Explainable AI for Transparent Decision-Making in Cybersecurity Operations for Autonomous Vehicles

By Dr. Amira El-Shafei

Associate Professor of Computer Science, Ain Shams University, Egypt

1. Introduction

Connected autonomous vehicles (CAVs), integrating connected vehicles and autonomous vehicles, demonstrate a use case of how advanced simulations and machine learning were combined to improve the decision-making capabilities for a holistic system CO3. CAVs are influenced by results obtained from adversarial research done on common sensor modalities (radar, camera and Lidar) integrated into them to take decisions related to perception and fusion. The output of perception acts as an input to control decision-making algorithms governing the movement of the CAV. Vulnerabilities have been identified in each common sensor modality which can change the decision-making of the algorithm [1]. The top three modalities (camera, radar, lidar) used in self-driving vehicles are shown to be inter-dependent with correlation analysis, the same information can be used to fool the system to trigger misclassification by both adding different real-word noise. Different factors that affect perception and decision-making were individually tested and demonstrated to affect the decision of the CAV. Videos of adversarial attacks test on each modality are shown to fool the algorithms by altering objects in a test environment such as different holograms or labelling them as different objects, creating realistic models of the road environment that are falsely identified etc. The different approaches were shown to effectively change the decision of the algorithms and thus the simulated behaviour of the CAV from stopping to moving or vice-versa. The results of the paper demonstrate the vulnerability of the testing algorithms and sensors to adversarial manipulative attacks. The effect of adversarial machine learning (AM) based attacks on the decision of a Control Decision Making Module for Connected Autonomous Vehicles (CAVs) is articulated in the paper as an example of what is common in cyber-physical systems today; to provide model user-centric information to users, critical malfunctions need to be addressed along with their interaction too as part of abductive reasoning. Recommend actions in simplest to complex domains are given for control systems

This work is licensed under CC BY-NC-SA 4.0.

including evasion generated by using Perception or sensing to evade on decision-making sensitivity, behavioural eviction, stability in state domain affected by rearranged adversarial actions, prediction affected by adversarial actions etc [2]. Interpolated state and mode dependent correlations are used to interpret multi-modal adversarial manipulation of CAV sensors for three common real-world object information. Smart system query function has no guarantee in diagnosis and critical state identification, further we propose, Abduction based Critical State Identification to identify the explain truth value given Substitute malfunctioning result.

Recent technological advancements in the area of Artificial Intelligence (AI) such as machine learning and deep learning approaches have demonstrated high-level autonomy and the potential to change the way algorithms based on explainable AI (XAI)derived decisions through their outcomes are integrated within systems. The level of autonomy in a complex system like connected autonomous vehicles (CAVs) is increasing, as is the potential impact of algorithms on their operational decision-making. However, as a consequence of the evolving nature of performance of CAVs, governed by mutually dependent human and technical systems, the notion of verifyability and understanding of AI-derived decisions become critical, especially for safety assurance [3]. In this article, the challenges in using Machine Learning (ML) based decision-making in cyber-physical systems (CPS) like autonomous vehicles (AVs) are discussed from the perspective of related applications in cyber-security, where resources implementing adversarial behaviour are increasingly making use of ML algorithms and evasion under changes in the input, to artificially change the decision of an algorithm.

1.1. Background and Significance

AI applications in general have witnessed many successes, but they are mostly treated as a black box. Low human interpretability makes it difficult to understand and verify such models. [3] Researchers working on XAI has communications that are common in the field, for example, they talk about improving post hoc model visualization and validation, human-interpretable settings for standard machine learning algorithms, human-in-the-loop design of model explanations, common brainstorming of diverse case studies and confounding manipulations, model explainer service frameworks, and much more. All such improvements

This work is licensed under CC BY-NC-SA 4.0.

can raise the bar for AI models to avoid, or at least proactively help remedy, unintended negative effects on human cognition, motivation, and behavior.

Explainable artificial intelligence (XAI) is a recent and inter-disciplinary research trend that has largely originated within the confines of the artificial intelligence (AI), machine learning, and natural language processing (NLP fields. AI models for cybersecurity are often treated as a black box, and currently, it is difficult to interpret and understand the decision-making process. Several interpretability, explainability, and transparency tools have been introduced by the AI community to enhance user experience in terms of understanding the reasons behind AI decisions. [4] In general and specifically for cybersecurity, XAI serves multiple purposes; it helps domain experts' understanding trends by providing more robust, interpretable, transparent, and precise model explanations. To the best of our knowledge, only a few review articles have touched on the application of XAI methods to cybersecurity domains.

1.2. Research Objectives

To enable human decision-makers to understand and trust the deployed AIs, the AI components must be transparent and explainable. Further, a user's knowledge structures understanding of the AI component, the task concepts, and the internal structure and the decision-making be informed by these knowledge structures. Along this continuum of user understanding, the explanations might need to convey different types of information and might be generated using different methodologies for generating explanations [5]. To make the AI environment even more transparent and explainable, XAI employs human-understandable concepts, which makes analysis easier and more effective. Further, an interactive XAI system should be visually interpretative for good problem-solving performance. The visualization is particularly important in domains like cybersecurity and autonomous vehicles where the consequences of a model decision can be dire.

