Explainable Reinforcement Learning Models for Adaptive Cyber Defense in Autonomous Vehicles

By Dr. Andrés Páez

Professor of Industrial Engineering, Universidad de los Andes (UNIANDES), Colombia

1. Introduction

[1] Cybersecurity continues to be an increasingly critical challenge for autonomous cars, and intelligent attackers can develop adversarial real-time attack strategies that are difficult to anticipate or mitigate. With conventional rule-based cybersecurity strategies and solutions, such as those used in traditional ITS (intelligent transportation system) scenarios, autonomous cars will not be able to anticipate all these possible real-time adversarial strategies to which they will be subject, thereby receiving suboptimal attack resilience. On the other hand, deep reinforcement learning (DRL)-based cyber defense strategies are able to model and optimize dynamic, multi-stage, and uncertain strategic games against adversarial black-box attacks. Unfortunately, this solution space over the state-action space of an autonomous car's DRL policies contains adversarially optimized policies or black-box adversarial actions, and their stop effect can catastrophic.[2] There has been tremendous growth and advancement of technologies used in intelligent traffic systems (ITS), autonomous vehicles (AV), and intelligent transportation systems. This advance has made it possible to carry out various routine activities such as traffic management, dynamic route decision making, and adaptive cruise control. However, this smart mobility system employing intelligent transportation technology has the potential to be exploited by the cyber attackers. The exploitations can result in devastating outcomes such as the violation of data privacy and digital security, road traffic congestion, and disruption of local services due to malfunctioning traffic lights and automatic cars. Hence, it is vitally important to understand the inherent cyber vulnerabilities, and study the potential impacts of cyberattacks to the mixed traffic flow of various levels of connected and automated vehicles. The appropriate framework for evaluating the prevailing infrastructure and requirements of cyberattack resilient control strategies is intensely demanded.

1.1. Background and Motivation

An effective strategy for ensuring the security, trust, and safety of CAVs in the future must put the user at the center, with a complementary explanation of all ML-based applications decision-making in complex, sensitive settings—such as driving—a crucial requirement to achieve this goal. This is increasingly critical when we consider fully-autonomous and highly-automated vehicle domain applications, where human drivers' control, preparation, and interaction with vehicle systems have become secondary and co-dominated by the appropriate instructions prescribed by the decision-support algorithms that are enshrined in ML applications integral to AD. The improve the security of CAV ML decisions by integrating contextual information and using the Dempster-Shafer Theory for evidential reasoning, with the latter proving to be effective to both mitigate the problem of imbalanced attack classes and ensure that all different signals can be suitably treated, thus producing a level of explainable prediction that will be even more suitable to provide it to trst. Such an application will be the key suitable jump forward to allow a best-practice demonstration in the language of the main potential stakeholders and to make necessary the creation of a new model to maximize RFI. Knowing the vulnerability of current systems to adversarial attacks, the prediction of these models will be fed with perpetually updated data so as to make precise countermeasures that can change according to the attack support by detecting which time window can be of most help with AD intelligence [3].

Connected Autonomous Vehicles (CAVs) have benefitted from the increasing reliance on Machine Learning (ML) algorithms to assist or make fully automated decisions, but their applications open new vulnerabilities in form of adversarial attacks such as spoofing. To ensure the security and robustness of CAVs against these kinds of attacks, it is important to build ML-based applications that are able to interpret the signals they receive and their inner functioning to ensure a prompt and effective response to defend the vehicle's security. The explainability of models is essential for creating short- and long-term security applications in the field of automotive security [4]. In this article, the work focuses on the use of Reinforcement Learning (RL) to create reactive applications in association with an explainable model named SHAP, that supplies a subset of general relief feature selection criteria according to which countermeasures at an application level can be derived. The focus on this research is twofold: On the one hand, we aim to develop and test an explainable model for detecting attacks against connected vehicle sensors, while on the other hand, we intend to use the same

model to provide the required intelligence to make this system an Adaptive Learning Intrusion Detection System (ADLS).

