaluating residual risk prior to deployment.

4.3 Availability and applicability of regulations and standards

Regulations and standards for automated driving functions The UNECE Working Party on Automated/Autonomous and Connected Vehicles has been preparing recommendations for harmonised regulations for safety aspects of ADS. These have included an UN regulation on Automated Lane Keeping Systems (ALKS) [28] as well as the Functional Requirements on Automated Vehicles (FRAV) [29], both of which define a number of high level safety requirements. These include the need to safely manage the dynamic driving task in all situations, including recognition of the boundary of the ODD, whilst complying with all relevant traffic rules within the country of operation (in this respect, there is a strong similarity with the proposed CDEI approach). Additionally the systems must recognise failure scenarios and be able to revert to fallback functionality. The regulations also contain requirements on the interaction with human operators including under foreseeable misuse. In August 2022, the European Union published regulation 2022/1426 [27] on the type approval of fully automated driving vehicles which contains similar requirements. Type approval is defined in terms of a fixed set of scenarios and associated tests an assessment of the ADS safety concept, credibility assessment of the toolchain for virtual validation and pre-requisite on in-service reporting.

In addition to the established safety standards ISO 26262 and ISO 21448, a number of standards and related specifications are currently under development to address safety-related aspects of automated driving. These include the definition of Minimal Risk Maneuvers (ISO/AWI 23793-1), scenario-based safety evaluation framework (ISO 34502), design, verification and validation aspects of ADS (ISO/AWI TS 5083), ergonomic aspects of driver monitoring and system interactions (ISO/AWI TS 5283) and safety ethical considerations (BS ISO 39003). Many of these standards are work in progress and some aspects require more specific guidance than is currently available (e.g. standard behaviours in reaction to emergency vehicles).

Regulations and standards for AI The EU commission has also begun work on developing a set of directives and associated initiatives related to trustworthy AI. Initially, a set of ethical guidelines for trustworthy AI was presented in 2019 [30]. The guidelines define Trustworthy AI as being: lawful, ethical and robust from the technical and social perspectives, where all three components should ideally work in harmony. The following requirements for trustworthy AI were identified in the report:

1. Human agency and oversight,
2. Technical robustness and safety,
3. Privacy and data governance,
4. Transparency,
5. Diversity, non-discrimination and fairness,
6. Environmental and societal well-being and
7. Accountability.

The guidelines recommend that these requirements are met through both technical and non-technical means. The report suggests that existing (safety) regulation may be sufficient to ensure that the requirements on trustworthiness are met, but new regulation should be considered if required to protect society from adverse impacts. Further, an agile approach to the continuous evaluation of the effectiveness and adaptation of the regulation is recommended. Governance frameworks (internal and external) are proposed to ensure accountability for the ethical dimensions of AI trustworthiness. A risk-based and multi-stakeholder approach [31] is recommended, including due consideration of the level of autonomy in AI-based decision making. It is also recommended that traceability and reporting requirements are introduced to enable auditability of the systems before deployment and oversight on an ongoing basis.

ADS that make use of AI technologies would fall under the risk category "high" of the EU AI directive, requiring them to fulfil mandatory horizontal requirements on trustworthy AI and undergo assessment before being placed on the market. However, for sectors where existing regulatory approaches are in place, the directive does not apply, with automotive falling into this category. When adopting, implementing or delegating regulation in these sectors, the requirements should nevertheless be taken into account. It is therefore to be expected that there will be an attempt to align regulation and standards for AI-based automated driving with the requirements of the EU AI directive. We see a remaining gap between the requirements for trustworthy AI as outlined in [30] and a set of technical criteria against which these requirements can be evaluated which would need to be addressed by sector specific regulations and standards. The CDEI report [26] might form a starting point for such regulations related to ADS.

The EU commission is currently drafting implementation decisions that will propose a number of standards to be defined at European level (CEN, CENELEC and ETSI) including standards for the risk management and robustness of AI systems. However, in parallel, activities have begun at an international level to fill this "standardisation gap". For example, within the joint ISO/IEC JTC1/SC42 sub-commitee on "Artificial Intelligence", a number of standards and technical reports have been developed or are in progress, including ISO/IEC TR 5469 "Artificial intelligence — Functional safety and AI systems". Within the domain of road vehicles, a sector-specific publicly available specification ISO PAS 8800 "Road vehicles - Safety and Artificial Intelligence" is currently under development. These standards are expected to go some way to close the current gap between societal and regulatory expectations and technical standards for ensuring safety. However, further guidance will still be required to provide use case and AI technology-specific guidance in the implementation of these documents. In particular, a sector and functionality specific operationalisation of the trustworthiness requirements will be essential in order to determine acceptable levels of residual risk of the systems. Further, it is not clear that these standards will address the need to consider both pre-deployment safety evaluation and in-service monitoring of safety, nor adequately cover the TNG layers.