Explainable AI (XAI) focuses on the design of AI models (e.g., neural networks and ensembles of classifiers) that can help human experts/users understand their decisions [6]. User interaction guarantees that results are not only comprehensible but also satisfactory from a domain perspective. This process of involving human judgment has many benefits, such as allowing for the ability to involve knowledge-based reasoning directly and indirectly, avoiding data-driven solutions, ensuring that the system is efficient, flexible and reliable,

This work is licensed under CC BY-NC-SA 4.0.

reducing the underlying data volume, and the generally improved performance and acceptance of the system. In the context of AI-based decision support for critical applications and human–AI task collaboration in various fields, the explanation of the decision-making process and the reasons behind the provided output outcomes are crucial [7].

2. Foundations of AI in Cybersecurity

This current outlook paper has refrained from putting emphasis on some ethical, philosophical, and social concerns related to AI and AV (Katzourakis, 2022; Baum and Bombaro, 2019; Teufl et al., 2020; Lagorio et al., 2020; Borys, 2021; Romn-Rodriguez et al., 2021; Wu and Zhou, 2022) simply because a clear picture of ethical considerations is also present in the rapidly evolving guidelines for the ethics of machine learning systems which could in principle be viewed as the focus of another future document (Susha and Belli, 2017; Kumar et al., 2019; McRae et al., 2020; Hain et al., 2020; Wu et al., 2021; Lee et al., 2021; Antonelli et al., 2021). Ethical considerations are relevant and are discussed as external factors of the system to demonstrate respect of relevant project impact, see (Peglow et al., 2022; Correia et al., 2020), and references therein to show the aptness of the used technology to generate an explanation for a given decision made by the DNN (Jha and Mudliar, 2019) and to meet standard guidelines for protecting sensitive data like complying with the General Data Protection Regulation (GDPR), see Aggarwal and Correia (2022).

This review is based on several articles and books as well as contributions and work of the NanoSec Project Conrad et al., 2022, [Link] d S.Garcia-Galan et al. Some elements come from the Deliverable D2.6.1 of the NanoSec H2020 project, which is on architecture and strategies for detection and response for the Internet of Things and Autonomous Vehicles, available at [Link] In view of its significance, the avoidance of the total dependence on dangerous AI algorithms is framed as an abiding principle in human-robot interaction that sets out significant ethical and scientific concerns with AI and explains the human-in-theloop framework in which this issue is addressed (Hall and Lemaignan, 2020). It was only after a long discussion about the appropriateness of deploying black-box models in safety-critical situations, in the absence of comprehensive interpretability, that a consensual decision-making principle was offered (Lagorio et al., 2019). Thus, machine learning for autonomous vehicles (Jung et al., 2020) is treated from the experimentally estimable model to the classical

This work is licensed under CC BY-NC-SA 4.0.

data-driven view, focusing on the previ obtained when learning to drive inside a simulated environment, with semantic segmentation pertaining images as the gold standard.

2.1. Machine Learning and Deep Learning Basics

The ELCS project had a major goal to improve the trust in the AI models and algorithms and to contribute to decrease their black-box nature improving their transparency and consequently their applicability. The following three main aspects were focused on: (i) The development in a self-stabilized and supervised intelligent security solution. (ii) Developing an improved, minimal, and safe compromise machine-learning system for efficient features extraction and data analysis, as well as an optimization method through multi-view/multi-modal analytics, DL models, and multi-task learning for generation and selection of the best features for the construction of a secure AI-based system. Due to the EICS approach (explainable intelligent computer system) development, saw the possibility of black-box nature minimization by the extraction of rational information. Therefore, once the system is working an explainable interface is proposed by the connection between explainable inferential multi-modal analytics (EIMA) and Real Analytics (RA) and Deep learning (DL) vision [8].

For cybersecurity, the new technologies in use are machine learning (ML) algorithms as they are able to implement a trustable, resilient and secure infrastructure through intelligent analysis and processing. Be it deep learning (DL) models in different fields, multi-task learning (a statistical learning strategy in which multiple tasks are solved jointly), or multi-view/multi-modal analytics (algorithms combining data from various sources/models and modalities), and other ML approaches can contribute to increase the level of 'explainable AI' and to improve risk management and security posture in the current global-serving systems and solutions. Regardless of the available and commonly used neural and non-neural learning strategies, the main weakness in the security-oriented models refers to the black-box nature of DL, which do not allow for rational arguments and explanations of their decision actions. Well-designed networks always work better than other types of classical server pulling-databases, as well as expert systems and simple off the shelf model architectures [9].

2.2. Explainable AI Techniques

This work is licensed under CC BY-NC-SA 4.0.

Algorithmic explanations may be used to annotate models with human understandable and meaningful summarization of the deployed AI, e.g., explaining the AI proof. This strategy may help human end "AI user" decision-makers to "trust" the AI-understood decision, based on the Rochester survey in. This is actually an approach that we are pursuing in our previous research work. To improve interpretability, the AI model, however, should be open enough to support a meaningful extraction of an explanation, i.e., to be complex enough and to take into account unpredicted events or nonlinear time evolution, interpretability can simply be underestimated. Therefore, we ensure to develop XAI interpretability in both ex post and in vivo mechanisms even with adverse conditions, in different cybersecurity and adversarial AI scenarios that we are going to explore already by the SR explanation packet in the next paragraph.