1.2. Research Objectives

Tygar's seminal paper on "Autonomous cyber defense: A research agenda" envisions "reinforcement learning-based infrastructures" in the first chapter of the Defense Life Cycle as a highly desirable long-term research goal. This has become increasingly timely considering the transportation industry is likely to include artificial intelligence based solutions in future autonomous vehicles. In this context, in this work we examine the security characteristics of these AI-based solutions and compute an adversarial-free training policy using a model-based reinforcement learning agent [5]. We also discuss how to analyze obtained policies with more complex game theory models and introduce further challenges to be covered by future research.

Reinforcement Learning (RL) [6] has been successfully used in the autonomous vehicle industry to aid with tasks such as maneuver and route planning and learning drive-by-wire control systems. Determining the safety and security risks of AI-based and autonomous systems has been a hot topic in the research community. However, when it comes to the cybersecurity aspects of RL models for adaptive cyber defense in autonomous vehicles, very few works have touched on this research area, the direct incorporation of a threat model in RL training has yet to be touched on.

1.3. Scope and Organization of the Work

article_id: 7c961118-d939-4041-a07f-95b5ebded0e7 article_title: Secure power injection and data exfiltration via autonomous vehicles in V2X article_main_idea: Autonomous Vehicles (AVs) communicate with other entities in their vicinity for a number of reasons. This article delves into the different ways AVs can unintentionally help or intentionally act as agents for attackers. It highlights these vulnerabilities and offers evidence-based recommendations for mitigating their risk or impact.

article_id: c228ba7d-2975-48e1-bf8b-27b84aaac1b4 article_title: A Game Theoretic Approach to Address Smart Grid Cyber-Physical Security Challenges article_main_idea: Integrating multiple renewable energy sources on Smart Grid (SG) systems, leading to changes in resistive characteristics and evolving digitalized control systems, introduces a highly volatile and

complex cyber-physical environment which can lead to cyber-attacks or exploitation of vulnerabilities. Therefore, the effectiveness of SG depends greatly on ensuring reliable and secure functionality. The main objective of this article is to identify industrial gaps and performance metrics in SG, discuss the advantages of game theory as solutions to identified issues, and propose fundamental needs for research. The study considers the game theoretical approach as an effective scheme in cybersecurity, related secure control and coordination problems in SG.

Articles: article_id: b238c17e-bde0-4055-8f47-4d9508e9b036 article_title: Design and Analysis of a Secure Fog-Aided Edge Computing Framework for Autonomous UAVs article_main_idea: Edge and fog computing have been widely discussed for enhancing performance of contemporary autonomous vehicles. However, the security aspects of utilizing fog computing for such vehicles remains unexplored. To this end, this paper presents a comprehensive fog computing security analysis for Secure Fog-Aided Edge Computing (SFAEC) for fast data acquisition and efficient data analytics in high level UAV applications. The results show the essential role of fog servers in providing fast and secure UAVs data acquisition, analytics and situational awareness in real-time. Also, the communication link and the edge-fog architecture plays a significant role in security performance of the institutions.

[7] This work focuses on the design and implementation of an explainable adaptive cyber defense architecture for manufacturing industries. In this context, we investigate deep reinforcement learning models that perform an anomaly detection and response strategies in order to protect the vehicle against cyber-attacks. To design engagement policies in autonomous cyber defense, we have to address several issues: the anomaly detection and classification, the design of proper engagement policies of the Autonomous agents, the availability of a proper dataset with proxing of cyber-attacks and defense strategies to apply. To this aim, in this paper we will dissect the components designed in the architecture and solve each of these issues separately [].

2. Autonomous Vehicles and Cyber Security

The situation of the technologies and skills needed for cybersecurity in autonomous vehicles and connected cars are shown in Appendices A and B respectively. Furthermore, as shown in Figure 3, their needs of industry and research with this technology are current conditions and

can be seen the requirements of industry and the academic sector for the future autonomous vehicle and connected car. In the field of cybersecurity vehicles, in Figure 3, it is seen that that who will supply the need of the cybersecurity of vehicles. As a consequence, these risky areas necessitate a particular focus on cybersecurity. Involvement of the automotive sector and relevant sector in contemporary approaches is also needed to fulfill these specific cybersecurity requirements. Overall, catered safety levels for vehicles and participants are identically important performance factors. Transport safety potentialities are evoked by systematically making technical and non-technical advanced security experiences and knowledge available for industry development.

