Explainable AI for Transparent Risk Assessment in Cybersecurity for Autonomous Vehicles

By Dr. Alexandre Vieira

Professor of Informatics, University of Porto, Portugal

1. Introduction

The deployment of autonomous vehicle systems holds great promise for our society, but also raises many concerns related to their proper operation and validation. We argue that the assurance and integrity of these systems cannot rely solely on different forms of verification, particularly machine learning (ML) produced by deep learning. Verification alone struggles to handle the complexity of ML-based systems. In the cybersecurity of autonomous vehicles, one general requirement is to make the cyber risk understandable and predictable by using explainable AI (XAI) models. This article proposes, from the perspective of autonomous vehicle cybersecurity risk management, to join the flourishing area of XAI with the well-assessed realms of security risk assessment. We envision trustworthiness and reliability in the relationship between data and predictions, as well as data validity for training the ML models, as the most important factors to balance in any autonomous vehicle XAI solution for the cybersecurity risk assessment case.

1.1. Background and Motivation

The Capability Maturity Model (CMM) for a goal system lifecycle calls for increasing rigor as well as for managing risk in each of these phases. This is especially needed for AV cybersecurity due to the increasing connectedness and interdependent operation of systems and services, a higher attack profile leading to path vexation, timeliness of attacks at the speed of machine perception and actuation, potential mass adversarial goal states from empty streets to geo-political damage, the Boneh-Deerwester curse of classification with high-dimensional data sets that exhaust training data in volume and variety, and accumulating potentially inconsistent and noisome training and updating sets. PV&V, subject to Reynolds' emergent complexity, provides a disciplined, process-focused, rigorous means to assure and guide the premise that no current system state can predict the properties of a future state.

PV&V provides for risk assessment of individually quantitative PM attributes that are fixed at design time and is not system architecture-aligned. The AV CMM provides rigor beneath risk assessment that may be qualitative and also is not C&C security architecture-aligned.

Artificial intelligence (AI) and machine learning (ML) have been increasingly used to automatically learn models using potentially large volumes and varieties of data in complex systems. Anomaly detection is a classic use case and application. One of the difficult and challenging concerns in the usage of AI implementations is the lack of transparency, a "black box" effect, identifying the causes conforming to the consequences. Explainability and transparency of models and derived actions form a major capability that needs to be enabled. We motivate the notions of AI that is transparent and explainable, namely Explainable AI (XAI), for model accuracy and risk assessment in cybersecurity for autonomous vehicles. We propose a novel Layered Semantic Data Model (LSDM) for graph analytics that combines predictability, causality, and trustworthiness for transparent learning that may lead to effective risk categorization.

1.2. Research Objectives

This paper will address the technical issues related to the development of state-of-the-art AI for assessing and mitigating the risks associated with the use of autonomous modes of transportation. More specifically, it will contribute to the literature on Explainable AI by designing and implementing a number of methods that will allow for very efficient and accurate explainability of AI systems. These AI systems shall be able to continuously assess and mitigate risks and threats in the context of autonomous vehicles, by making use of both formal methods and simple model-based AI planning and scheduling techniques. They shall be capable of putting vehicles in safe states through coordinated behavior planning or adjusting the execution of safe driving actions when a risky situation emerges and will rely on human control or unavailability of formal defensive and dynamic fault recovery actions. Throughout this paper, the term "defensive actions" is used instead of "collision mitigation", because we believe that the safety assurance requirements for an autonomous vehicle are less by far the most natural and meaningful approach to achieve that, even if this may require fundamental changes in the way safety goals are measured and how AE is achieved for autonomous vehicles.

2. Fundamentals of Reinforcement Learning

Machine learning can be divided into supervised learning, unsupervised learning, and reinforcement learning. In supervised training, the algorithm is provided with a large set of input-output pairs, called training examples. The algorithm is trained to map the set of inputs to the expected outputs. The training procedure fits the internal parameters of the model to optimize the model's output. Unsupervised training (or automatic learning) extracts structure from the input data. Clustering is a typical example: the algorithm separates the data into groups based on a principle that is yet to be defined. Reinforcement learning (RL) consists of an agent (typically a physical or virtual entity) learning from interaction. The agent learns to solve tasks through trial and error, optimizing cumulative reward or some appropriate performance measure. It is a feedback-based training, where the feedback is not given by a supervisor or waypoints, but by the outside environment. This is why RL is related to both supervised and unsupervised training. Ideally, it should be seen as a TL-based framing, where the best actions need to be selected in response to the continuously developing environment. In other words, the autonomous vehicle (controlled by the algorithm) must be able to see, predict, and understand the traffic and surroundings to make the best decisions. Recognizing common patterns is very important. However, with years of research, the more fundamental aspects of such a dynamic learning problem have been defined in the classic RL framework.