Availability of applicable standards as a prerequisite for deployment The technical and organisational (i.e. in terms of safety management systems and operating procedures) approaches for implementing the requirements in the standards and regulations described above are still unclear and require solving the open challenges outlined in Sections 2 and 3 of this report. Published international safety standards represent industry consensus on the state-of-the-art for "conventional" systems. Therefore, for liability purposes they can be considered as the lower bound of best practice and compliance to the standard can therefore be reasonably expected. By providing a basis for conformity assessments and certification, standards can support regulation and secure an independent level of trust.

There is inevitably a lag between the development and application of new technologies and the development of standards. Standards require time to be developed and due to the processes of international consensus building, have to be of a limited scope in order for pragmatic progress to be made. This means that standards are either missing, in progress or lack sufficient detail to provide specific guidance to manufacturers, suppliers and operators on achieving and maintaining the safety of highly automated driving. We are not currently aware of standards, either in existence or in development, that are adequately addressing the topics of emergent complexity and resilience as defined above, nor do they span development and operation to a sufficient degree.

In general, the existence of, and conformance to, appropriate safety standards is currently a key barrier to the introduction of cognitive CPS for safety-related applications. This leaves regulators with the following choices when defining conditions for the release of highly automated driving systems:

1. Apply existing standards. This may include restricting either the scope of the functionality or the context of operation in such a way that existing standards can be used to assess the overall safety of the system. Third party assessment against the standards could then be defined as a pre-requisite for safe deployment.
2. Apply equivalent measures. Although directly applicable standards with technology-specific guidance may not yet exists, measures can be justified to achieve an equivalent level of residual risk, oriented towards the principles and objectives of existing standards. The development of *outcome-based standards* which define the set of criteria that the system should fulfil and expectations on associated evidence and argumentation strategies can provide more agility when introducing new technologies. This would result in the onus being on the manufacturers and operators of the system to provide a safety assurance argument that the system is adequately safe, despite the lack of detail in the standards. Nevertheless, a residual level of assurance uncertainty will remain and will need to be compensated for when defining the conditions for deployment and continuous monitoring of the performance of the system in operation. The approach proposed by the CDEI [26] is intended to provide such a framework.
3. Prohibit systems which cannot be argued to achieve an equivalent level of safety to existing standards or state-of-the-art. If no standards exist and alternative methods for achieving tolerable residual risk have not been justified, such systems should be prohibited. This might seem draconian, and contrary to the desire to enable responsible innovation, but this is one option, and perhaps the ultimate sanction, for regulators[6].

5 Continuous safety assurance and regulation

As described above, deployment decisions should be based on the availability of operational procedures for ensuring continuous safety. Inspired by the model of self-adaptive systems described in Section 2, in this section we describe a model for the continuous monitoring and adaption of the system, its operating procedures, safety management system and regulatory approach in order to continuously manage the risk of systemic failures due to emergent complexity and uncertainty within the system.

[6] We are aware of regulators outside the driving domain who took this as their starting point, but there now seems to be a growing move towards more flexible approaches.