[6] [8] Transport of people and goods of all kinds is an essential part of our daily lives. One of the most important aspects of future transportation is automation, self-driving cars specifically. However, as we know, along with automation comes a transformative growth in cyber security threats ranging from common malware to advanced persistent threats. Advanced autonomous vehicles and connected cars, which link autonomous cars, are new areas of risk for the automobile industry and automotive cybersecurity Researchers and analysts have only recently begun to take note of the dangers posed by known and newly emerging cyber-attacks on autonomous vehicles. However, still there is no broad information on the functionalities and potential results of different cybersecurity threats on autonomous vehicles. Almost all these research aspects in the literature are focused on the detection and prevention of conventional cyber risks by using intrusion detection systems and firewalls. As such, there is limited focus on advanced cybersecurity threats and minimizing their dangers, e.g. polymorphic malware, denial of service, zero-day attacks and the weaknesses of the security model.

2.1. Overview of Autonomous Vehicles

The driverless car, as a major scientific invention, results in great joy for human society. By turning around, the car can realize safe and smooth driving in indoor parking. Irobot is a very successful company that develops vacuum cleaners for consumers' daily life and business cleaning. The Touran robot of Irobot Vacuum Cleaner adopts sensors to sense the environment, intelligent identify relative locations of the room, avoid staircases and walls, automatically clean 2 The company website: [Link] ver 1.1, Dec .assemble the robot vacuum cleaner system, and charge automatically when it runs out of power. This is a very typical

example of successful indoor autonomous vehicles. Enabled by the development of information and communication technologies, driving automation has experienced rapid progress recently. The connected autonomous vehicle (CAV) has yet to be integrated into everyday traffic, but there is little doubt that this is only a question of time. With this newfound mobility there also come new risks and vulnerabilities [1–6]. When cyber-physical systems are involved, the main attack vector is digital and not physical— as advanced driver assistance systems and automated vehicles are equipped with a huge number of sensors and communication channels, attackers can use a variety of channels to impact vehicles, for instance, GPS in the case of location spoofing, digital radio signals (DSRC) in the case of paralyzing the whole traffic situation, or hacking into the sensors and injecting false observations.

In, a flexible connected-automated driving simulation environment for connected autonomous vehicles (CAVs) is presented that supports the rapid development and evaluation of driving automation features. It takes into account the significant influence of CAV technology on the development of the V2X ecosystem. However, they do not consider any security solutions for the transmitted data in the connected-automated driving simulation environment. Another significant solution in discusses the security and privacy issue with regard to remote cyberattacks in CAVs using connected vehicular networks. It highlights several critical privacy-related characteristics associated with different network technologies that Captivate CAVs. discusses the simulation of connected autonomous drone systems with the help of a hybrid simulation network. Moreover, the secure spectrum sharing in the use case of connected—and partially—autonomous drones systems presents the essence of 5G's simplest network slicing concept.

article_id: 5fe0a7fd-73af-4cb9-a010-110b0050486c

In this work, four interdependent services present autonomous machine attacks. These target the AV's camera sensor, DoS attack—an implement on the AV control command, GPS spoofing—the impersonation of an incorrect GPS signal deception, and infotainment malware so that an unauthenticated member inside an AV network be transformed as authenticated control as shown in Fig. 1. A Multilayer security agent that provides adaptive security arrangements under both adversarial and normative tracking. MagicMock blockchain technology for the establishment and steady maintenance of valid AV control connected

