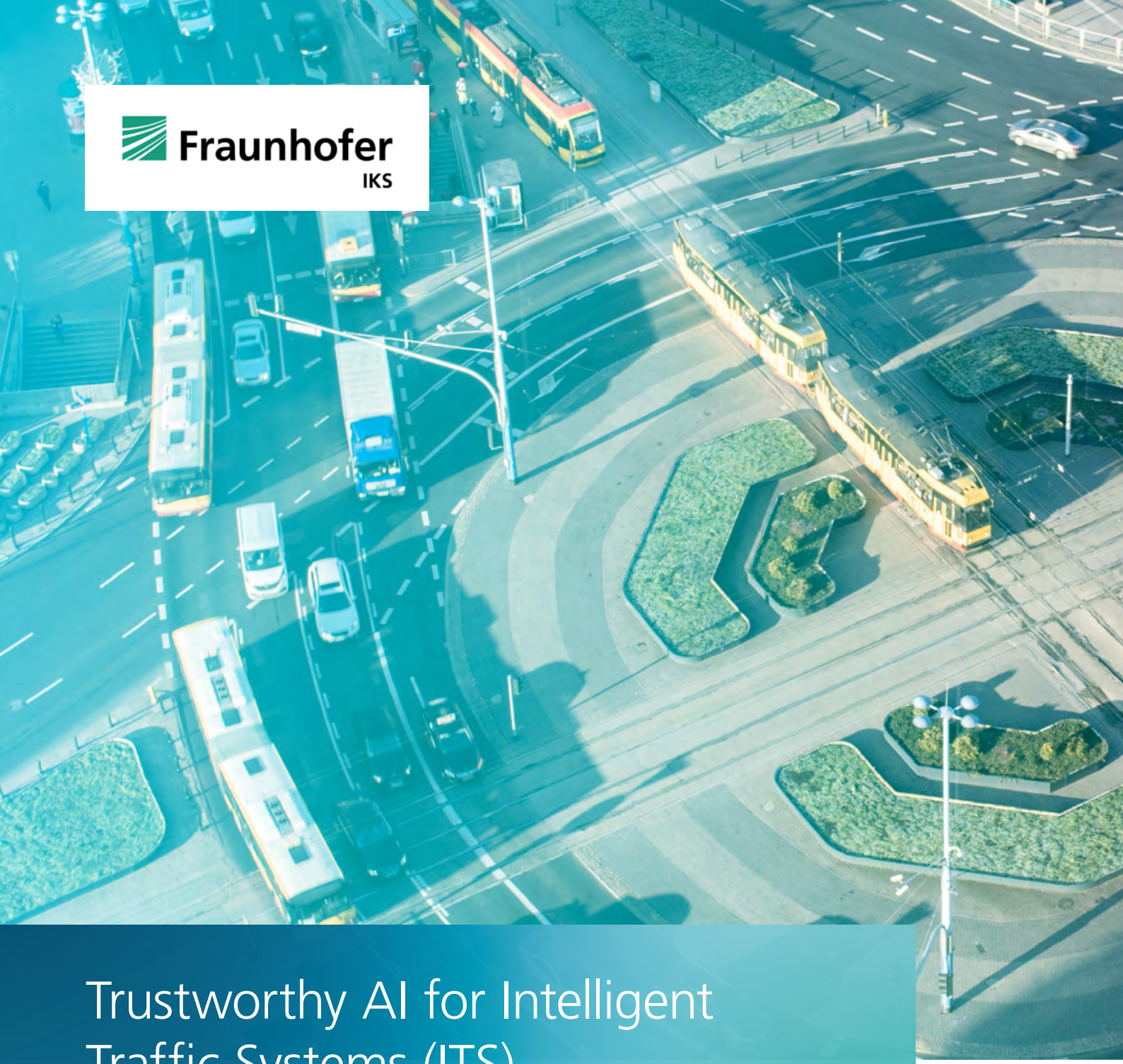

Fraunhofer IKS

Trustworthy AI for Intelligent Traffic Systems (ITS)

Fraunhofer IKS in cooperation with the Huawei Research Center Munich

#SafeIntelligence

Executive Summary

AI-enabled Intelligent Traffic Systems (ITS) offer the potential to greatly improve the efficiency of traffic flow in inner cities resulting in shorter travel times, increased fuel efficiency and reduction in harmful emissions. These systems make use of data collected in real-time across different locations in order to adapt signaling infrastructure (such as traffic lights and lane signals) based on a set of optimized algorithms. Consequences of failures in such systems can range from increased congestion and the associated rise in traffic accidents to increased vehicle emissions over time. This white paper summarizes the results of consultations between safety, mobility and smart city experts to explore the consequences of the application of AI methods in Intelligent Traffic Systems. The consultations were held as a roundtable event on the 1st July 2021, hosted by Fraunhofer IKS and addressed the following questions:

- How does the use of AI fundamentally change our understanding of safety and risk related to such systems?
- Which challenges are introduced when using AI for decision making functions in Smart Cities and Intelligent Traffic Systems?
- How should these challenges be addressed in future?