The global trend toward complex AI systems and Deep Learning (DL) methods has led to the realization of algorithms bicameral, like a human brain, without a transparent reasoning. Since their recall of patterns or knowledge is often non-trivial to interpret by humans beyond a dataset, understanding and thus trust in their predictions are essentially limited [10]. In this respect, the AI community has addressed the issue, and numerous research and development efforts have materialized in the exploration and design of Explainable AI (XAI) techniques [5]. XAI represents an AI transparent by design, i.e., AI that is naturally explainable, as it satisfies the White-Box principle, where the internal mechanisms of the AI result understandable by human beings in terms of features, decision process, and impact [11]. In several real-world applications, like autonomous vehicles, AI provides safety-critical decision-making, therefore, trust and confidence in AI are significantly crucial. Thus, the need for understanding and psychologically ease the reasoning of the AI has been visible used in XAI areas, even with the enhancement of XAI tailored for preventing any possible adversarial AI attack (particularly to autonomous vehicles). In this setting, we focus our study on the application of transparent AI solutions for cybersecurity in autonomous vehicles. Consequently, AI fail and potential bias (lack of robustness and adversarial vulnerability) can be timely countered and foreseen avoiding AI accidents.

3. Autonomous Vehicles and Cybersecurity

Preliminary contributions include an overview of a collaborative study from industry, academia, and major automotive stakeholders, aimed at providing domain-specific

This work is licensed under CC BY-NC-SA 4.0.

challenges and guidelines for introducing and regulating XAI for AV cybersecurity in public roads. Furthermore, this article explores unique Security and Safety gaps and transdisciplinary concerns arising in national and international regulatory infrastructures. Advanced mobility solutions will enable deployable AVs in our interconnected mobility networks, which will necessitate complete redefinitions of transportation regulations. New paradigms for XAI and safety regulation should ought to commence, integrating lessons from past technology innovation 'waves' to create societal rules and protections extending to all citizens.

[12] [13]Autonomous vehicles (AVs) have entered the public consciousness in recent years as a plausible commercial transportation option for the future. With current mainstream technology developments and potential widespread adoption, cybersecurity and privacy threats are salient concerns. For AV cybersecurity, Explainable Artificial Intelligence (XAI) has been proposed as a means to improve attack detection reliability, transparently guide AV operations, and provide explanations for AV actions to improve trust. However, to enable XAI on critical systems such as AVs, there are concerns and challenges related to functional safety, timing predictability, and certification through regulation. Thus, we highlight key considerations for AV cybersecurity and XAI to commence meaningful interdisciplinary discussions on regulating this field [].

3.1. Overview of Autonomous Vehicle Technology

Whilst the performance of machine learning has rapidly increased in such tasks as image recognition or game playing, transparency—the ability to describe how these decisions were reached—often takes a backseat. For autonomous vehicles to be widely deployed, the industry must develop explainable AI (XAI) so that end-users can trust decisions made by the vehicle's artificial decision-making systems [14]. Fortunately, the immediate environmental-sensing nature of autonomous vehicle technology could also make it an interesting entry point for transparency in decision making. The inductive nature of classical AI and the reductive nature of traditional cybersecurity models allow for leveraging logical or physical dependencies between inputs and outputs to explain decisions from an electromagnetic, environmental, logical or safety point of view [15]. However, the vulnerabilities open to exploits of these Always-Aware vehicles must be explored in an age of mass Open Source Intelligence

This work is licensed under CC BY-NC-SA 4.0.

detection and high-quality Mobile Artificial Intelligence (Ai) assistance, leading to a discussion in decision vulnerability domain.

The widespread adoption of autonomous vehicle technology has the potential to radically reshape a range of industries, including transportation, public safety, and emergency health services. Today, many organizations—particularly those providing transport services—are making significant investments in developing this technology. To ensure their widespread adoption, stakeholders must address the challenges of both operational and decisional transparency. This form of transparency, most commonly known as explainability, is essential to gaining consumer and regulatory approval for this technology.

3.2. Cybersecurity Threats and Vulnerabilities in Autonomous Vehicles

To address these issues, the concept of Explainable AI (XAI) has been developed to provide insurance, traceability, and transparency and thus improve security throughout various processing stages of XAI. Based on this architecture, the cybersecurity framework for AVs for the detection of cybersecurity threats and vulnerabilities can provide a security safety net in case of attack. With the goal of achieving a secure and transparent automated process, it is crucial to make the decision-making system of an automated vehicle traceable. An XAI system contributes by allowing for the verification and understanding of the relations between situational factors and related decisions. Cyber threats and typical XAI processes in AVs, like preprocessing, responsibility analysis, and visualization, are explained separately along with possible countermeasures [16].

Modern automobiles are equipped with multimedia technology, connected services, and autonomous features. The increasing digitization of automotive functions results in a higher number of security-critical components and a more complex attack surface. Particularly in the context of autonomous driving, cybersecurity is a crucial factor, as a compromised autonomous vehicle (AV) can potentially result in a loss of safety diligence [15].

4. The Need for Transparency in Decision-Making

When AI decision-making is made more transparent in cybersecurity tasks, this has implications for the accountability of the deployed systems with respect to end-users, law and regulations, for independent public assessment of actions, and for risk insurance. Here we present an accessible guide to the societal relevance of AI transparency and accountability,

This work is licensed under CC BY-NC-SA 4.0.

with empirical examples demonstrating why these issues lead to trust in a cybersecurity context. In sum, beginning with the problem and its historical and timeless basis we lay the groundwork so that the cybersecurity community may strengthen its ability to trust both itself and outside observers. Subsequently, we demonstrate the potential link between mutual trust and transparency and accountability using the examples of explaining black-box decisions with traditional machine-learning classifiers, and of using AI for JSON Web Token encoding and user-agent string identification.