network. The adversary models have been considered on the basis of the AV's network, channel, and designed security agent evasion nature. By tailoring alone, sweeping through the immunity till inflow, adversary real-time adaptability, finally onboard all connected AV. Regulatory-based customdeveloped devices required to effectively segregate the entirety of danger and recovery process. The vehicle's manoeuvrability is best when the dependency of Driver is totally eliminated, and it actuates safely adapting itself to Environment and fellow entities, and over adaptive control. If any coexisting vehicles fume situation occurs, while current four levels are transferring control over to the machine completely with less user discretionary autonomous behaviour, directly working on these, will affect Infotainment normative device. Further, with trainable physical model, a UKF-based AV infotainment fault analyzer is used as a covert channel to mitigate the ad hoc concept. This malicious infotainment is only present to create an illusion of its previous integrity and is not a part of any infotainment console. An on-ground attack in this scenario is a double-dealer. While one is infotainment attack, a GPS spoofing interferes to continuously create Maneuver errors that reveal the true manifested adversary.

article_id: 2842b5b4-4ed9-4c43-8259-81eae66a65df

The level 5 autonomous vehicle (AV) is self-driving without any human intervention. In the Level 4 AV, the AV's guardian technology provides backup in case of an emergency or system failure. Physical road tests in real-world driving environments, Driving (Hence) are the only means for validation and testing that expose AVs to a variety of changes in driving environment and, thus, sharpen their software and hardware design to be transferable to a wide range of driving scenarios. As current and future generations of AVs offer various value added services, including connectivity with an outside communication network, the global connected AV market's growth is anticipated to be exponential in the forthcoming years.

article_id: 2f91accc-fc26-4f0a-b6a5-015cdfe7b43a

2.2. Cyber Security Threats in Autonomous Vehicles

[9] The primary components of the CAV include the main unit, such as the autonomous vehicle itself, and external subsystems such as automotive suppliers, communication networks, and service providers. Each of these systems is susceptible to specific structural, configuration-based, and procedural vulnerabilities; some may not even be cyber-related.

This diversity multiplies the potential threats to CAV security. In the automotive domain, the clear distinction between passive entry points, such as diagnostic ports designed for servicing purposes, and fully protected system areas according to the OBDS-II or vehicle-to-everything standard has been washed out. Cybersecurity for 5G technology is also heavily influenced by the application area. 5G wireless systems are crucial industrial infrastructures, with immediate impact on economy. The security and privacy of fifth-generation (5G) systems like regular customer communication, communication with cars and drones, and internet of things (IoT) communication will be impacted especially. The repetition of well-known security weaknesses when deploying novel 5G technologies would be a big mistake. The most important precautions for better network security and privacy are using certified 5G communication components and access control mechanisms and defining closer security requirements for future 5G products.[10] Just like conventional vehicles, CAV are not only the means of getting from any location A in the world to location B, but serve a number of purposes: economic, social, and recreational. Today's cars are complex mechatronic systems empowered with help from on-board software (which typically is a tailored Linux system), and they possess a number of electronic control units (ECUs) interlinked using communication buses. These units support, e.g., drive system control, maneuvering, and collision prevention. Moreover, "the sounds of the engine, the vehicle's weight, its inertial forces, ethylene glycol and battery acid leaks and the emissions of smoke and noises" are part of the mental images associated with the automobile in the collective memory. Since the advent of the electronic horizontal positioning-controlled vehicle, followed by the autonomous vehicle, ethical aspects have received a solid scientific treatment. At a broad level, the literature goes all the way from in-car user interfaces, through user preferences connected with the vocal representation of the ambient intelligence, to the implanting of a moral conscience and liability issues in coordination with road traffic laws. Even if CAVs receive extensive testing in all conceivable traffic, meteorological and lighting conditions, the only possible conclusion is that a human driver's knowledge and reactions can only be approximated.