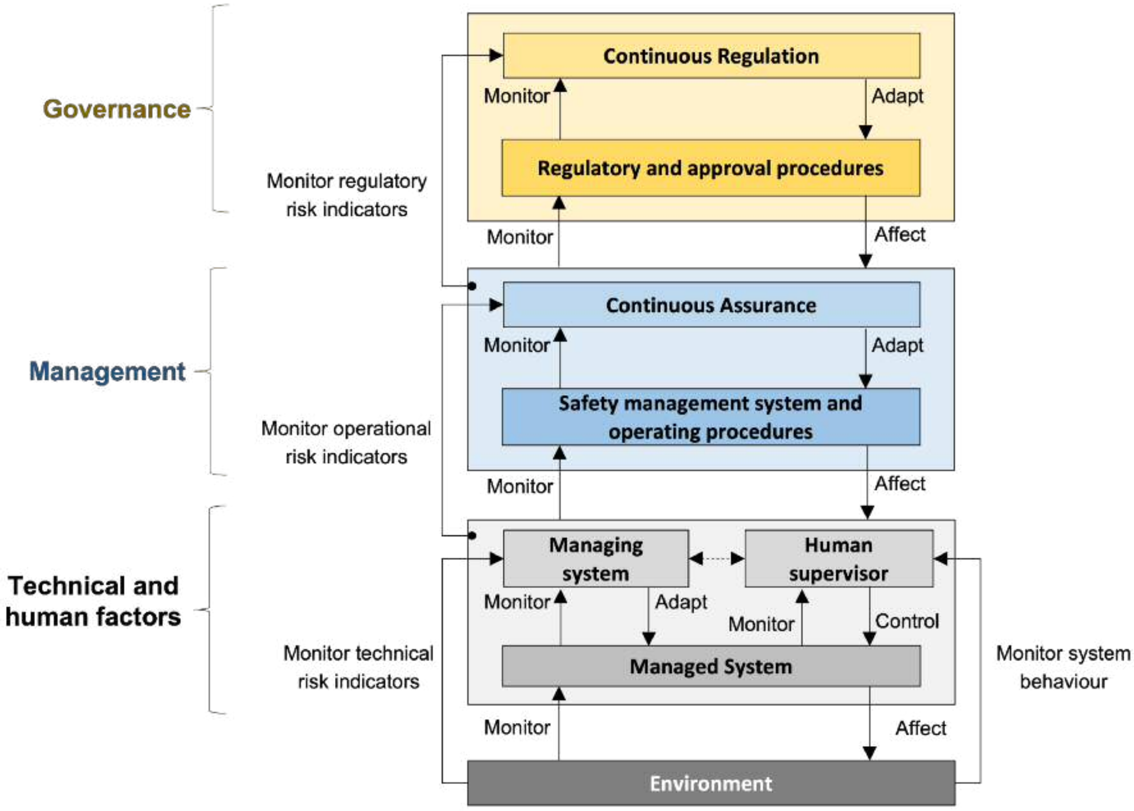

Figure 4: Hierarchy of continuous safety assurance

The objective of this model is to ensure resilience against emergent risks based on a systematic view of complexity and uncertainty and the associated controls as described in the SCS Framework (see Section 2). This includes a definition of a number of observation points at all three layers of the TMG framework. Ideally, these observation points will act as leading (as opposed to lagging) indicators based on observed properties of the system and its environment. Note, it may not always be possible to identify leading indicators in advance, and in many cases they may only be identified by analysing behaviour of the system *post-hoc* (for use in future iterations of the system)[7]. Based on these observations, e.g. the above-mentioned as $10^{-7}/h$ hazardous events, and an associated decision model, adaptions to the technical system, safety management system, operating procedures and regulatory approaches should be selected in order to reduce the residual risk associated with the system.

This model can be applied across all three TMG layers and requires the following questions to be answered at each layer:

- What observation points are required to monitor the emergent technical, operational and regulatory risk? This includes identifying leading indicators of deviations in safety performance.
- How can genuine deviations in safety performance from predicted or acceptable levels of variation be distinguished from transient effects which will self-correct?
- How can an effective, evidence-based risk management model for the interpretation of observations and selection of suitable measures be developed - despite the inherent complexities, uncertainties and exacerbating factors associated with the system? This, of course, links to concepts such as reducing risks As Low As reasonably Practicable (ALARP), GAMAB or other criteria such as Minimal Endogenous Mortality (MEM).

[7] The CDEI report introduces the concept of notifiable events which give the regulators the ability to specify these observation points and to adapt their definition over time.

We define a hierarchy of dynamic safety management, continuous assurance and continuous regulation as described in Figure 4. At each of these layers, the principles identified above will be applied and the same fundamental questions must be answered

5.1 Technical & Human factors layer

At the technical and human factors layer, increased resilience against technical uncertainty e.g. (residual SOTIF unknown triggering events) can be achieved through either technical measures in the system (see Section 2.4 on self-adapting systems) or through human supervision and intervention.

It becomes clear from the representation in Figure 4, that there is the potential for conflicting actions between the monitoring/managing system components and the human operator. Therefore the amount of self-adaptation allowed by the system should be balanced against the potential for competent and effective human supervision for monitoring residual risk. For example, where close human supervision is possible and the human operator is expected to detect and override anomalous behaviour in a timely enough manner to avoid accidents (e.g. for a lane keeping system), the amount of self-adaptation in the system should be carefully designed to avoid mode confusion and conflicts between the human and technical supervisors. For higher automation levels where less direct human supervision is possible, more self-adaptation, or at least dynamic safety management may be allowed or even required in order to ensure the necessary resilience.