[4] [17]Determining who made what decisions is decisive for transparency and accountability, often leading to the phenomenon of 'opportunistic interpretation': the latter does not consist of consciously attributing values for those purposes, but rather emerges blindly from the simple social game. We refer to this as 'opportunistic interpretation by social attribution', which is different from, but definitely not incompatible with, opportunistic interpretation in the narrow sense. Explainability is a strategy for AI to signal where, rather than how, decisions are made in order to build human trust, and more generally to situate itself in the 'ethical space' opened by transparency. Explainability is a means to govern decision-making with AI and reveal it to us, proposing a way for us to establish the shared reality that forms the basis for trust or accountability. In concentrating decision attribution from joint into separable and imputable causes, explainability also reaches the heart of the problem of responsibility imputation by creating a clear divide between human and non-human causes.

4.1. Importance of Explainability in AI Systems

AI systems used in AVs must be explainable. AI systems are fast-growing technology particularly in the context of AVs, and cybersecurity operations in AVs in particular [18]. This booming technology has the promise to revolutionize the modern world, especially by the time when fully autonomous vehicles will be a reality. However, the trustworthiness of AI algorithms is still a challenging problem in both the fields of AI and AVs. According to a 2020 study by Deloitte, the vast majority (63%) of AI professionals concede there are understandable limitations or human-audiences that should not interact with AI at all in its current, pure form. This is why the work of understanding and improving the explainability of AI is becoming increasingly important in both fields [4].

Explainable AI (XAI) refers to an AI model's ability to provide explanations as to why it made a certain decision, in a manner comprehensible to human beings. This is essential for ensuring

This work is licensed under CC BY-NC-SA 4.0.

both transparency and the trustworthiness of the AI system. Human acceptance of XAI-based decisions is significantly influenced if the decision-making process is transparent and understandable [12]. As a matter of fact, one of the main components of trust in intelligent transport systems in general and AI-based systems in particular is the visibility or transparency as to how a particular decision has been made. In the context of Automated vehicles (AVs), the process of arriving at a decision consists of several steps, such as perception, localization, mapping, planning, reasoning, and control.

4.2. Challenges in Achieving Transparency

Since autonomous vehicles use several AI mechanisms, the request for transparency also comes from them. The main way to achieve transparency from AI is to search for understandable AI. Explainable AI is a common phrase that is mainly used to be identical with transparent AI. Also, AI safety and ethics is a rapidly growing area of research. One of the ethical issues concerning it is the need for system transparency [19]. From the point of view of AI ethics, transparency here means the ability to interpret the inner mechanism of AI that is used around us. But ethical challenges surrounding explainable AI are complex, and there isn't necessarily a straightforward answer.

Although effectively explaining AI systems can be a highly complicated issue [20], it is necessary to resolve this issue given the ethical responsibilities of AI developers and operators. In many cases, AI is viewed as a black box with a complicated inner mechanism, and they may act in a counterintuitive manner, making the trust of potential users, including clients and the public, extremely delicate [21]. It is generally accepted that we need transparent AI to re-build the trust that engineering AI originally broke. Transparency could be described as a process that makes the inner mechanism of an AI visible as much as possible.

5. Explainable AI Approaches in Cybersecurity

A novel approach is presented where AI models are designed to be interpretable by humans [22]. This is inspired by the case of autonomous transportation, where cars and trucks are expected to have some explainable and transparent predictive capacity—more stringent even than human drivers in offering predictive responses. Specifically, this required predictability and trust in automatic decisions become instrumental when autonomous vehicles operate in open upredictable environments. In particular, the article provides practical motivation for

This work is licensed under CC BY-NC-SA 4.0.

Explainable Artificial Intelligence (XAI) in the context of autonomous transportation. This involves adversarial deep leaning (for attacks on AI systems used in autonomous vehicles), evolutionary robotics (for the continual adaptation of vehicles to environmental changes), and cross-modal explanations (for the fusion of largely independent streams in multimodal transportation contexts). The upshot here is that different modes of explainability are appropriate for different settings in transportation-related AI decision-making processes. The reliability, trustability, and transparency of AI models become particularly important when we cope not only with autonomy integration in particular wearable devices or robots but also when we need to address truly open and critical settings such as in autonomous vehicle operations [1]. The first XAI setting introduces an anomaly in light of the fact that we would favor the most transparent and explainable decisions by AI systems in general—and in a critical open setting, such as autonomous transportation in particular—whereas the second setting would highlight ducked and duller protective capacities of AI in such settings, where rooms for technical ambivalence remain. It would appear that usable AI systems oftentimes already demonstrate these proactive (anticipative) and interactive (reflective) explanatory abilities that remain to become fully fleshed out in a generic and intensive manner. Then, by canvassing and compiling pertinent literature resources, this article views the requirements from various perspectives, detailing a novel set of XAI solutions at the interface between visual perception and decision making in autonomous vehicles.