Based on these discussions, the white paper summarizes current and future challenges of introducing AI into Intelligent Traffic Systems in a trustworthy manner. Here, special focus is laid on the complex, heterogeneous, multi-disciplinary nature of ITS in Smart Cities. In doing so, we motivate a combined consideration of the emerging complexity and inherent uncertainty related to such systems and the need for collaboration and communication between a broad range of disciplines.

Contents

1. Introduction

Smart Cities as a concept was first developed in the 1990s and can be defined in various ways. In [1] the authors define a Smart City as "a concept that integrates information technologies into urban areas, to overcome urban challenges, improve sustainability in cities, and enhance citizens' quality of life". Intelligent Traffic Systems (ITS) can be seen as a sub-domain of Smart Cities and offer the potential to greatly improve the efficiency of traffic flow within a city. Distributed sensing and centralized analysis and control can be used to continuously analyze the current traffic situation and process large amounts of real-time data to support advanced optimization strategies. Free from the computational resource limitations of current road-side infrastructure, the cloud-edge computing continuum and modern communication technologies enable the exploration of novel paths unavailable to legacy systems.

Despite their advantages, failures of such systems will have wide-ranging consequences. These range from short term traffic disruptions to long term impacts such as increased accident rates or emissions. Therefore, for such systems to be accepted by the general public and city authorities, their utility must be clearly demonstrated and risks emerging from the introduction of the systems carefully managed. However, there are several characteristics inherent in ITS that create significant challenges when arguing the trustworthiness of the systems. ITS operate within a complex socio-technical context. They consist of interconnected technical systems-of-systems that interact with many different human stakeholders. These stakeholders, that include both city authorities and the general public, may pursue multiple, often conflicting goals. For example, getting across the city as fast as possible vs. safety and environmental considerations. This leads to difficulties in defining an exact specification of "desirable" behavior. The use of AI components that make opaque, imprecise decisions further exacerbates the challenges involved in developing and deploying trustworthy AI-based ITS. This also includes evaluating and communicating the utility and risk against such a diverse of expectations.

This report summarizes the deliberations of a safety expert round table, hosted by Fraunhofer IKS on the 1st July to examine the challenges and potential solutions to trustworthy AI in the context of ITS. The round table consisted of safety, mobility and smart city experts representing academia, industry, and public authorities. Although we primarily address this topic from the perspective of traffic systems, many of the concepts will equally apply to other AI-based smart city applications. We therefore consider the topic from the perspective of complex socio-technical systems which are likely to become more and more reliant on the use of advanced AI-based technologies.

The report is structured as follows: After a brief review of previous work on the topic of risk in Smart City applications, we describe an example of how AI can be used to optimize traffic light signal phases to optimize traffic flow across a city. We then investigate the impact of emergent complexity and uncertainty within the system on the ability to argue trustworthiness. This leads to the identification of several key challenges that need to be overcome when introducing such technology. We conclude this report with a collection of ideas for future directions of work in this area.

2. Consideration of risk in smart cities and Intelligent Traffic Systems

A systematic literature review [1] reviewed work on Smart Cities between the years 2000 and 2019 in order to identify the origins, trends, and categories of risk occurring in Smart Cities. The authors divided the risks into three categories:

- Technological: e.g., cyber security, threat of data loss, etc.,
- Organizational: e.g., absence of required competencies of authorities in a Smart City, and
- Social: e.g., stakeholders' conflict, mistrust of society to new technologies, etc.

Among the surveyed papers, 52% of the articles investigated technological risks, 32% addressed organizational risks, and only 16% considered social risks. In their conclusion, the authors highlighted that there is an urgent need for further research in this area, especially with respect to social risk.

A similar study [2] focused on safety and security issues and provided the concept of a Safe City. They defined a Safe City as "a city, that by the integration of technology and natural environment increases the effectiveness of processes in the field of safety, in order to reduce crime and terror threats, to allow its citizens to live in a healthy environment, provide simple access to healthcare, and to achieve readiness and quick response to threatening or arising emergencies". The authors pointed out that to evaluate Smart Cities, first the main components of every feature and every system needs to be defined. Based on this evaluation, strengths and weaknesses of the corresponding city system can be identified. Moreover, the authors stressed that the education of citizens about the use of such systems is also crucial. However, this will be a non-trivial task, as Smart City infrastructure may consist of a collection of loosely coupled technical and social systems including many legacy components.

Another study on Smart Cities was conducted in [3]. Their research showed that there is an urgent need for security principles and standards, as well as regulations to improve the safety and security of Smart Cities. The authors also pointed out that there is a need for further investigation on the collaboration between safety and security disciplines to understand and mitigate risks and vulnerabilities. Overall, there is a need to establish functional standards, responsibilities of liabilities and practices cross-countries. Despite this growing awareness for the consideration of safety and security risks within a Smart City context, we are nevertheless not aware of a significant body of work into methods for assuring the trustworthiness of such systems and especially not for the topic of Intelligent Traffic Systems. As we will see later, we believe that this is due to several significant paradigm shifts required to address these issues.