One could argue that the ability to develop truly *self*-adapting systems for complex tasks is limited by the semantic gap. Specifically, if it was possible to define the conditions under which system adaptations should be made then the managed system itself should be developed to cope with these conditions and react accordingly, negating the the need for the managing system. However, this leads to systems that are robust against well understood situations and risks but not resilient against unforeseen situations or risk.

Therefore self-adapting technical systems make use of a managing system that utilises a fundamentally different model than the managed system and can make observations and associated deductions which the managed system cannot. In other words, resilience against factors that the technical control system cannot reason about itself (from the perspective of the control system). This analysis appears to suggest that the application of machine learning techniques for developing such advanced models for managing systems would bear promise. The research challenges associated with this idea are explored further in Section 6.

Still, there maybe some compelling arguments for applying this architecture paradigm, despite these restrictions. For example, where heterogeneous redundancy can be argued between the managing and managed system, or where the focus on the managed system would be to optimise utility of the function where the managing system would optimise safety, imposing functional restrictions where a suitable level of residual risk cannot be guaranteed by the managed system. Whilst these arguments may be valid for basic models of system failures, for systemic failures and complex systems with high levels of uncertainty, it will be difficult to escape the trap of the semantic gap.

One important consideration at this layer is how to analyse the safety of the (potentially dynamic) allocation of function between human and system, particularly where the system is carrying out functions autonomously. One approach is to extend classical models of the task to identify where the function is undertaken – human, system, or shared – and which responsibility each actor has – undertake (do), monitor, ensure safety. This approach has been adopted in the context of autonomous driving and ADS [17]. This enhanced model allows safety analysts to assess the impact of shared control, taking into account the capabilities and limitations of each actor. This approach has been used to extend STPA with additional prompts for failures that can arise due to the complexities of function allocation[8], and thus to provide a richer assessment of the human-system interaction than arises from the control-theoretical model underlying STPA alone.

[8] The method and examples will be published in Helen Monkhouse's PhD thesis, to appear in 2023.

A more direct consideration of human control over autonomous functions distinguishes operational controls from those that must be implemented at design-time, and considers the need for assurance in each case [32]. *Inter alia*, it identifies the conditions for effective human control especially for an over-ride function, which include understanding of: the limitations of the system, the consequences of actions, and the attendant ethical issues, as well as sufficient time to evaluate these issues. To be effective, there needs to be sources of information (knowledge) independent of the system and the ability of the human to carry out the action, especially if it is now rarely undertaken due to the level of autonomy of the system[9]. Further, aspects of human behaviour such as cognitive biases that might lead to an operator 'approving' a recommendation from the system without proper evaluation (confirmation bias) need to be considered. Such factors should be evaluated and suitable arguments and evidence provided in a safety case that the human-machine system is acceptably safe.

5.2 Management and Operations

An assurance case provides a convincing and valid argument that a set of claims regarding the safety of a system is justified for a given function based on a set of assumptions over its operational context. However, due to the impact of complexity and uncertainty described elsewhere in this report, the ability to construct a convincing argument, before deployment into the target domain, that remains valid over time is limited.

This issue has been addressed in work on continuous assurance [33] and dynamic safety cases [34]. These concepts[10] address remaining and emerging insufficiencies [35] of the assurance argument, and thus potential residual risk by providing a framework for the collection of evidence and re-evaluation of the assurance argument. This could be necessary as the system is adapted over time (e.g. due to software updates), the operating environment system evolves or previously undiscovered risk factors (e.g. unknown SOTIF triggering conditions) come to light. The fundamental difference between continuous assurance and dynamic safety management measures applied at the technical level, is that continuous assurance has the objective of identifying (emerging) flaws in the safety argumentation itself rather than maintaining the reliability of the service at a technical level. However, the two concepts go hand-on-hand as shown in Figure 4 where adaptions made to the system at the technical level must be evaluated in the assurance argument.