2.1. Definition and Basics

The input of risk assessment in the context of cybersecurity has been described as well. In practice, the risks are due to inputs or scenarios that, under certain assumptions, could lead to undesired consequences. In the case of autonomous vehicles (and more generally in the case of automated systems), the vehicle cannot be suspended if an AI-based component for path planning or other important functionalities does not behave well. The AI-based component itself cannot be easily checked in an offline environment, nor can it be verified during operation. These aspects stress the importance of having transparent AI-based solutions for autonomous vehicles, especially when they are risk sensitive. At present, many AI-based solutions that are widely used are not transparent, and the search for comparable solutions that are transparent and exhibit similar performance is one of the main research directions.

The artificial intelligence, and in particular the subfield of machine learning, plays a crucial role within risk assessment in the context of cybersecurity for autonomous vehicles. Several

assessment activities can be performed by relying on different AI paradigms starting from the basic concept of supervised learning. AI can also be used as a tool to analyze security in general or to support risk assessment definition. In the first case, AI-based solutions can automatically assess whether the autonomous driving function and its surrounding ecosystem can resist different types of attacks without taking into consideration the environment.

2.2. Key Concepts and Algorithms

2.2.1. Automatically Defining a Confidence Machine The use of confidence machines for estimating the risk of an action has been covered quite a bit in the research literature. The basic intuition behind using a confidence machine, which acts like an "oracle" predicting the probability of being correct for any given decision, is in order to reduce the risk of an unintended or unsafe action. In the theory of reducing the risk of an event that has a low probability of occurring, causal decision theory argues that it is best to ignore such high consequence, low probability of happening events. Further, expected utility from decision theory suggests that even if those events happen with regularity, low probability multiplied by high consequence is still less than the expected utility of ignoring such events. In summary, for very unlikely events such as crashing an autonomous vehicle into another car, neither optimizing the likelihood nor the consequences of the event are justified.

Three key concepts and algorithms that are necessary in order to build transparent risk assessments for safe navigation of autonomous vehicles are discussed: (1) automatically defining a confidence machine, (2) fighting the confidence machine through bagging in order to make it more conservative, and (3) force convergence of the model towards a randomly weighted ensemble.

3. Explainable AI in Cybersecurity

Proliferation of autonomous algorithms for decision support in critical cyber-physical systems and Intelligent Transportation Systems (ITS) increases dependence on the quality and trust of cybersecurity approaches. These are traditionally designed based on fragile and incomplete expert knowledge. AI and machine learning could fundamentally improve security and reliability, especially when leveraging the power of big data analysis. However, lack of trust into black-box models and the negative perception of adding complexity is currently a

counter-fact to a wider adoption of intelligent methods in the main industrial cybersecurity domain.

To reach trust in intelligent decision support systems (IDSSs) for cybersecurity, an approach towards explainable AI is presented. The Fuzzy Intrusion Prevention and Detection Systems (FIPDS) in Intelligent Transportation Networks are used as a test example for critical cyber-physical systems. The proposed model comprises well-interpretable decision rules created by the genetic algorithms, custom fuzzy membership functions, and possibility distributions. Hybrid integer-based encoding for genetic algorithms is combined with expert encoded features in fuzzy sets. The explainable rule-based model is evaluated on expert-preprocessed cybersecurity dataset. The presented approach aids the cybersecurity expertise and provides transparency for operational management and risk-aware design in complex autonomous systems.

3.1. Importance and Benefits

Transparent risk assessment can provide several benefits, particularly in a setting like autonomous vehicles where the presence of an AI is explicit and its safety implications are large. The benefits target three roles for verification and validation (V&V): ensuring that developer-designed requirements and constraints are conformed to, which we call requirements-validation (Re-V); preventing safety problems (preemptive safety); and root causing of safety problems Dur-V. Transparent risk assessment consists of five risk assessment questions, which we will discuss in more detail in the next section: (1) identification of knowledge gaps; (2) understanding the requirements-validation relative to the system design; (3) alignment of safety requirements/validation with normal operation; (4) estimation of level of safety in the presence of attacks; and (5) creation of transparency and maintaining re-v and dur-v properties.

In several critical domains, artificial intelligence (AI) in the form of complex and often opaque models has demonstrated its strength in recognizing patterns and allowing us to make predictions. Cybersecurity for autonomous vehicles is one of these domains, where such AI models are being increasingly used for risk assessment and decision support. In cybersecurity, one must reason about untrustworthy adversaries who can exploit vulnerabilities, as well as about the goals of data confidentiality, integrity, and availability. While testing can identify many vulnerabilities, it is often impractical to test for all potential security problems. As a

result, it is necessary to perform risk analysis using models of the system. For complex autonomous vehicle systems, much of the practical testing occurs in simulation and real-world tests, making it even harder to assess the potential range of risks.

3.2. Challenges and Limitations

The main challenges associated with explaining causality in AIs are significant, comprising technical, practical, and ethical issues. Causality XAI concepts derive from fundamental concepts in physics, which are difficult to apply meaningfully to AI. AI predictions are primarily conditioned to densely connected neural weights, learnable by gradient descent and tech-assisted adjustments, so conventional calculus encapsulates the significance of those complex mappings to inputs and parameters—making end-to-end explainability difficult if at all possible. Furthermore, deep neural networks supervised via noisy label and uncertain human reasoning are probably not how our general intelligence works, so general causality concepts need to be defined, understood, and quantified in an AI-unique way. Lastly, the presence (or lack thereof) of causal relations across AI model components or between models and actual systems outputs does not adequately circumscribe how much an input must be changed to favor or impede some model outcomes—with this quantification being a necessary aspect of explainability. More generally, AI causality is ruled by the work of building a qualitative connection across trained models' features and human-known concepts/attributes/ontological entities, which is currently largely unexplored for AIs.