5.1. Rule-Based Systems

It should be also noted that there is also an increasing research interest on the importance of domain-specific explanations in AI-based cybersecurity systems. In this article we focus on providing a human-readable explanation that is either from the point of human/vehicle system interaction (determining if the decision can be 'unobservable' or 'observed') or the cyber complexity behind the attack prediction (why a decision is valid). If a decision is observable, it can be either a UVNC where a User or Automaker are informed about the ongoing or potential attack or an OVNC where the Vehicle System itself is aware of the cyber manoeuvres. Nevertheless, we do not aim to generate morally implicated explanation structures besides considering data sought by users for decision making. We use the human/vehicle system interaction as well as cyber complexity as the real-world examples, on seeked data by users. Further examples on different target groups also might be developed and used as context to previous research.

This work is licensed under CC BY-NC-SA 4.0.

Using rule-based explanations, explanations are provided by uncovering and presenting the : • Relevant system parameters and their specified values and their values used in the query of the decision process as the top-level explanations. • Combined clustering and Matryoshka technique can be used to present the explanations at different hierarchical levels when different thresholds are crossed. • In the case of rule breaking, the VNC system can list the violated system specification as a type of explanation. • We can also provide the top-level feature based explanation based. This form of presentation is in contrast to the explanations of unconstrained AI-based solutions where one would provide a list of all states/features (thus irrelevant ones as well) and their associated weights and biases as an input feature(s) as explanation. Another characteristic of rule-based explanations is that they are, in all practical terms, time deterministic. That is within restricted time frames, the rules used for decision making are time-invariant. This is contrast to explanations obtained fro dynamic AI-based systems where the quality of explanations can depend on the current state of the system and the underlying design, if not for the hash-based or extremely rare patterns, these explanations are in general time-variant [23].

There are different ways AI systems can be engineered to explain their decisions. In the rule-based approach, decisions are based on a set of rules that humans can easily understand. These rules can be extracted from model-based AI solutions and are then presented to the user [7]. For the transparent AI-based cybersecurity system, currently we are using traditional rule-based knowledge-driven AI systems. Take as example a scenario where a Vehicle Networking and Cybersecurity (VNC) System has detected a potential cyber attack. The VNC system explains its decision based on welldefined handcrafted rules that are easily understandable for the user. For example, VNC explains its decision by stating 'User imposes zombie production request'.

5.2. Local Interpretable Model-Agnostic Explanations (LIME)

It is considerably significant to understand the decision-making process of deep learning applications in diverse fields. Local Interpretable Model-Agnostic Explanations (LIME) is an influential explanation approach being able to provide interpretability to opaque machine learning models. This method operates at the data level to offer explanations for the decision-making process of black-box models. Since LIME is attribute to a local surrogate model for the purpose of understanding the effects brought by the local neighborhood, LIME is quite

This work is licensed under CC BY-NC-SA 4.0.

sensitive to the local neighborhood data points. As a result, it might not be able to generate proper explanations in certain situations. [24] In IEEE Access 6 (2018), p. 13298-13313, the original formulation of LIME-axis that LIME-axis improves the classification performance in the car stopping time task. In Proceedings - 2017 IEEE International Conference on Software Quality, Reliability and Security, QRS 2017, Institute of Electrical and Electronics Engineers Inc., 2018, p. 387-393, LIME is selected as an experiment target to evaluate [25] PhilaeX.

A local surrogate model is trained to approximate the target classification function by using [26] Local Interpretable Model-Agnostic Explanations (LIME) to generate interpretable explanations. However, there are certain limitations of LIME, such as LIME can only support binary classification and multi-class classification tasks, and it cannot be directly applied to handle the regression problem. In this paper, an optimized approach is proposed to refine the explanation results obtained by LIME using the uncertainty-based strategy. The experiment results demonstrate that our approach can improve the contributed weights from the positive sampling data and is less sensitive to the information perturbated by the local neighborhood. Moreover, it can accurately and efficiently reflect the influences brought by different local neighborhood data samples.

6. Case Studies and Applications

The need for explainable AI (XAI) has been emphasized in [2], suggesting that the challenge of XAI in AI-based systems is to balance the inherent complexity and explainability associated with these systems. We present five common AI concepts employed in autonomous vehicles and their corresponding case studies. First, we consider that autonomous vehicles will rely heavily on AI-based systems dedicated to processing data from various sensors used in the control and operations of both the vehicle and the sensor. Then, we consider AI systems dedicated to providing communications connectivity and managing interference from various RF jammers and other communications attacks. For this case, we consider an AI system that supports each vehicle in predicting a future interference scenario, to switch when necessary, to predict the tactics of adversaries, and to ensure the recovery of communications. Next, we consider AI systems dedicated to the protection of both the vehicle and the passengers inside, to be able not only to detect an adversary but also to resist any motors, including ramming, that may be launched by this adversary. Then, we consider AI systems dedicated to personalized traffic jam problem resolution for each passenger using different objective

This work is licensed under CC BY-NC-SA 4.0.

functions such as travel time, energy, comfort, etc. Finally, we consider an AI anti-spoofing system specialized in continuously determining the quality of the video and image streams in each sensor and detecting, predicting, and protecting against different types of spoofing attacks. We seek to address these five use cases of AI systems dedicated to mission-critical systems in autonomous vehicles by adding more intelligence to them through XAI. This AI capability aims mainly to increase the transparency of such distributed AI systems, thus improving their direct technologies, immeasurable robustness, and predictability. Additional metrics for such indirect metrics, such as passenger safety, well-being, quality of life, and human ethical concerns, could be motivated by using XAI capabilities. The update of railway level crossing is about automated AI in terms of safety, predictability, and increased use for passengers. The business prospects go further with growing safety, cost reduction, and high-speed business in non-shared routes, requiring AI-centric external information and internal management. The additional business to provide internal facilities, the containment of ecological angles, should have exceptional transparency. In other applications and human-robot teamwork use cases of a similar autonomous vehicle, human consequences must be considered as the consequences of AI decisions also vary. Our examples of AI use cases focused on autonomous vehicles aim to provide high-level guidance for decision-making processes and XAI solutions across AI operations, both in centralized and decentralized AI systems running for safety-critical applications. [3]