In [4] the authors provided an overview of the main challenges facing the implementation of intelligent traffic systems. These included the integration of heterogeneous data stemming from different systems, the management of big and knowledge representation as well as the identification of hazards caused by malfunctions of the system. Other work, such as [5] has directly addressed the issue of safety associated with traffic management systems. In this work, safety analysis tools such as Hazard and operability (HAZOP) analysis and Fault Tree Analysis (FTA) was used to evaluate the risk of hazardous events associated with dynamic speed restrictions on all-lane running smart motorways in the UK. Despite a successful trial study which demonstrated an increased level of overall road safety, the UK government decided to pause the overall rollout of such systems due to residual safety issues. In particular, concerns were raised regarding the impact of changes to system parameters on safety (e.g., distance between emergency refuge areas and response times to breakdowns in live lanes) that were not fully considered during initial deployment. A more in depth analysis of the safety issues and their causes related to the deployment of smart motorways in the U.K. can be found in [6] (Annex C.2).

The work summarized here has highlighted that risks associated with smart cities and intelligent traffic systems can emerge due to their complex interdependencies on their environment and deployment parameters. This suggests that a comprehensive and holistic consideration of their deployment context is required to ensure the trustworthiness of such systems.

3. AI-enabled Intelligent Traffic Systems

ITS consist of various sub-fields including Advanced Traveler Information Systems, Advanced Traffic Management Systems, Advanced Public Transportation Systems and Emergency Management Systems [14]. AI has the potential to support many of these applications including the application of digital twins in town planning and consultation processes [7], AI-assisted location of available parking spaces, and infrastructure supported automated driving applications.

To illustrate the topics discussed in this white paper we shall focus on real-time traffic flow optimization systems as these demonstrate many of the challenges involved in arguing the trustworthiness of AI-based ITS. The primary objective of these systems is to optimize traffic flow whilst at the same time minimizing negative impacts on road safety and the environment. A city-wide approach to optimization must react to adverse, unforeseeable events (accidents, weather conditions, disruption to public transport, failures of technical infrastructure) and adapt signaling schedules in real-time. Previous generations of such systems were based on localized sensing systems limited to vehicles and public transport and bound to localized optimizations of specific intersection or a corridor linking intersections. The focus of much work over the last few years has been on the use of AI for both traffic planning optimization (e.g. [8],[9],[10]) and traffic forecasting [11]. The next generation of AI-based ITS rely on data collected throughout a city through different forms of sensors and the cloud-edge continuum to compute optimal configurations in real-time across a city. Such systems target not only the flow of passenger vehicles but consider the mobility infrastructure as a whole. These systems will therefore build upon a hierarchy of both new and legacy systems such as the existing intersection control that ensure that traffic lights are not switched to an inconsistent pattern (e.g. green in conflicting directions) but are only able to make localized decisions regarding switching patterns.

Adaptive Traffic Light Optimization systems (ATLO) are a form of traffic flow optimization that dynamically adjust traffic light configurations across many intersections, optimizing multiple weighted traffic objectives (e.g., throughput, delay, number of vehicle stops or queue length). At each intersection, cameras collect video data which is subsequently processed by a stream compute unit (SCU) using machine learning (ML) methods such

Intelligent Traffic Systems are composed of a set of application and management tools to improve the overall traffic efficiency and safety of the transportation systems.«

Convolutional Neural Networks to classify and detect the types and trajectories of vehicles. The SCU then generates discrete information summarizing the state of traffic over time. This information is fed to forecasting and optimization algorithms and used to adapt the global traffic light plans across intersections. The system must consider both local and regional properties within its optimization space, leading to the need to ensure a complex and dynamic equilibrium between these perspectives. Once selected, the updated traffic light plans are transmitted to the intersection specific controllers, thus closing the loop to the traffic flow to be optimized. Machine Learning algorithms therefore support this application by both accurately sensing the current state of traffic using Convolutional Neural Networks to detect and classify traffic objects and by predicting the future state of the traffic in terms of multiple traffic metrics based on forecasting and optimization algorithms.

Due to the complexity of the operating environment, the dependability requirements for this system may not be immediately obvious and may vary based on various stakeholder perspectives. Furthermore, the lack of predictability and transparency of many ML techniques lead to concerns regarding unforeseen failures and unwanted side effects, which in turn could hinder public acceptance. The topic of dependability and trustworthiness of AI-based ITS therefore requires a more systematic investigation. However, a comparison and evaluation of work in the area of AI-based ITS is difficult due to the use of different performance indicators, synthetic scenarios and traffic simulations being used to test the systems [12]. Furthermore, maximizing the performance of ITS must be seen as a multi-goal optimization approach where trade-offs must be made between different evaluation criteria and traffic metrics. Various solutions, both theoretical and empirical, have been explored to analyze the behavior of different traffic metrics and their dependencies (i.e. [13]) and to develop more precise measurements by unifying traffic performance metrics under a common formulation [14].