Analogous to the dynamic safety management and self-adaptive systems concepts at the technical layer, continuous assurance requires both a model of observations that can be made of the environment, the technical systems, and the effectiveness of the safety management system (including the assurance case) and operating procedures. We refer to these observations as *operational risk indicators* and include the following classes of properties:

- Properties of the environment: These properties can be used to confirm or invalidate assumptions made on the environment within the assurance case. This can include the types of expected events (including previously unknown triggering conditions) and their distribution (e.g. to detect distributional shift). These observations would different from those made by the managing system at the technical layer in that automated adaptations to the managing system cannot or have not been defined and therefore require measures at the management and operation layer to maintain an acceptable level of risk to counteract the emergent properties of the environment. It should be noted that some of these observations may be made through the filter of the technical system itself (e.g. black-box recording or similar mechanisms).
- Interactions between human users and the system: These observations can be used to infer properties related to the interactions between human users and the technical system which may provide indicators for emergent risk. For example, an increased frequency of interventions by human operators could indicate a change in the operational conditions and increased risk.

[9] This shows the need for linkage between the layers in the TMG model. One mitigation for this issue is a policy at the Management layer, or a requirement at the Governance layer, that critical risk mitigation activities are rehearsed.

[10] Both concepts are similar and use slightly different terms, within this report we will use the term "continuous assurance and continuous assurance cases" to better differentiate between these concepts and those of dynamic safety management applied at the technical layer.

- Properties of the technical system and its behaviour: These observations provide details of the status of the technical system itself. This can include, on-board evaluation of technical uncertainty (e.g. out of distribution detection in ML components), general diagnostics and health monitoring of critical components as well as details regarding the frequency of interventions by the managing system within a self-adaptive model.
- Properties of the assurance case: Finally the assurance case and operating procedures shall be described in such a manner that their ongoing validity can be analysed given the observations described above. This will require the definition of a set of conditions that are able to describe bounded tolerances of the validity of assurance claims and associated evidence. This includes a definition of the frequency with which the assurance case is reviewed or which set of triggering events would lead to a review.

Deficiencies discovered in the assurance case based on the observations above can lead to a number of measures as part of the adaptation of the safety management system and operating procedures. These can include:

- The generation of additional evidence (e.g. tests) to reduce assurance uncertainty.
- Specification of technical adaptations to the systems (that would require a re-appraisal of the assurance case).
- Adaptations to the operating procedures of the system. This can include restrictions to the ODD as well as operating and maintenance procedures required to reduce residual risk. This is not uncommon in other industries, e.g. increasing the frequency of maintenance check intervals for a suspected new or worsened hazard cause.

Whilst we are not aware of (public domain data on) such monitoring and feedback being undertaken in the automotive domain, this is an area where AI/ML might be useful in analysing data and identifying patterns of behaviour which are at variance with what was predicted in the safety analysis. An example from healthcare using Bayesian Belief Networks to analyse healthcare data and to identify where causal factors in practice deviated from those identified in the hazard and safety analysis [36] indicates what might be possible.

5.3 Regulation and Governance

Regulatory guidelines and objectives such as those proposed by the EU for the use of AI [30] rely on standards to provide more concrete recommendations on how the objectives of the regulations can be achieved. However, due to the increased pace of technological change as well as the inherent uncertainties within assurance, it is increasingly difficult to develop a set of regulatory guidelines and associated standards that are sufficient to protect the public from emergent risk of cognitive CPSs whilst ensuring that the societal benefits of innovative technologies are not inhibited. This implies the need for an agile approach to regulation with the support of standards.

We propose to extend the model described so far to describe continuous assurance and dynamic safety management to include the concept of *continuous regulation*. This implies that at the layer of regulation and governance, a model of observations and adaption measures is required. The observation points must be selected in order for regulators to react to emergent risks in a timely fashion, whilst the decision model must take into account an understanding of the effectiveness of current regulatory measures and potential adaptations that could be applied to counteract emergent risks. *Regulatory risk indicators* could include the following types of observations:

- Divergences between the predicted level of residual risk as defined by safety management systems and associated standards and the actual achieved level of risk for new classes of technologies. This should include a comparison of risk achieved using new classes of technologies and previous technologies (according to the GAMAB principle), for example are an increased number of incidents observed involving new vehicle functions?
- Systematic *post-hoc* activities should be applied to evaluate root causes of accidents (and preferably near-misses) and continuously identify improvements in the safety management

and regulatory procedures. For post-hoc analysis the SCS Framework could be applied to gain a holistic understanding of the root causes of incidents.