6.1. Real-World Examples of Explainable AI in Cybersecurity Operations for Autonomous Vehicles

Most of the conversations and analyses on autopilots have neglected an outlier analysis, emphasizing the inability of explaining (for) abnormal or "violent" user behavior, which is essential for training. In March 2021, a complete new Tesla Model 3 (M3) headsets to travel from coast to coast in the one-step process (CA, US) using the AutoPilot drive (AP), as fully autonomous driving of the vehicle in other. However, it drove an entire CP after picking up some goods from the launch of the store's parking light ramp. In China, ROBORIXBRAVE achieved full marathon and achieved full grade by using hardware and software power, including full-tense management and recharging. An autopilot "bitch" hardware chip based on FMOD-SOC, Expert Board, and MFG software platforms is private.

This work is licensed under CC BY-NC-SA 4.0.

[12] [27]In-vitro experiments were compared to real-world examples such as autonomous driving software within the cyber-physical systems. The commercial autonomous systems are facing roadblocks because the decision intelligence in such systems is often not explainable and there is a need for basics of explainability in such cyber-physical systems. One such autonomous vehicle system is Tesla FSD, which has been driving unsupervised through cities for an extended period; swapping a human-in-the-loop (HITL) with an AI-in-the-loop (AITL).In 2019, Elon stated that "Tesla vehicles will soon talk to people if you want. This is real."Tesla engineers later implemented a Tesla honk gesture that is an explicit and spontaneous method of communication between the autonomous system and the human driver. Following the 2020 introduction of Explainable AI (XAI) visualization tools, the next logical step is for Tesla to incorporate public trunk drivers into the FSD road testing by enabling XAIC to produce real-time explanations.

7. Evaluation Metrics and Performance Assessment

A new type of ETR on vehicle features was tested using an AI-enhanced assessment designed to increase traffic safety and to enable a close link between human and automated vehicle driving and risk measures to control the operations. This new approach to the safety target of the Automated Vehicle (AV) is tested with the goal of always reaching safety. Beyond reliability, accuracy, precision, and other commonly used performance metrics, Explainable AI enhanced risk was tested considering the cases of False Threat Associated to Future Collisions (FTAFC) and False Threat Associated to Current Collisions (FTACC) [28]. Evaluated risk proved to possess a high capability to mirror both the expected evolution of the risk with the same output and the overall dynamic correlation with the parameters. While vehicle risk is mainly influenced by the velocity in case of FTACC, the noise voltage is the main issue when handling a FTAFC event.

A Cyber-Physical System (CPS) is a system spread in the physical world, which consists of computing capabilities and actuators that control the behavior of the physical world components [29]. This work introduces methods and algorithms to design automated vehicles that belong to this category and focus on explaining the decision-making models in Autonomous Systems (AS). Among the main paradigms to achieve explainability, introspective methods and post-hoc methods are included. The development focuses on road

This work is licensed under CC BY-NC-SA 4.0.

vehicle applications and, as a specific application, to offer a tool for the explainable adaptive Road traffic lane detection algorithm described in [15].

7.1. Interpretable Metrics for Model Performance

Performance metrics such as false positive rates for the detection of invalid inputs are overly simplistic when considering systems that make dynamic sequential decisions. This is particularly the case when their focus is failure detection, i.e. minimising the consequences of sub-optimal system performance and improving overall system transparency. This article proposes a novel set of context-dependent metrics for evaluating adversarial robustness performance in real-world AGI domains is proposed, which allows decision-makers to better understand the trade-offs between adversarial robustness and test-time performance [30]. Due to examples that can be seen as a universal attack, which works regardless of the policy in place we will concentrate on that kind of attack. It is worth noting that the robustness and adversarial robustness problem is more general than the human perception model. However, for the task of designing a system capable of taking in sensor information from the real world and making decisions then an attempted universal attack model can still give a focus on some of the challenges in achieving robustness.

Evaluating the trustworthiness of neural network models in context-specific domains is a challenge in the research community [17]. Evaluation of how robust a model is and assessing how it transitions between modes plays a big role in this context [31]. Metrics for evaluating robustness have been developed for a wide range of scenarios to hit this note, however, often these metrics are not context-dependent. This is often not helpful in making decisions or can provide misleading interpretations of system performance. In the domain of demonstrating the robustness of machine learning models to perturbations of their input data, common metrics are outperformed by domain-specific metrics that aim to capture a more true sense of system robustness. This is because intuitions for robustness are not necessarily intuitions for robustness for a decision-making system in real applications, which is often the case for machine learning models. Analysis of adversarial robustness, for instance, fails to capture some fundamental insights about adversarial attacks in the automotive domain. Adversarial robustness measures the robustness of a predictive model only with respect to crafted adversaries and not with respect to errors that can be obtained in the real world. This reduces

This work is licensed under CC BY-NC-SA 4.0.

the practical real-world impact that such measures have for decision-making in the context that eliminate risks.