4. Emergent complexity

Complex systems theory defines a system as complex if some of the behaviors of the system are emergent properties of the interactions between the parts of the system, where the behaviors would not be predicted based on knowledge of the parts and their interactions alone.

From the perspective of complexity science there are a number of characteristics that characterize complex systems [15]. Based on these definitions, ITS can also be viewed as a complex system, based on the following observations [16]:

- **Semi-permeable boundaries**: The boundaries between the system of interest and its environment may be fluid and dynamic. For example, the ATLO system described above can be considered as a closed loop control system. However, to fully understand the impact of, and therefore potential risks associated with its behavior, it must be considered within the context of interactions with emergency services and city or highway infrastructure as well as public transport systems.

- **Non-linearity, mode transitions and tipping points:** The system may respond in different ways to similar input depending on its state or context. Small changes in the behavior of traffic participants or minor anomalies (e.g. temporary road works) may lead to rapid changes of state, exacerbated through coupled feedback between system components and road users. The seemingly spontaneous occurrence of traffic jams and stop-start traffic on motorways are examples of such behavior within traffic systems.

- **Self-organization and ad-hoc systems:** The adaption of the behavior of human road users in response to automated vehicles is an example of self-organization, where humans become part of a larger ad-hoc system. The ability of traffic to spontaneously respond to approaching emergency vehicles, even at complex intersections, is an example of ad-hoc self-organization. Self-organization and ad-hoc systems can also emerge because of unplanned integration of new systems with legacy components.

The consideration of ITS as a complex system leads to observations regarding our ability to reason about the trustworthiness of such systems. The emergent complexities of such systems and functions will have multiple layers of impact. For instance, changing the signal phases of traffic lights could have an impact on broader patterns of traffic flow, including how human actors plan their journey. Thus, automated decisions within complex system could lead to indirect and accidental behavior with unpredictable consequences. In general, there is a need to understand the behavior of the systems we regulate, but there will always be gaps in our understanding. How do we ensure that these gaps in understanding do not lead to unacceptable levels of risk? Instead of trying to control all aspects of the systems, it may therefore be better to allow for the positive opportunities of emergent properties to arise in a shepherded manner, thus ensuring that they do not lead to unacceptable risk. This will lead to a need to define and ensure quality of service guarantees that can be monitored in real-time despite the emergent complexity and unpredictability of the system.

The term "Semantic Gap" refers to the difficulty in expressing, often implicit, expectations on a system as a complete set of technical requirements. For autonomous, AI-based systems, the emergent complexity within the environment and the system itself, as well as the transfer of decisions previously made by human actors to the system can lead to semantic gaps [17] which make the task of defining suitable system behavior all the more difficult. Existing methods of iteratively deriving technical requirements from a precise definition of a set of safety goals for the system are unlikely to lead to a sufficiently complete understanding of the required behavior. Instead, an iterative process of reflecting system behavior onto the stakeholders of the system will be required to refine a set of technical criteria by which to measure the trustworthiness of the system. These stakeholders will include groups such as city planners, traffic planners, municipalities and city authorities, public transport providers, emergency services, logistic providers, traffic participants, standardization and regulation agencies, businesses in the city, technology suppliers, insurers, and further stakeholders. It is also to be expected that some stakeholders will inevitably experience negative effects whilst the system strives to achieve a global optimum. Approaches to communicate the overall utility of the system despite potential side-effects and the need to respect accepted definitions of human agency are therefore also required.

5. The impact of AI on uncertainty

Different manifestations of uncertainty threaten our ability to reason about the dependability (safety and reliability) of a system and therefore, it is of utmost importance to be able to identify their sources and apply appropriate mitigation approaches to reduce these uncertainties.

A system can be defined as open context if it operates within an environment which cannot be fully defined during design time. This can be due to either the inherent complexity and unpredictability of the environment or the way in which the environment evolves over time. ITS demonstrate properties of open context systems in which the environment they are attempting to control (traffic flow) is inherently complex (also in the sense described above), dependent on very many variables that cannot be precisely modelled (including human behavior and interaction with other systems) and evolves over time. In order to control such systems, multiple points of sensing are required and, increasingly, AI techniques are used to extract a suitable model of the environment from the unstructured sensor data and to derive suitable control strategies.