- Whilst the above observations would involve lagging indications of risk, leading indicators cannot be defined so easily and require more research. However, one approach could involve requiring the reporting of selected and consolidated (e.g. at fleet level) operational risk indicators such as the number of interventions required by human operators during system monitoring, number of technical system updates and required changes to the assurance case.
- A framework should be put in place to ensure that the proposed regulations for automated driving can be rapidly adapted based on experience in initial deployments and emerging risk associated with the technologies as well as developing standards in this area. This involves making use of different standards mechanisms such as publicly available specifications (PAS) that allow for a faster development of recommendations, but also the active participation of regulators in the development of these standards to ensure an early alignment of principles. This may require a re-qualification of systems at regular intervals to incorporate the latest updates to standards and best practices.
- An evaluation of public perception of risk associated with the systems.

Whilst it doesn't use the term continuous assurance explicitly, the recommendations of the CDEI Report [26] can be seen as an attempt to define such an assurance process.

The ability to reflect upon the effectiveness of regulations and standards in order to identify additional measures will require some model of the way in which the regulatory guidelines "work" so that the impact of changes can be evaluated before changes are made. As part of the EU AI directive, an agile approach to regulation, including the use of regulatory sandboxes has been recommended that would accompany pilot projects to support the development and evaluation of regulatory approaches [37]. Here, we also see the potential for the application of the SCS framework to gain a holistic understanding of the risks inherent in the system and the potential benefit of measures across the TMG layers [2]. An area of future research would therefore involve the development of a model of risk uncertainties that emerge over time during the introduction and successive roll-out of such technologies. This model would be deployed within the "managing system" of the continuous regulation cycle of Figure 4. The development of such a model would require a strongly interdisciplinary definition of acceptable risk as well as a definition of the observation points (regulatory risk indicators) required to evolve the regulations.

Open questions remain regarding the distribution of responsibilities between manufacturers, operators and regulatory authorities for the safety of automated driving systems. This includes a clear definition of accountability and a pro-active closing of liability gaps [6] related to the complexity and resulting semantic gaps in the system.

6 Remaining research challenges and conclusions

This report has summarised the challenges in assuring the safety of highly automated systems that operate in complex environments, with particular focus on automated driving. Definitions of complexity and uncertainty were introduced in order to motivate these challenges. In order to structure approaches to addressing these challenges from the perspective of technical and human factors, operations and management and regulation and regulation and governance (referred to as the TMG layers), two models were introduced. The first model (see Figure 2) describes a causal approach to understanding systemic failures across the TMG layers and the impact of design and operation time measures for reducing risk. The second model (see Figure 4) describes a dynamic perspective of how these layers can interact as part of a continuous regulatory and assurance approach. In particular, this perspective allows us to reason about the observation and decision models required to maintain a tolerably low level of residual risk. These models also allow us to identify a number of areas of research that are still required in order for such a scheme to be feasibly implemented:

- Criteria for the evaluation of the systematic task complexity of a particular function operating within a particular context should be defined. These criteria would be used to evaluate the potential for system failures and identify appropriate measures to be applied at the technical and human factors layer for mitigating this risk. These criteria can also be based on reference systems which can be used as a comparative benchmark for complexity. For example, measures that have been demonstrated effective for an automated lane keeping system operating in congenial weather conditions on well-mapped sections of road could be transferred to functions where similar levels of complexity can be argued to hold (e.g. adaptive automated braking or overtaking assistants).

- System-theoretic safety analysis techniques should be extended to consider uncertainty and emerging complexity within the system and in the interactions between the system, its environment and human operators. Combinations of approaches from FRAM and STPA may provide some inspirations for such approaches, in particular when modelling the interactions between human drivers, other traffic participants and the automated driving system. The work on analysing function allocation between human and system [17] is an example of what needs to be done.
- Remaining intrinsic complexity within the system may lead to rapid changes of state due to tipping points and non-linear behaviour. Therefore, näive approaches to system health monitoring and diagnostics may not be able to react in a timely enough manner in order to mitigate the risk of systemic failures due to this complexity. More research is therefore needed on how to define observation and decision models for leading indicators of risk and to derive appropriate actions across all the TMG layers.
- When applying the self-adaptive systems paradigm to dynamic risk management and continuous assurance of ADS, models for the “managing system” component are required that are able to deduce the most appropriate adaptations based on a set of observations regarding the system state and the environment. One approach could include making use of explainable ML techniques to provide a salient prediction of system dependability based on environmental conditions and systems state. Some initial work in this area has also been undertaken for building dependability models of ML components [38]. Data-driven approaches to such models will require the systematic collection of relevant data across all TMG layers. The principle objective of such models would be to identify leading indicators of risk such that a system adaptation can be applied in a timely enough manner in order to reduce potential hazards.
- An alignment of a technical understanding of complexity, uncertainty and residual risk with ethical and societal definitions of risk is required in order for regulatory authorities to make appropriate deployment and adaption decisions whilst finding a balance between utility and risk. This will include the definition of a common language to allow technical safety specialists and ethicists to converse and in particular to discover which which ethical questions can be reasonably imposed upon the system and where the limits of either the technical system itself or the assurance of the system prevent certain ethical questions being answered.