8. Ethical and Legal Implications

The autonomous vehicle may malfunction when hit by malware, adversarial attacks or may itself be guided towards strategic intentions by the sudden exposure to new scenarios, skill training and motivations by the owner. Hence, the vehicle must employ XAI mechanisms to justify every decision. This chapter comes to the fore with comprehensive reviews, evaluation and the development of the practice of AI in cybersecurity. We skim-care survey reports, papers, think tank discussions, policy papers, and new laws formulated by various countries, to identity critical success factors to support the practice of AI in the security domain. [14] [4]

AI has achieved remarkable outcomes in almost all domains, including cyber security, healthcare, public safety, autonomous vehicles, finance, legal sectors, and more. AI has revolutionized the existing scenarios, driving cyber safety to the next level for controlled or autonomous systems. The recent trend of intelligent malware introduced by complex algorithmic behavior has made life difficult for mass security professionals. Attack surfaces have been increased and a huge window of vulnerability is an inevitable outcome. This chapter reviews diverse literature, including random articles, scientific reports, survey reports, think tanks discussions, policy papers, mainstream media analysis, blogs, and new directives by national/international security authorities, and provides several ethical considerations of AI for cybersecurity. Thus, autonomous systems e.g. autonomous vehicle need some degrees of explanation for every decision drawn by it.

8.1. Bias and Fairness in AI Systems

Biased decision-making of the model in hindsight can vitiate robust decision making potentially endangering the autonomous vehicles and its occupants. As human operators and co-drivers are progressively becoming removed from the decision process of autonomous vehicles, the Model-aided Geometrical Shaping of Dual-polarization 4D Formats performance of the AI model becomes increasingly crucial. In the context of pixel-level picture-perfect semantic segmentation in Urban Automated Driving, the importance of a five-module DCNN-based learning and scanning tool is highly exploited [32]. Furthermore, since the bulk

This work is licensed under CC BY-NC-SA 4.0.

of the detection task is executed on the basis of the spatial geometry of the objects, this network fundamentally consists of three elements.

Deep learning models are the backbone of modern computer vision systems that play a crucial role in determining the comprehensive safety of future autonomous vehicles [17]. However, the preponderance of deep learning models translates into a decreased level of understanding in terms of the reasoning process and outputs of the model, thereby attributing to confounding decision makings that might create an environment of doubt, as well as reducing the overall safety and reliability of the decision-making outcomes ultimately present to the end users. As an instance, confrontation of biased data could generate obscured biases within the learning process of the deep learning models, and further amplify disparities in the predictions of the models that coyly rely on such biased information which essentially will result in discriminatory decision outcomes and actions in several applications like health care, law, and employment [33]. Therefore, it is vital to understand the cause behind a model's decision in order to both measure and eliminate the potential biases and to increase the interpretability, as well as the acceptability of the model's outputs.

9. Future Directions and Emerging Technologies

On the other hand, the decisions produced by competitive AI systems are justified by their nature, but their complexity limits their transparency drastically. For instance, applying decision trees with down stream uncertainty-aware models to reinforce causal reasoning beyond learned look-up tables was assumed to be a step forward in this regard thanks to the system's competency as well as its interpretability. The decision process is more transparent due to the computation of scores and splits based on Weibull probability densities. Furthermore, the uncertainty-aware nature of target shaping elements grants the system the ability to assess and communicate its predictive uncertainty [22]. In this study, it will be shown that applying such uncertainty-aware down steam models has the potential to increase the decision process transparency of earlier explained-agent control pipelines (or any other known state-of-the-art RL architecture) without notably affecting their competency. This is a fundamental step toward making such AI agents' decisions more transparent in general.

Explainable Artificial Intelligence (XAI) systems, classifiers and decision trees, used in explaining the actions of AI agents in [2] and presented in detail. The decision trees are used as a back-end computation model when a large stream of data arrives. When faced with a

This work is licensed under CC BY-NC-SA 4.0.

decision to make, the AI vehicle must generate an objective quality estimation of each possible option. Decision trees are simple, interpretable models that can be used to automatically score and provide explanations that summarize why a choice was made.

9.1. Advancements in Explainable AI Research

The breakthrough in acceptability and transparency of artificial intelligence techniques is currently the subject of exciting scientific investigations in various communities. It is vital to close this gap, as the safety of these connected and autonomous systems depends on trust in their decisions. Autonomous vehicles (AVs) have gained tremendous public attention in recent years for their potential to minimize driving fatalities, save congestion time, and improve overall road safety through efficient and safe transportation. The need to deploy AVs has raised significant issues and concerns. Currently, AVs require regulatory-compliant operational safety and real-time decision explainability. Explainability of an AI system-specifically a black-box architecture-is still a formidable research challenge. In response to this challenging issue, explanations of certain key perceptual tasks are a necessary measure for improved human-AV interaction. A significant failure of AI safety assurance for autonomous transportation can be traced back to the failure of the design and testing of learning algorithms. Moreover, the lack of an acceptable margins of uncertainty assessment, especially in real-world environments, is still an open challenge in reallife AI deployment and especially in real-time self-driving AD applications [14].