The issue of uncertainty is closely related to the topic of emergent complexity described above and can be defined in terms of a lack of knowledge about the system, its environment, and the impact of its actions. This inevitably leads to gaps in our ability to determine the trustworthiness of the system, for example, the creation of a complete and convincing safety assurance case. Generally, uncertainties within a system can be classified into one of the following categories:

- **Aleatoric uncertainty:** Inherent uncertainty associated with the randomness of a process. This could, for example, include the unpredictable nature of the traffic that is to be managed by the system, including unforeseeable events due to human behavior, accidents or unknown interactions with other systems.
- **Epistemic uncertainty:** Lack of precision regarding a process. This could include the limitations of a machine learning model to accurately forecast possible future states of the system due to deficiencies in its training data or modelling approach.
- **Ontological uncertainty [18]:** Complete unawareness of factors influencing a process (also known as deep uncertainty). This could include a lack of awareness of critical factors in the environment or system context that impacts the behavior of the system.

Definition of deep uncertainty

In these situations, experts do not know or the parties to a decision cannot agree upon (i) the external context of the system, (ii) how the system works and its boundaries and/or (iii) the outcome of interest from the system and/or their relative importance [19], [20].

Uncertainty can therefore be caused by the complexity and unpredictability of the environment in which the system operates (aleatoric and ontological uncertainty) but can also stem from the fact that the AI/ML components used to model the system provide only a limited approximation of the desired target function (epistemic uncertainty). Thus, we can distinguish these three categories of uncertainties within the complex intelligent system of systems. Furthermore, the Sense, Understand, Decide, Act (SUDA) model [20] can be used to identify the sources of uncertainty in complex intelligent systems. Figure 1 summarizes the sources of uncertainty and complexity in these systems based on the different components within the context of the ATLO example. As noted above, the system may consist of a hierarchy of subsystems each of which may follow its own SUDA model (e.g. local intersection traffic light switching).

Compared to traditional software, AI techniques and machine learning exhibit properties that lead to uncertainties in their calculations as well as difficulties in assuring their dependability. These can be summarized as follows:

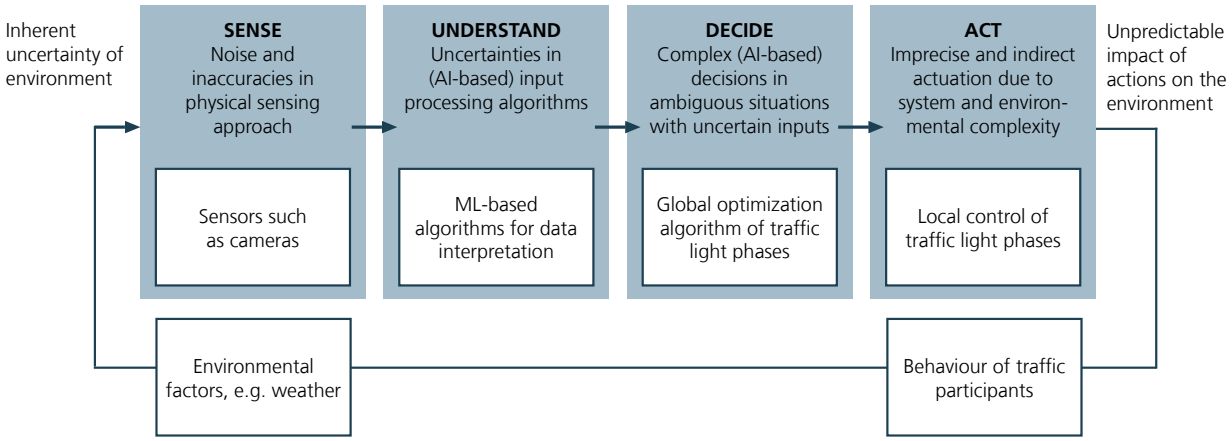

Figure 1: Sources of uncertainty in complex intelligent systems.

- **Generalization:** Generalization errors can be caused by degradation of accuracy when input data is not within the distribution of the training data or overfitting of the model to spurious correlations within the training data not related to the target function.

- **Robustness:** Errors caused by sensitivity to small perturbations in the input data. These perturbations can be naturally occurring [21] (due to aleatoric uncertainty in the inputs, e.g. sensor noise, weather conditions, etc.) or caused by adversarial attacks [22].

- **Unreliable confidence estimations:** Many machine learning algorithms provide a confidence score (as a value between 0 and 1) for each result. The confidence that the machine learning function indicates for a particular result does not necessarily match the actual probability that the result is correct [23].

- **Fairness:** Imbalanced training sets can lead to an unequal probability of errors between various semantic classes within the input space.

- **Explainability:** The decision making of the ML algorithms is often opaque, leading to a lack of explainability of the decisions leading to the results [24], [25]. The features from the data are extracted automatically and therefore, the causal relationship between the features and the output is unclear.

An additional challenge related to the use of AI and machine learning algorithms is the issue of semantic gaps [17] described above. AI, and in particular ML algorithms are often used to process unstructured data whose properties cannot be algorithmically described. As a consequence, it is often not possible to precisely describe the desirable dependability properties of the function, leading to a paradox where machine learning is used in place of specificied behavior but a specification of dependability properties is nevertheless required in order to gain trust.