The most dominant and pressing limitation surrounding the current state of XAI for AIs is its inability to explain causality in the model's decision-making. XAI's principal shortcoming is providing users with correlated insights about the trained AI, instead of guiding them on how to change the input, the algorithms, or the model configuration parameters to change the model's output. Essentially, this means that XAI solutions aim more at predicting what the model will not answer, instead of genuinely and consistently explaining how the model functions, which features contribute to the model's outputs, and how to truly fine-tune the training process for a model to genuinely constitute what humans want a particular predictive model to represent. For derived insights to be truly informative and useful, the behavior of a predictive model must be associated with actionable inputs that actually have causal and not merely correlative attributes.

4. Adaptive Cyber Defense in Autonomous Vehicles

The term "adaptation" in an autonomous vehicle can formally be defined as "the ability of a system to adjust its behavior to changes in its environment and system." A system is said to be adaptive if it acts, learns, and reasons in varying or uncertain environments or operates as part of a group to achieve desired ends. The concept of adaptation is fundamental to autonomous vehicles; they are adaptive since they actively learn and improve performance, are self-organizing, self-monitoring, self-protecting, and self-healing while continually interacting with their environments. This concept of adaptation is a fundamental challenge for the engineering of autonomous vehicles, since the dependencies of the mission and decisions are deeply integrated into the functionality of this vehicle, including dependability requirements. The design of autonomous systems as adaptive systems must face the definition of the adaptability level of the system architecture and of the system functions and their implementation techniques. In addition to the specific tasks provided by autonomous vehicles, adaptability and vehicle recovery are interpreted as a further way to improve the vehicle and their service, either allowing to remotely switch to another vehicle or to the control station in case of the presence of obstacles in the path of the vehicle, or also by mixing all the possible technologies on different autonomous vehicles of the same network.

4.1. Overview of Autonomous Vehicle Security

Autonomous Vehicle Security poses unique challenges due to the multi-modal risk factors that Autonomous Vehicles (AVs) face. These risk factors range from non-malicious, unpredictably adverse events that occur independently of human intervention, to adversarial risk factors, which have not been thoroughly characterized. This paper focuses on predictably adverse risk factors introduced by malicious attackers through manipulation such as actively adversarial components. Such components could introduce evaluation distortion risks, remote control hazards, integrity attacks on stored data, or passenger onboard impact. We present a methodology for semi-transparent decision-making through programmable hardware inferences on visible components and their interaction with the underlying system. More clearly, our IoT methodology not only uses programmable hardware-implemented Protocol Adapters for data transport, but also data processing and decision-making elements in the vehicle's physical data storage and architectural components.

We discuss the various threats faced by autonomous vehicles that require protection. We differentiate between non-malicious risk factors that are unpredictably adverse, unpredictable

network connectivity and infrastructure dependability, special attributes of the network and infrastructure that are predictably adverse and can be used to be utilized, and lastly, adversarial capabilities. This classification helps identify different kinds of risks posed to autonomous vehicle security. While some of the threats to autonomous vehicles are similar to other networked computer systems, others are unique to the on-road environment. We analyze the various components of the vehicle software stack, the properties of external infrastructure required, and potential communication patterns. We present a simple prototype where a shape changing mechanism of the vehicle itself is used to detect a security violation so as to protect the vehicle from such a violation or could detect against a misuse of certain vehicle capabilities.

4.2. Cyber Threats and Vulnerabilities

The new risk assessment concept will be instantiated with supervised document classification methods that can model known specific and adversarial scenarios from available training and validation data. Contextual explanation enables governance-informed design time and operational decision-making. More specifically, proprietary, confidential corporate manufacturers' specifications that reveal system architecture and technology product specifics need to be characterized with conjunctive cyber-heuristics, both developed in this paper, that can reveal opportunities to infiltrate AST or V2X interfaces with spurious AT-Liances (authorization of 4th generation cellular and 5th generation cellular to service alliance adaptation, business, networking service, and security policy layer).

The motivation behind using a risk assessment technique for the autonomous vehicle in cybersecurity is the fact that not only has human safety to be guaranteed, but the resilience of the cyber-physical vehicle itself and ITS (Intelligent Transportation System) environment, including additional authorized services and actors, has to be preserved.

The main contribution of this paper is to present a new concept to address the data-driven limitations of risk assessment of autonomous vehicles at design time and while in operation. This concept inhabits design-time and operational-time risk measures that are not only deep learning model-enabled, but can also account for multi-model scenario-specific learning. It also provides a roadmap to sufficiently assess it without engaging in a deep learning "blackbox" approach that lacks transparency and governance. To the best of the authors' knowledge, no such explainable AI metrics currently exist.