Finally, the analysis laid out in this report highlights the need for action at the level of standards and regulations on the safety of AI-supported ADS, which can be summarised as follows:

- Safety standards should acknowledge the inherent issues of uncertainty and emergent complexity in systems and the impact on risk. This has an impact on the way that safety requirements are expressed, the role of system theoretic safety analyses and the acceptance of potential confidence deficits in assurance arguments.
- Regulation should define rigorous pre-requisites for safe deployment (see Section 4) and ensure continuous assurance principles are in place (see Section 5). Deployment pre-conditions should include an impact assessment of systematic task complexity and an analysis of the suitability of existing standards to manage risk in the system. Continuous assurance must go beyond periodic road-worthiness tests by including a continuous evaluation of functional insufficiencies in the system and a monitoring of the safe deployment conditions (e.g. to address distributional shift in the environment).
- Regulation needs to define conditions for acceptable levels of residual risk for common au-

tomated driving use cases that go beyond simple references to existing accident rates as a direct comparison between human-related errors and systematic machine-induced errors may not be ethically tolerable (see research recommendation above). Instead, regulation shall define concepts of risk that can be evaluated before deployment and that stand up to robust ethical judgement whilst considering the technical limits to the nature of evidence that can be used within such a judgement.

Whilst we have articulated and illustrated these issues in the context of autonomous vehicles, we believe that the overall considerations and need for evolution in the approach to regulation applies across domains. Of course there will be differences, e.g. in terms of risk acceptance criteria, but the need for changes in regulatory practices and mechanisms is, we believe, common across many domains.

References

[1] John Alexander McDermid. “Safe, Ethical & Sustainable: A Mantra for All Seasons?” In: *Safety Systems* (2022), pp. 5–10.

[2] Simon Burton et al. “Safety, Complexity, and Automated Driving: Holistic Perspectives on Safety Assurance”. In: *Computer* 54.8 (2021), pp. 22–32.

[3] John McDermid et al. *Safer Complex Systems – An Initial Framework*. `https://raeng.org.uk/media/4wxiazh3/engineering-x-safer-complex-systems-an-initial-framework-report-v22.pdf`. Royal Academy of Engineering, 2020.

[4] Peter Erdi. *Complexity Explained*. en. Springer Science & Business Media, Nov. 2007.

[5] Warren E Walker et al. “Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support”. In: *Integrated assessment* 4.1 (2003), pp. 5–17.

[6] Simon Burton et al. “Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective”. In: *Artificial Intelligence* 279 (2020), p. 103201.

[7] Roman Gansch and Ahmad Adee. “System theoretic view on uncertainties”. In: *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE. 2020, pp. 1345–1350.

[8] Carl Bergenhem et al. “How to reach complete safety requirement refinement for autonomous vehicles”. In: *CARS 2015-Critical Automotive applications: Robustness & Safety*. 2015.

[9] *ISO 26262: Road vehicles - Functional Safety, Second Edition*. Tech. rep. Geneva: International Standards Organisation (ISO), 2018.

[10] Rakshith Amarnath et al. “Dependability challenges in the model-driven engineering of automotive systems”. In: *2016 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. IEEE. 2016, pp. 1–4.

[11] International Organization for Standardization. *Road Vehicles — Safety of the Intended Functionality*. Standard ISO/PAS 21448:2019. ISO, 2019.

[12] Robin McDermott, Raymond J Mikulak, and Michael Beauregard. *The basics of FMEA*. SteinerBooks, 1996.

[13] Glenn Bruns and Stuart Anderson. “Validating safety models with fault trees”. In: *SAFECOMP’93*. Springer, 1993, pp. 21–30.

[14] Nancy Leveson. *Engineering a safer world: Systems thinking applied to safety*. MIT press, 2011.

[15] Asim Abdulkhaleq et al. “A systematic approach based on STPA for developing a dependable architecture for fully automated driving vehicles”. In: *Procedia Engineering* 179 (2017), pp. 41–51.