These properties manifest themselves to different extents between various types of ML algorithms and application domains, e.g. depending on the dimensionality of the input data. However, in order to argue that the dependability properties of the function are fulfilled and the epistemic uncertainty introduced is within acceptable bounds there is a need to understand the causes of these properties, such that appropriate measures during development and test and can be applied and measures at the system level can be defined for limiting the propagation of the resulting uncertainties within the system.

What is the definition of a trustworthy AI-based system?

According to the EU Guidelines for Trustworthy AI, there are seven requirements that AI systems should meet in order to be deemed trustworthy [26]:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

6. Challenges in the introduction of Trustworthy AI-based ITS

The complexity and uncertainty properties of AI-based ITS lead to a multitude of factors that must be considered whilst developing, deploying and operating such systems. We have identified key challenges related to the introduction of Trustworthy AI-based ITS and grouped these into two main categories – those associated with the system and application as whole and those associated specifically with the use of AI.

System-related challenges

Challenge 1: How to define the scope of the system of interest, despite unclear system boundaries and unanticipated interactions with other systems? This includes the identification of all possible stakeholders that could be impacted by the system.

Challenge 2: How to define the system-level expectations that adequately cover the trustworthiness expectations of all stakeholders and balance the needs of all stakeholders whilst working towards a global optimum? A strategy is required to develop and maintain trust amongst stakeholders. This will include ensuring that human agency over the system is maintained. This will inevitably include the need for a level of transparency and explainability in the decisions made by the system to allow for human operators within the city authorities to determine whether or not the actions taken by the system are leading to desirable behavior or whether an intervention is required.

Challenge 3: Each city has its own set of unique properties and legacy systems. Solutions for one city might not work for others. Therefore, general standards and regulations will need to be tailored to the needs of the specific set of stakeholders for the city accordingly.

Challenge 4: How to determine the impact of uncertainty within the system and, in combination with the system-level requirements, derive specific performance criteria for the technical (AI-based) components in order to constrain the uncertainty caused by technical components of the system?

Challenge 5: How to define suitable acceptance criteria that confirm the performance of the system in terms of its direct impact on the observable traffic metrics to be optimized? How can the acceptance criteria on the system be confirmed based on multi-objective metrics and a variety of representative (quantity, quality, diversity) scenarios that can also reproduce disruptions or anomalies in the system?

Challenge 6: Which risks are inherent to the system and to what extent can tolerable levels of residual risk be quantitatively agreed upon, e.g. based on the direct impact on traffic metrics, and which arguments must be supplied to confirm this level of residual risk? How to evaluate the positive risk balance based on the contribution of the system to technical risk and its ability to reduce the inherent risk associated with the traffic conditions?

AI-related challenges

Challenge 7: At present, regulations are under development to ensure the trustworthiness of AI [26]. Nevertheless, these regulations will inevitably be generic in nature and therefore difficult to directly measure their fulfilment. There is therefore a need to map trustworthiness requirements, such as robustness, transparency, and accountability to measurable properties of the ML components in the system.

Challenge 8: The ability to manage the emergent risk associated with the use of AI is constrained by the initial understanding of the system and relevant interactions therein. The (possibly unintended) behavior of the AI-based system may also be dependent on the scope of the data used to train and test the systems with an unintentionally wider than required scope possibly leading to unwanted emergent behavior. In general, the performance of the AI-based systems will depend on the choice and availability of training data. In particular, negative effects on the robustness, generalization, fairness and prediction accuracy of the function must be minimized. Depending on the target function and ML techniques used, specific criteria will need to be defined to ensure that training and validation data lead to a function with properties required for a trustworthy system.

Challenge 9: Many AI algorithms and techniques exist and new approaches are continuously being proposed in this dynamic field of research. This makes it challenging to identify the "right tool for the job". The techniques vary not only in their absolute performance given any given task, but also in the nature of evidence that can be collected to argue their dependability properties.

Challenge 10: How to prevent the misuse of the massive amounts of data required to train and test the AI functions? This challenge must consider data storage and proper usage of data.

Challenge 11: Which measures can be applied during the design and operation to counteract the emergent uncertainty resulting from properties of the environment as well as the ML algorithms and the data themselves to ensure that a tolerable level of residual risk is met? This choice of risk control measures needs to be made based on a fundamental understanding of the causality relationships between sources of uncertainty and errors in the system and their impact on overall system dependability.

Questions to ask during system design

A key question that needs to be asked when deploying AI in ITS is therefore, for which tasks, with which impact, within which levels of the system is the AI function being deployed? Does the system allow for a fully automated decision making by a single AI component or are there other checks and balances in place? To what extent do human stakeholders retain agency over the system and its impact on its environment?