5. Development of Reinforcement Learning Models

To improve the explainability, the present work proposes an extensive experimental infrastructure design. It would make the user aware of the reliability levels from which a trustworthy reinforcement learning-based risk assessment model is made. The extensive reinforcement learning-based experimental procedure along with various trustware models can be extended to other relevant cyber-physical systems. Such a modeling approach could enable the use of the present approach to establish trustworthy explanations and better performances for model predictability. A possibly large volume of algorithms exists in theory for various problem spaces, but they tend to spend time developing sophisticated techniques to provide solutions in practice for, e.g., autonomous vehicles in order to pass the driving certification tests in real-world environments across various governments and geographic locations.

This section presents a set of reinforcement learning-based multi-agent models, including an intensive model analysis, development, and experimentation for achieving explainable and transparently trustworthy risk assessments for the autonomous vehicle's cybersecurity. The originality in the current study resides in the design, development, and extensive experimentation involved with trustware reinforcement learning modeling backed by trustware algorithms. The model development follows a systematic reinforcement learning approach. A trusted awareness level is made highly critical in establishing the trust of any AI-based system, especially critical in the context of cybersecurity for AVs. The transparency becomes additionally vital in malik applications. In traditional reinforcement learning modeling, there is no trace of user satisfaction levels and explanations - both are vital in identifying a trustable model in a domain like autonomous vehicles' cybersecurity. To address this shortcoming, this work proposes the concept of trustware reinforcement learning modeling.

5.1. Data Collection and Preprocessing

The basis of the human-machine interface is the possibility to understand how well the intelligent machine understands the situation and what it's doing. By knowing the level of the machine's understanding, a human can take control of the situation, verify ongoing processes and activities to keep the machine within chosen constraints. If complex cybersecurity solutions are introduced to offer cross-secure protection against various types of attacks and

their possible future derivatives, it is quite natural that our ability to understand the 'safety' picture becomes limited. Humans have only a limited ability to understand highly complex pattern recognition and correlation AI models, most notably those employing deep learning. However, there is a clear and growing appreciation that human interest in transparency and explainability is linked to a range of ethical, security, legal, and social issues. For some applications and situations, the requirement for an explainable system is established through relevant laws or regulations, such as the European Union's General Data Protection Regulation, but many other use cases still wait for applicable standards. For accountable AI support in high-stakes security-critical situations, ongoing development, refinement, and external checks of transparent and explainable AI cybersecurity systems are required, both from a system and algorithmic perspective, and also concerning the underlying data quality.

Explainability in Transparent Risk Assessment

As explained in Section 4, the cybersecurity of connected and autonomous vehicles in the broader scope of future mobility systems is an example of a high-stake security-critical system that faces risks and vulnerabilities with potential severe consequences. Today's highly complex and largely autonomous risk assessment and cybersecurity systems are not designed for continuous human oversight and intervention. The necessary shift from classical cybersecurity systems to AI-augmented, cost-efficient, and more accurate AI cybersecurity systems presents the challenge of establishing an AI-supported and human-mediated feedback loop of transparent and explainable risk assessment for trustworthy and secure risk-optimized operation. We achieve this by creating human-in-the-loop AI systems to enable transparency for human-informed data-driven risk assessment and explainable diagnostics of the predictions and evolving risk picture. Building transparent and understandable cybersecurity systems provides necessary information for continuous human control and support of closed-loop control.

Human-in-the-loop AI into the Loop

5.2. Model Architecture Design

At the beginning, the input camera images are processed with 3 2D Convolutional layers which utilize 24 and 36 filters with sizes of 15, 9, and 5. After this initial stage, two additional 2D Convolutional layers are utilized with filter sizes of 3x3. After the dropout layer, 3 fully

connected layers of sizes 220, 120, and 15 are used. After the penultimate fully connected layer, the final dropout is employed. The model then applies the activation function which is given by the Sigmoid modification that is demonstrated in Eq. (5). In all layers, one-dimensional biases are used. After output neuron calculations which involve the sigmoid activation function, the total model cost, i.e. cross-entropy, is calculated.

Design of the architecture has a direct effect on the performance of the model and the quality of saliency maps. For complexity and computational reasons, non-linear activation functions are preferred to be used after the fully connected layers. However, in order to enable the utilization of LRP based pixel level sensitivity maps, these activation functions should be modified. Here, ReLU and Sigmoid activation functions have an underlying assumption of positive relationship i.e. if we add a small positive constant, the result will not change. Therefore, a modified version of ReLU and Sigmoid (shifted ReLU-Sigmoid) functions are utilized. The slight modification does not affect general properties of these functions such as continuity and differentiability. Therefore, even by changing the inner workings of these functions, the functionality of the model does not impoverish. Thus, the modified functions enable the usage of LRP.