Trust in AI methods is affected by a lack of interpretability due to the complex structure of typical AI architectures. To address this, a scientific effort has emerged to demystify the inner workings of AI methods and instill community-wide trust in their use [10]. This emphasis on explainable AI is gaining traction with funding agencies and has seen specific success in various fields, such as interpretable machine learning in the weather community. Transparency is a key requirement of explainable AI and is especially important in human-AI interactions in safety-critical systems such as AVs. In autonomous driving, established concepts like accountability of AI and algorithmic causality are discussed, while theoretical concepts like counterfactual causes, such as interventions or given the dependency of (some part of) the action on the causes, as well as vivid explanations and explanations by examples introduce innovative techniques aiming to explain the black-box-like perception and decision-making behaviors in safety-critical applications.

This work is licensed under CC BY-NC-SA 4.0.

10. Conclusion

Where, what, and how should be analyzed better to describe the forecasting process in greater detail and to clarify the underlying logic provoking the input-output mapping. In view of other areas, research attention is more and more focused on the interpretability of the black-box models of the machine learning and deep learning in various applied areas. As yet, the interpretability issue in the domain of smart techniques to be used in many applied areas was simply ignored, because science and technology were focused exclusively on maximizing the forecasting accuracy. This new paradigm, i.e., interpretable and transparent models, seems very important also because one of the most important goals of machine learning and deep learning in many areas is to forecast more effectively and to convey the results to the end-users both easily and naturally. The new approach, interpretable and transparent machine learning, is called XAI, that is, Explainable AI. It is a collection of techniques used to understand or interpret the internal mechanisms of black-box complex models in this field and finally to produce transparent and interpretable forecasting patterns in applied areas. Finally, any transparent and interpretable algorithm unfolds the mechanism of the market agents, which enables experts to follow what causes what, and, what should be done when something is happening. This is so-called inferential statistics in the context for which the present study has been launched [14].

Regression analysis or machine learning serves as a tool for predicting quantitative values within a continuum with X over the whole of its particular universe of instances. A common situation in which this approach is used involves a response variable—a deterministic characteristic of the universe under study (e.g., level of a feature, occurrence of an event) depending on other characteristics (features) of this universe. Regression analysis is universal and is widely used in numerous fields of science and technology including biomechanics, economics, ecology, politics, psychology, marketing, sociology, and law. Mathematically, Ordinary Least-Squares (OLS) and Maximum Likelihood Estimator (MLE) are the most popular methods in regression analysis. Nevertheless, many scientists note that assessing variable importance and determining the logic of the functioning of the background processes are the important and non-routine tasks in various areas [1].

10.1. Summary of Key Findings

This work is licensed under CC BY-NC-SA 4.0.

In this chapter, we aim to provide a summary of the key findings reviewed in the various sections, reflect on the contributions, and suggest potential future work. We emphasize that people should trust deployments of AI and AV industry that meet higher standards, especially because AI can make a difference for improved security in Autonomous Vehicles; In high impact industrial domains such as AVs, it is even more critical that the development of AI-based systems, indeed when they are used, it is transparent, cyber resilient, safe, and secure; From the textual analysis of main international policies and regulations from the European Union (EU) and United States, we can deduct multiple key insights for safe and secure AI in Autonomous Vehicles, and Intelligent Transportation Systems (ITS).

[34] The main contribution of the paper is to provide a better understanding of how AI-based systems for Autonomous Vehicles with AI can be developed using Explainability AI (XAI) techniques. It is showing the relation between concepts AI for Cybersecurity, Software Engineering, and Policies related to AI deployment in the society. [7] We propose designing AI towards more Safety, Security and Trustworthiness with security and transparent reporting. This is in line with our earlier work on Desiderata for building transparent Explainable AI based cybersecurity systems. The paper is dealing with AI-based Cybersecurity and Autonomous Vehicles (AVs) Systems, there is also a strong focus on policies for AI deployment, development, and certification. The paper present better results in software engineering Irina B, Gaël G.

Reference:

1. Tatineni, S., and A. Katari. “Advanced AI-Driven Techniques for Integrating DevOps and MLOps: Enhancing Continuous Integration, Deployment, and Monitoring in Machine Learning Projects”. *Journal of Science & Technology*, vol. 2, no. 2, July 2021, pp. 68-98, https://thesciencebrigade.com/jst/article/view/243.
2. Shahane, Vishal. "Optimizing Cloud Resource Allocation: A Comparative Analysis of AI-Driven Techniques." *Advances in Deep Learning Techniques* 3.2 (2023): 23-49.
3. Abouelyazid, Mahmoud. "Comparative Evaluation of SORT, DeepSORT, and ByteTrack for Multiple Object Tracking in Highway Videos." International Journal of Sustainable Infrastructure for Cities and Societies 8.11 (2023): 42-52.

This work is licensed under CC BY-NC-SA 4.0.

4. Prabhod, Kummaragunta Joel. "Advanced Techniques in Reinforcement Learning and Deep Learning for Autonomous Vehicle Navigation: Integrating Large Language Models for Real-Time Decision Making." *Journal of AI-Assisted Scientific Discovery* 3.1 (2023): 1-20.
5. Tatineni, Sumanth, and Sandeep Chinamanagonda. “Leveraging Artificial Intelligence for Predictive Analytics in DevOps: Enhancing Continuous Integration and Continuous Deployment Pipelines for Optimal Performance”. Journal of Artificial Intelligence Research and Applications, vol. 1, no. 1, Feb. 2021, pp. 103-38, https://aimlstudies.co.uk/index.php/jaira/article/view/104.

This work is licensed under CC BY-NC-SA 4.0.