7. Preparing for the consequences

To prepare for the consequences of AI-based decision making in ITS, holistic and inter-disciplinary approaches will be required to derive reasonable expectations on complex intelligent systems. There is a need for a better understanding of the diversity of the public understanding and acceptance of risk associated with the deployment of such systems. This understanding should be used to formulate and calibrate regulations (see e.g. [26]) which determine the desirable properties of "trustworthy AI" and may restrict their use for some applications. Despite the legal, technical and ethical challenges associated with the systems, there must also be an understanding that AI opens opportunities that cannot be solved with traditional technology. The risks associated with the introduction of such systems must therefore be carefully balanced against their potential utility.

At the same time, the technical capabilities need to be developed in order to create trustworthy AI-based systems that fulfill properties of safety, cyber-security, explainability, fairness and robustness. This includes a set of V&V methods that confirm that such properties are met. The diverse perspectives of social acceptance of residual risk balanced against utility and the ability to evaluate and ensure critical technical properties of the system must be combined within an assurance process capable of convincing a wide range of stakeholders.

The role of standardization is expected to play a major role in the deployment of AI in ITS. Standardization can build the bridge and act as a translation between regulatory, ethical and contextual issues, to unfold and close the semantic gaps among stakeholders involved in the systems. It also enables interoperability between different components of systems. Hence, we require shared definitions of system boundaries and system stakeholders, defining quality and safety standards for AI in the context of the application scenarios which include definition of interfaces between systems, technical qualities, quantitative and qualitative testing and common criteria for ethics and safety on the use of AI. Harmonized standards would define common criteria for the safety assurance of the systems. However, there may be a need to adapt the standards according to specific needs of the city.

Inevitably, it cannot be reasonably expected that technically perfect AI-based ITS systems will ever be developed that are able to continuously meet all stakeholders' expectations. This leads to a number of important considerations:

- There will need to be a process of ongoing, continuous assurance, to evaluate the system's dependability with respect to changes in the environment, stakeholder expectations and emergent behavior.
- Systems will need to be designed to be resilient against yet unknown sources of failures and remaining uncertainties within the system.
- There is a need to ensure that humans maintain authority and control over the systems including the ability to evaluate whether the system is performing its task in a trustworthy manner.

8. Conclusions and next steps

Introducing AI into Smart Cities and ITS is a complex, heterogeneous, multi-disciplinary problem requiring collaboration and communication between a broad range of disciplines. This paper illustrates the need to consider the emerging complexity and inherent uncertainty related to such systems and identified key challenges associated with ensuring their trustworthiness.

Despite these challenges there are a few immediate steps that can already be taken. This includes greater collaboration with standardization bodies and Research and Development (R&D) within lighthouse projects to allow for a co-development between the safety assurance approaches and associated goal-oriented standardization and regulation. Forums (such as a series of peer-reviewed workshops) should also be established whereby members of different communities can participate in detailed discourse and work together to learn from experiences in different academic fields and domains. In addition, simulation-based experiments and case studies, with the associated data need to be made freely available in order to provide common benchmarks against which to compare the effectiveness of different solutions. Standardized scenarios could be defined to support this analysis thus reflecting the combination of technical specifications, reference implementations and conformance test successfully applied with telecommunications standardization initiatives.

After these initial immediate steps, there will be a need to increase the understanding of the general public for possibilities and limitations with the help of the tech community. This will include making use of the inter-disciplinary dialogue and existing standards to form consensus - establishing an iterative dialogue between engineering-informed ethical considerations and ethics-informed engineering. This should accompany a move towards system-thinking approaches to goal-based regulation that can keep track of the rapid technical advances expected in this field.

References

[1] S. Shayan, K. P. Kim, T. Ma, and T. H. D. Nguyen, 'The First Two Decades of Smart City Research from a Risk Perspective', presented at the Sustainability, 2020.

[2] M. Lacinák and J. Ristvej, 'Smart City, Safety and Security', presented at the Procedia Engineering, 2017.

[3] S. O. Johnsen, 'Risks, Safety and Security in the Ecosystem of Smart Cities', Risk Assessment, 2017.

[4] A. M. de Souza, C. A. Brennand, R. S. Yokoyama, E. A. Donato, E. R. Madeira, and L. A. Villas, 'Traffic management systems: A classification, review, challenges, and future perspectives', International Journal of Distributed Sensor Networks, 2017.

[5] A. J. Arlow, C. J. Duffy, and J. A. McDermid, 'Safety Specification of the Active Traffic Management Control System for English Motorways', presented at the 2006 1st IET International Conference on System Safety, 2006.

[6] 'Safer Complex Systems - Royal Academy of Engineering'. https://www.raeng.org.uk/global/international-partnerships/engineering-x/safer-complex-systems (accessed Mar. 12, 2021).

[7] F. Dembski, U. Wössner, M. Letzgus, M. Ruddat, and C. Yamu, 'Urban Digital Twins for Smart Cities and Citizens: The Case Study of Herrenberg, Germany', Sustainability, 2020.