6. Explainability Techniques in Reinforcement Learning

6.1 Reinforcement Learning Explainability for reinforcement learning (RL) is in its infancy since deep learning is a precursor for bespoke RL. There are models that can explain convolutional layers at the pixel level in image classification tasks. The learned model's actions, observations, and policies are understood by providing oversight on a restricted sub-network in response to predictions. General concepts are learned from clustered activations over multiple time steps to explain recalled policies. These forms of explanations are not useful in our cybersecurity domain. Another popular strategy is to force predicted explanations to be simple and abstract by using L1-norm regularization with sparsity constraints. Penalizing temporal activations of explanations yields too simple and insufficient informative representations. Reinforcement learning behaviors are identified heuristically in a network implementation. These techniques provide insights for optimization and provide a walkthrough of what features matter, hence opportunities for refinement. The types of behaviors these models may not align with risk indicators needed for cyber assessment.

Risk assessments in cybersecurity typically assume worst-case behaviors. Sequential behaviors in adaptive agents create entities with more complex and sophisticated environments and patterns. The use of explainable AI can provide mechanisms to identify these witnessed behaviors and help refine risk assessment calculations. Claims made on observed behaviors can lead to better outcomes, minimizing false negative/positive results. We surveyed the landscape of AI techniques for cybersecurity, explainability techniques for complex agent-based systems, and relevant interactive machine learning techniques, showing that these existing techniques do not meet challenge requirements. We set forth an agenda for research to address technology gaps with temporal constraints, dynamic environments, and domain understanding that are requirements of autonomous vehicle cybersecurity.

6.1. Interpretable Models

Local Interpretable Model-agnostic Explainer (LIME) is arguably the best-known work in the latter category, which utilizes local approximators. Furthermore, More et al. propose an automatic interpretation scheme that applies symbolic techniques to automatically organize the knowledge discovered by interpretable models. UnsupportedOperationException is tackled by linear programming to guide the learning of local approximation models. Besides uncovering vulnerabilities, explanation capabilities can further pinpoint adversaries for defeating or fine-tuning policies to avoid adversarial behaviors.

Interpretable models convert an ensemble of models or a complex model with non-linear operators into a relatively simple model such that it can be easily understood. In general, there exist two strategies in interpretable models. First, an interpretable model produces globally linear decision surfaces mimicking the decision boundary learned by complex non-linear models. Specifically, Ibrahim et al. propose a linear approximation of an ensemble of decision trees for a general domain. This approach optimizes the linear approximation within the group sparsity concept, which they call tree-constrained group lasso. Imtiaz et al. address a similar problem using stochastic learning on the inverse mapping. Second, an interpretable model identifies the regions in the input space that are most confusing to the base model. These local interpretable models allow us to understand why the model behaves a certain way (e.g., decision for samples) within certain regions of the input space (e.g., group of samples). Due to recent research efforts, the latter strategy has achieved various methods which correlate the decision of complex models with organized, spatially explicit data.

6.2. Local and Global Explanations

The approach used to extract from Model Card Toolkit answers whether the model: 1) Provides a global overview of the intended use, performance, and characteristics of the model, 2) Provides examples of different inputs, 3) Provides overviews of the model, including relevant decisions, choices, assumptions, and limitations, 4) Provides ethical considerations and trade-offs underlying model decision-making? 5) Includes robustness and error-reasoning, is the model output trustworthy? 6) Includes strategy, guidelines, and limitations for users to interact with a model responsibly. For the creation of the global explanation of XAI recommendations for autonomous vehicles risk analysis, would few sensitive information that could compromise cyber protection by disclosing: restricted operational details, algorithmic explanations, or access to the data when it comes to global explanation within the model card paper, to have a basic amount of minimal standard of care.

Explainability can be defined as the extent to which humans can understand the cause-and-effect relationship between the decision-making rules, the input data, and the decisions made by members of an AI system. For machine learning models, explainability comes in two main kinds of communication strategies: local explanations that explain how individual decisions were reached by the model and global explanations that describe how the entire model generally works (what features are relevant, what influence do they have, etc.). Local explainability of the process is an important factor in decision making, especially responsible decision making, because it improves trust among users by providing them with good information and a better understanding of the agent's behavior. When agents use specific models to make sense of the world, dragging those models into daylight in order to understand what they are doing is vital.

7. Case Studies and Experiments

Cybersecurity Exercises: The risk reduction supporting effect was demonstrated through four drills, and its robustness was examined. Since Red Team operations were not part of the planned program, the planned content was selected as scenarios close to actual reconnaissance. On the other hand, the characteristics and traits of insider attack scenarios were utilized in each actual Red Team operation exercise. Five Red Team operations also took place during the last two days, executing the "power aerial control situation experiment". These Red Team operation monitoring exercises are called Cysho_experiment_II.

Red Team Operations: The purpose is to understand the risk levels calculated by the risk assessment system during Red Team scenarios. Six Red Team operations for attacks on information confidentiality were conducted. Then, a presumed countermeasure was taken, and 5-10 days later, another Red Team operation was conducted to assess the effectiveness. These Red Team operations are named Cyshoexperiment_I. A Red Team operation is an exercise in which professionals simulate attacks on a level similar to an actual invasion by an attacker to verify and improve the organization's defensive capabilities.

We performed case studies for two cybersecurity exercises using the proposed risk assessment system. In the case study experiments, defenders attempted to reduce the calculated risk scores. The average risk scores from the case studies were analyzed, and the system availability explained in the section titled "Evaluation of System Availability" was based on such critical risk experiments.