[16] Siddartha Khastgir et al. “Systems approach to creating test scenarios for automated driving systems”. In: *Reliability engineering & system safety* 215 (2021), p. 107610.

[17] Helen E Monkhouse, Ibrahim Habli, and John McDermid. “An enhanced vehicle control model for assessing highly automated driving safety”. In: *Reliability Engineering & System Safety* 202 (2020), p. 107061.

[18] Jens Rasmussen. “Risk management in a dynamic society: a modelling problem”. In: *Safety science* 27.2-3 (1997), pp. 183–213.

[19] Erik Hollnagel. *FRAM, the functional resonance analysis method: modelling complex socio-technical systems*. Ashgate Publishing, Ltd., 2012.

[20] Betty HC Cheng et al. “Using models at runtime to address assurance for self-adaptive systems”. In: *Models@ run. time*. Springer, 2014, pp. 101–136.

[21] Henry Muccini, Mohammad Sharaf, and Danny Weyns. “Self-adaptation for cyber-physical systems: a systematic literature review”. In: *Proceedings of the 11th international symposium on software engineering for adaptive and self-managing systems*. 2016, pp. 75–81.

[22] Mario Trapp, Daniel Schneider, and Gereon Weiss. “Towards safety-awareness and dynamic safety management”. In: *2018 14th European Dependable Computing Conference (EDCC)*. IEEE. 2018, pp. 107–111.

[23] Joseph Sifakis and David Harel. “Trustworthy Autonomous System Development”. In: *ACM Transactions on Embedded Computing Systems* (2022).

[24] Rogério de Lemos et al. “Software engineering for self-adaptive systems: Research challenges in the provision of assurances”. In: *Software Engineering for Self-Adaptive Systems III. Assurances* (2017), pp. 3–30.

[25] Jan Reich and Mario Trapp. “SINADRA: towards a framework for assurable situation-aware dynamic risk assessment of autonomous vehicles”. In: *2020 16th European Dependable Computing Conference (EDCC)*. IEEE. 2020, pp. 47–50.

[26] *Responsible Innovation in Self-Driving Vehicles*. `https://www.gov.uk/government/publications/responsible-innovation-in-self-driving-vehicles/responsible-innovation-in-self-driving-vehicles`. Centre for Data Ethics and Innovation, 2022.

[27] *COMMISSION IMPLEMENTING REGULATION (EU) 2022/1426, Automated cars - technical specifications*. `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R1426&qid=1661841316689&from=EN`. European Commission, 2022.

[28] *Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping Systems, ECE/TRANS/WP.29/2020/81*. `https://unece.org/sites/default/files/2021-03/R157e.pdf`. United Nations Economic Commission for Europe (UNECE), 2021.

[29] *Functional requirements on automated and autonomous vehicles (FRAV), Consolidated ADS safety text pursuant to the 21st FRAV session, ECE/TRANS/WP.29/2020/81*. `https://wiki.unece.org/download/attachments/140710465/FRAV-21-05.pdf?api=v2`. United Nations Economic Commission for Europe (UNECE), 2021.

[30] *Ethical Guidelines for Trustworthy AI*. Tech. rep. Brussels: Independent high-level expert group on artificial intelligence, European Commission, 2019.

[31] *Policy and investment recommendations for trustworthy AI*. Tech. rep. Brussels: Independent high-level expert group on artificial intelligence, European Commission, 2019.

[32] John McDermid. *Responsible Innovation in Self-Driving Vehicles*. `https://www.york.ac.uk/assuring-autonomy/news/blog/human-control-ai-autonomy/`. Assuring Autonomy International Programme, 2019.

[33] Fredrik Warg et al. “Continuous Deployment for Dependable Systems with Continuous Assurance Cases”. In: *IEEE International Symposium on Software Reliability Engineering Workshops*. 2019.

[34] Ewen Denney, Ganesh Pai, and Ibrahim Habli. “Dynamic safety cases for through-life safety assurance”. In: *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*. Vol. 2. IEEE. 2015, pp. 587–590.

[35] Richard Hawkins et al. “A new approach to creating clear safety arguments”. In: *Advances in systems safety*. Springer, 2011, pp. 3–23.

[36] Yan Jia. “Improving medication safety using machine learning”. In: *AIME 2019* (2019).

[37] *Artificial intelligence act and regulatory sandboxes*. `https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf`. European Commission, 2022.

[38] Iwo Kurzidem et al. “Safety Assessment: From Black-Box to White-Box”. In: *2022 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. IEEE. 2022.