[8] X. Liang, X. Du, G. Wang, and Z. Han, 'A Deep Reinforcement Learning Network for Traffic Light Cycle Control', IEEE Transactions on Vehicular Technology, 2019.

[9] I. Abu-Shawish, S. Ghunaim, M. Azzeh, and A. Nassif, 'Metaheuristic Techniques in Optimizing Traffic Control Lights: A Systematic Review', presented at the International Journal of Systems Applications, 2020.

[10] T. Wu et al., 'Multi-Agent Deep Reinforcement Learning for Urban Traffic Light Control in Vehicular Networks', IEEE Transactions on Vehicular Technology, 2020.

[11] C. S. Sánchez, A. Wieder, P. Sottovia, S. Bortoli, J. Baumbach, and C. Axenie, 'GANNSTER: Graph-augmented neural network spatio-temporal reasoner for traffic forecasting', in 5th Workshop on advanced analytics and learning on temporal data, 2020.

[12] D. Krajzewicz et al., 'Towards a unified evaluation of traffic light algorithms', presented at the 5th European conference on transport research arena, 2014.

[13] R. Blokpoel, J. Vreeswijk, D. Krajzewicz, and T. Kless, 'Unified evaluation method for traffic control algorithms', presented at the ITS World Congress, 2014.

[14] X. Li, G. Li, S.-S. Pang, X. Yang, and J. Tian, 'Signal timing of intersections using integrated optimization of traffic quality, emissions and fuel consumption: a note', Transportation Research Part D: Transport and Environment, 2004.

[15] P. Erdi, Complexity explained. Springer Science & Business Media, 2007.

[16] S. Burton, J. A. McDermid, P. Garnett, and R. Weaver, 'Safety, Complexity, and Automated Driving: Holistic Perspectives on Safety Assurance', Computer , 2021.

[17] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, 'Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective', Artificial Intelligence, 2020.

[18] R. Gansch and A. Adee, 'System Theoretic View on Uncertainties', presented at the 2020 Design, Automation Test in Europe Conference Exhibition, 2020.

[19] R. J. Lempert, S. W. Popper, and S. C. Bankes, Shaping the next one hundred years: new methods for quantitative, long-term policy analysis. CA: RAND, 2003.

[20] V. A. W. J. Marchau, W. E. Walker, P. J. T. M. Bloemen, and S. W. Popper, Decision Making under Deep Uncertainty: From Theory to Practice. Springer, 2019.

[21] D. Hendrycks and T. Dietterich, 'Benchmarking Neural Network Robustness to Common Corruptions and Perturbations', arXiv:1903.12261 [cs, stat], 2019.

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy, 'Explaining and Harnessing Adversarial Examples', arXiv:1412.6572 [cs, stat], 2015.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, 'On Calibration of Modern Neural Networks', arXiv:1706.04599 [cs], 2017.

[24] G. Schwalbe et al., 'Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications', in Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops, 2020.

[25] O. Willers, S. Sudholt, S. Raafatnia, and S. Abrecht, 'Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks', in Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops, 2020.

[26] European Commission. Joint Research Centre., Robustness and explainability of Artificial Intelligence: from technical to policy solutions. Publications Office, 2020.

Imprint

Fraunhofer Institute for Cognitive Systems IKS
Hansastraße 32
80686 Munich

Authors

Huawei Research Center, Munich
Dr. Stefano Bortoli, HUAWEI TECHNOLOGIES Duesseldorf GmbH, stefano.bortoli@huawei.com
Dr. Margherita Grossi, HUAWEI TECHNOLOGIES Duesseldorf GmbH, margherita.grossi@huawei.com

Fraunhofer IKS
Prof. Dr. Simon Buton, Fraunhofer IKS, simon.burton@iks.fraunhofer.de
Marta Grobelna, Fraunhofer IKS, marta.grobelna@iks.fraunhofer.de
Yuki Hagawara, Fraunhofer IKS, yuki.hagawara@iks.fraunhofer.de
Philipp Schleiß, Fraunhofer IKS, philipp.schleiss@iks.fraunhofer.de

Image credits
istock.com/NateHovee, istock.com/simonkr, istock.com/MarcelConrad, istock.com/hanohiki, istock.com/ilbusca

Acknowledgements

We would like to thank the contributors to the safety round table held on 01.07.2021 for the valuable discussions which formed the basis of this document:

Prof. John McDermid, University of York

Dr. Philip Garnett, University of York

Dr. Claus Bahlmann, Siemens Mobility

Fabrizio Arneodo, 5T

Stefano Cianchini, Municipality Turin

Dr.-Ing. Rasmus Adler, Fraunhofer IESE

Dr. Götz Brasche, Huawei

Liming Lu, Huawei

Kontakt

Prof. Dr. Simon Burton
Safey Assurance
Tel. +49 89 547088-700
simon.burton@iks.fraunhofer.de

Fraunhofer Institute for Cognitive Systems IKS
Hansastrasse 32
80686 Munich
www.iks.fraunhofer.de