7.1. Simulation Setup

The radar sensor data of the lead car is utilized. The overtaking vehicle should pass the ahead one while avoiding the safety-critical zone of the ahead vehicle, minimizing the trajectory deviation and maximizing the comfort in the proper time window. The focused points in the research work are: What kind of risk is involved in executing the overtaking manoeuvre during deep reinforcement learning? What influences the decision of executing the action of the autonomous vehicles during overtaking the most? Although exploiting the Doppler effect, the ability of resistance to uncertainty and efficiency which radar provides in ranging and velocity estimation has given rise to a large number of research projects in the automotive fields. Crucially, a spare radar sensor for autonomous vehicles has the capability to provide reliable distance and velocity estimations for multiple targets concurrently. Such an observation is particularly important in the unstructured, highly cluttered, and dynamic environment. The fidelity of the sensor measurements is directly related to the safety and performance of the autonomous vehicle. An emergent action is propagated based on the deep reinforcement learning algorithm in response to the sensor measurements. Commonly, behaviorists can train the model with the huge volumes of real vehicle driving data. Without doubt, the simulation cannot copy the endless and unforeseeable conditions in the real driving field perfectly unless it contains all possibilities. Optimism bias cannot be ignored due to the lack of knowledge of some exceptional circumstances.

In the study, we consider a communication network comprising multiple vehicles implementing an array of applications/safety-critical Support Cooperative Automated Driving (SCAD). A deep reliant car is invoked to conduct two overtaking manoeuvres. One with realistic driving behaviour and the other with a worst-case scenario. Contrary to the well-known particle filter approach commonly employed in automotive fields, we use out-of-the-box reinforcement learning. We aim to demonstrate the overall effectiveness of the deep reinforcement learning approach for assessment of corresponding risk and SCAD in driving activities. Moreover, we aim to provide understanding and confidence in the operation of machine learning models through transparent interpretation based on SHAP. We use different behavioural representations of the overtaking manoeuvres to thoroughly demonstrate the benefits of the SHAP explanation in autonomous driving systems because real-world demonstration leading to some conclusions can have legal aspects.

7.2. Results and Analysis

Experiments are implemented with varieties. The weights become very small, which can affect the other weights during the learning. However, the average scores of real output results are high enough for the risks which are required to be explained for interests. And the number of training is decreased by the applied CNNs that have small sizes of parameters. The deep neural network can have sufficient performance under the limited comparing with the traditional deep neural network among the performance for the accuracy in the autonomous vehicle security. In particular, the scores of sensitive data discoveries are high, the scores of zero days attacks are also high, and the discovery of unsupervised adversarial attacks are also interpreted with the local explanation of LIME. The analysis results can guide the risk management in the cybersecurity for the autonomous vehicle using the multi-sensor fusion with CNNs.

The experiments are conducted with various types of data, where the types are simulated data and an autonomous vehicle's usage data. In the simulation experiment, the sensor types are modeled on 3 types of sensors which are LiDAR, camera, and radar. Two types of risks, sensitive data discovery and adversarial attack success model, are applied to the sensors. The performances of the deep neural network are validated under a white box and a black box model. We also investigate the frequency of changes of the outlier values which can affect the

performance of the deep neural network. And the performances with explainable AI technique are validated for getting the interpretation of the risk scores.

8. Ethical and Legal Implications

Part of the legislative requirements guiding autonomous and connected vehicles is section 44.1 of the Italian Highway Code introduced by the DM-30-07-2018. It states that "With regard to active safety systems, the vehicle shall have a transparent human-like behavior, capable of alerting the road users exposed to the risk generated by non-compliance with traffic rules and eliminating the risk if they do not react."

Although it is often beneficial to use ethics, legal compliance, and evaluation, we are early in the process of being able to provide guidance on producing working applications that take into account the core ethical and legal concerns of individuals and communities. Different jurisdictions and individuals adopt distinct ethical and legal normative frameworks and behave differently. One opinion is that all activities implemented by autonomous mission platforms mimic decisions made by human operators and therefore follow the rules of human decision-making processes. Several studies and articles linked to the potential of AI and autonomous developing technologies have been made. The requirements for ethical compliance that influence autonomous decisions were expounded, and several recommendations were provided. These include the ability of relevant external observers to identify that autonomous decisions were made, the ability of such observers to identify the general internal decision-making process that led to each decision made by the autonomous platform, the ability to modify the influence that key stakeholders have over the decision-making process that leads to any autonomous decision, and the ability to take into consideration the ethical boundaries established by the human operators.

8.1. Bias and Fairness Considerations

For instance, consider a high-stake risk score that is learned from an unconditionally fair algorithm. Now, due to the fact that objections such as "if it was my family in an error, I'd rather take my chances with something unfair" correlate with a conditional definition of fairness rather than an unconditional one. To better understand fairness and measure unfair biases in the data, it is important to consider the proposed application domain.

Moreover, multiple attributes exist for any individual. Which protected attributes, group membership, and variable interaction are pertinent? Fairness is context-dependent, and both an organization and the audience determine what relevance entails. Additional nuances to the fairness problem in machine learning rely on the fact that fairness is a multidimensional utility of conflicting goals. Consequently, the process of obtaining a fair application may not just involve adjusting the algorithm.

Fairness in machine learning entails operating systems in a way that ethically accounts for the allocation of fair opportunities, treatment, candidates, and outcomes. For instance, a fair selection system would treat two candidates with equal qualifications, characteristics, and background in an equivalent manner without any biases. An unfair system might tend to over-classify an individual based on a particular demographic that is highly correlated with a protected attribute (e.g. race, sexual orientation, religion). While the goal is to eliminate unfair bias and increase fairness in decision-making, it is important to note that fairness and bias are not universally defined.

8.2. Regulatory Compliance

The everchanging landscape of tailored legislation needs to be considered when developing tools and guidelines like an AI-based XAI approach that aims to be a cornerstone to build trust into the new era of autonomous vehicles. In summary, our AI-driven cyber psychological model is capable of providing explanations, risk assessments, and tailored suggestions to enhance self-awareness, trust, and interventions that spell out explainable privacy and security protocols, needed to define future, autonomous, and cognitive transportation environments and address the compliance and adherence requirements from frameworks, guidelines, and regulations.

While the Federal Motor Vehicle Safety Standard (FMVSS) number 150 does not directly regulate the privacy of information and cybersecurity risk assessment, the Definition of Vehicle Security-Related Terms in Cybersecurity Assurance Guide for J2964— SAE's Cybersecurity Guidebook for Cyber-Physical Vehicle Systems defines the terms that distinguish the whole lifecycle cybersecurity risk assessment from other lifecycle functions of a vehicle.

There are various standards and guidelines that address the protection of personal and other sensitive vehicle data. For example, the Motor Vehicles Data (Privacy and Security) Regulations (MVDSR) in the state of California defines requirements for original electronic equipment manufacturers (OEMs) and third-party service providers to protect information that they access, receive, store, or collect from vehicles already in California.

9. Conclusion and Future Directions

We have conducted experiments to validate the performance of the proposed methods and compared our proposals with other available explainability techniques. We have also reported the limitations of the proposed techniques while discussing different related work. Our study aims to promote transparent risk management related to class-conditional predictions in order to interpret learning models in different domains. Still, several avenues could be considered in the future.

Lastly, besides these specific use cases for autonomous vehicles, we employ the limit-based method to manage predictions of a variety of other deep learning models (e.g., image or audio classifiers or recurrent neural network-based models) by identifying confidence intervals that contain adversarial examples or other obstructions.

For the image classifier, the proposed method, in particular, has shown promising results. It enables risk managers to identify obstructed traffic signs from an abnormal distribution of gradients without significant performance impact. This could help manage the confusion of autonomous vehicles, which might otherwise classify images incorrectly due to ambiguities in a proposed security system of obstructing traffic signs.

Recent advances in Explainable Artificial Intelligence (XAI) have been proposed for both transparent risk assessment in cybersecurity for autonomous vehicles and for interpreting the behavior of such models. We have demonstrated the effectiveness of a family of risk assessment methods for identifying misclassifications in the behavior of an image classifier, which detects traffic sign obstructors, as well as for a semantic segmentation model, which detects dark pixels in a traffic image.

9.1. Summary of Findings

- With Guidelines: 1. Introduce guidelines for the adversary which define and then limit how far and in what direction adversarial search algorithms can search for adversarial perturbations. 2. Train risk prediction models using moderate amounts of randomized data to establish general robustness. 3. Use a hyperbolic tangent neural network activation function. 4. Use a hyperbolic defeasible rule as the base AGDS and for the adversarial attack risk detection model. 5. Design the AGDS in accordance with the moderate size with the training size used, with the constant value representing moderate size generally increasing with more training data. 6. Configure the Risk Model data to enter the adversary in holdout validation.

The results of the study are based on the best and most consistent neural network model performance, which are the final architectures used, the architectural-specific parameter sets -

- Model Robustness: Model architectures, the architecture-specific parameter sets, data sets, and amounts of data all play significant roles in determining model validity and robustness. In the adversarial vehicle attack problem, which is the focus of the autonomous vehicle (AV) cybersecurity domain and this AV cybersecurity model, suitability can be measured using common techniques. These common techniques are robustness to the detection of corrections in the generalized attack scenarios for real-world application to urban driving.

- Risk Assessment: The neural network AV cybersecurity model predicts values in the [0, 1] interval representing the risk of each possible input value being part of an adversarial attack. Explainable AI methods such as LIME show promise in making it easy for security professionals to understand why this prediction is made and to find adversarial scenarios without using additional algorithms or domain-specific knowledge. However, as adversarial scenarios with risk outcomes being zero also have a high probability of being discovered, it presents a significant challenge to risk-results validation and to the imminent danger posed by adversarial attacks, but it is more of a side benefit attached to the probability risk model-training process.

This study set forth its data, models, and findings regarding the effectiveness of using explainable AI methods to understand how risk assessment is performed in autonomous vehicle (AV) cybersecurity models. Specifically, it focused on the capability of the explainable

AI methods for adversarial scenarios and the performance across varying architecture designs, architectures, data sets, and training sizes. Findings presented are as follows:

9.2. Potential Areas for Future Research

Integration with Traditional Risk Assessment Methodologies: A deeper look at RAIM would reveal areas where traditional safety assurance and risk assessment methods would need to be supplemented, such as in determining cybersecurity-related mission impact in the mission-tree development. RAIM for Other Aspects of the Transportation System: Today AI is prominent in all modes of transportation: airlines, surface transport, and maritime. This paper focuses on the most complex mode of transportation: aviation, but future work might consider multimodal transportation risk assessment, as well as the emerging markets of small unmanned aircraft used in package delivery. The cybersecurity of analog components (e.g. secure, or air-gapped, flight control systems) in complex autonomous systems may also be illuminated by our use of AI in explaining their risk. Regulatory Agencies: Lastly, our inclusion of decision-ready text takes the RAIM ontology concepts of information and decision readiness a step further. Regulatory agencies require information that enables them to judge RAIM processes and procedures, output transparency and accuracy, model confidence in the integrity of their operations and products, decision transparency and timeliness, and standardized reporting for increased model interpretability.

As this paper demonstrates, explainable AI may both foster trust in AI and improve the performance of AI models. This paper is positioned at the intersection of several disciplines and tools: private blockchains, explainable AI, automated risk assessment, and, as an application, autonomous vehicles, as well as systems engineering, cybersecurity, and autonomous system safety. The research presented is wide-ranging and a bit shallow with regard to each topic covered, so future work might investigate the issues more deeply. This paper considers three areas, in addition to extending the blockchain for Khatri and Sangal's private blockchain and Persaud et al.'s integrated framework.

10. References

1. A. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

2. F. Doshi-Velez and B. Kim, "Considerations for evaluation and generalization in interpretable machine learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 979-985.

3. Z. C. Lipton, "The mythos of model interpretability," ACM Queue, vol. 16, no. 3, pp. 31-57, 2018.

4. P. Hall, N. Gill, and B. Schmidt, "Towards an interpretability evaluation framework for machine learning," arXiv preprint arXiv:1901.04592, 2019.

5. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, 2017.

6. F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 8, pp. 2674-2693, Aug. 2019.

7. Tatineni, Sumanth. "Compliance and Audit Challenges in DevOps: A Security Perspective." *International Research Journal of Modernization in Engineering Technology and Science* 5.10 (2023): 1306-1316.

8. Vemori, Vamsi. "Evolutionary Landscape of Battery Technology and its Impact on Smart Traffic Management Systems for Electric Vehicles in Urban Environments: A Critical Analysis." *Advances in Deep Learning Techniques* 1.1 (2021): 23-57.

9. Mahammad Shaik. "Rethinking Federated Identity Management: A Blockchain-Enabled Framework for Enhanced Security, Interoperability, and User Sovereignty". *Blockchain Technology and Distributed Systems*, vol. 2, no. 1, June 2022, pp. 21-45, https://thesciencebrigade.com/btds/article/view/223.

10. Vemori, Vamsi. "Towards a Driverless Future: A Multi-Pronged Approach to Enabling Widespread Adoption of Autonomous Vehicles-Infrastructure Development, Regulatory Frameworks, and Public Acceptance Strategies." *Blockchain Technology and Distributed Systems* 2.2 (2022): 35-59.

11. S. Rajalingham et al., "Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks," Journal of Neuroscience, vol. 38, no. 33, pp. 7255-7269, 2018.

12. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

13. R. Guidotti et al., "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, pp. 93:1-93:42, 2018.

14. K. G. Papakonstantinou, G. Korkinof, and S. T. Papadopoulos, "Risk assessment framework for autonomous vehicles," in Proceedings of the 12th IEEE International Conference on Intelligent Transportation Systems, 2019, pp. 2063-2070.

15. A. Holzinger et al., "From machine learning to explainable AI," in Proceedings of the World Symposium on Digital Intelligence for Systems and Machines (DISA), 2018, pp. 55-66.

16. J. Lin et al., "Explainable deep learning: A field guide for the uninitiated," in Proceedings of the IEEE International Conference on Big Data (Big Data), 2018, pp. 1676-1685.

17. A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020.

18. C. Rudin, "Please stop explaining black box models for high-stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.

19. F. Doshi-Velez et al., "Accountability of AI under the law: The role of explanation," arXiv preprint arXiv:1711.01134, 2017.

20. A. Torralba, A. A. Efros, "Unbiased look at dataset bias," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1521-1528.

21. A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018.

22. S. S. Shen et al., "Interpretable credit risk modeling for SMEs using machine learning with feature engineering," IEEE Access, vol. 7, pp. 164111-164121, 2019.

23. P. Gadepally et al., "The big data ecosystem at Lincoln Laboratory," IEEE Transactions on Big Data, vol. 2, no. 1, pp. 34-45, 2016.

24. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," arXiv preprint arXiv:1312.6034, 2013.