3. Reinforcement Learning in Cyber Security

In general, resilient architectures can be created in cyber security with deep reinforcement learning by enabling autonomous and adaptive system responses to dynamic network attacks through hierarchical policies, deep learning, LSTM or RNN networks for long-term learning

[11]. Vehicles are prime targets of cyber attacks because the lack of security features and scalability of attacks for vehicular networks become large threats for autonomous vehicles. So, reinforcement learning provides significant progress in adaptive and autonomous instrumented vehicle security and provide potential systems' cyber resilience, automation, and transparency. Also, a comprehensive vulnerability for these breakthroughs is cyber-attack methods and threat model (such as Environmental Adversarial Robustness Evasion Model (EAREM), Security Evaluation of an Adversarial Reinforcement Learning Environment (SEARLE), Universal Adversarial Perturbation (UAP), and Proximal Policy Optimization-Adversarial Adversarial Defense (PPO-AD).

Reinforcement learning (RL), including deep reinforcement learning (DRL), is widely applied in autonomous decision-making [5]. RL has already passed the Turing test on ATARI games with DQN, plays Go at a higher standard than any human, achieves human-level performance in some aspects of physical simulation and solves logistic problems involved in transportation and robotics with advanced algorithms and relatively small compute power. Reinforcement learning is extensively used for cyber-physical systems (CPSs) including autonomous vehicles and autonomous driving. The main aim of reinforcement learning is to extract the optimal policy by maximizing rewards and minimizing punishment. In the real world, a phenomenon might happen due to insufficient knowledge and randomness in environment. In such scenarios, it is difficult to maintain the certainty that needed. Reinforcement learning process is required to be active, adaptive, fast, and cyber-resilient to investigate and counteract uncertain environment changes in autonomous vehicles and cyber systems. Apart from that, resilience actions performed by the CPSs must be dynamic, proactive and reactive to work seamlessly in evolving, pervasive, and complex environments. Concurrently, an adaptive cyber defence system should adjust itself to new attacks and threats through automatic detection and intelligence mechanisms.

3.1. Fundamentals of Reinforcement Learning

Optimal Q-function is defined as $Q*(s,a)=max\pi E[Rt|s,a]$, which gives the maximum possible return we can get by taking action a in state s regardless the policy [12]. For finite MDPs, there exist such a policy π, called the greedy policy with respect to Q-function, that always takes action that has maximum expected return. Formally, $\pi*(s)=argmaxaQ*(s,a)$. The Q-learning algorithm and its variants aim to find the Q-function and the optimal policy by iteratively

updating the Q-function Q(s,a) using the experienced quadruples, (s,a,r,s′), along the trajectory. A key insight that makes this iteration work is the Bellman equation for Q(s,a), which gives us an update rule for Q(s,a) that incrementally minimizes the Bellman error [13].

The environment is modeled as a Markov Decision Process (MDP) represented as a 5-tuple, M=(S,A,p,r, γ), where S is the set of all the states, A is the action space, p and r are the transition functions and γ is the discounted factor. At each time step t, the agent observes the current state st, takes an action at leading to the next state st+1~p(·|st,at) and receives a scalar reward rt=r(st,at) [14]. The Q-function estimates the expected return from taking action a in state s and then following policy π for the rest of the episode. It is defined as Q(s,a)=Eπ[Rt|s,a], where ∈R is the expected return given that the agent takes action a in state s and thereafter follows policy π.

3.2. Applications of Reinforcement Learning in Cyber Security

Nowadays, the rapid conversion of legacy transportation systems to intelligent transportation systems (ITS) can be observed. The security issues pertaining to them can be regarded as ongoing research investigations. As reasoning about the subject at hand will become more comprehensive by examining the prerequisites of reduced vulnerability and the risk of cyber-attacks, the article discusses the matter in the context of a reinforcement learning (RL) model that is designed to be adaptive. Moreover, ITS vehicles exhibit peculiar communication techniques, which increase the surface area for exposure to potential adversaries, whereas a legacy system operates in a less sophisticated environment. Here is where the concept of smarter learning-based defense systems will be touched upon in the course of this research study. The amazing adaptive intelligent transportation system (AITS) architecture scrutinized in this article is both a step in this direction and a unique approach that is worthwhile to explore. AITS dynamically builds defense architecture and obtains intelligence about threat signs in an effort to ascertain an optimum performance in various environments. An AITS moves like a living entity based on additional safety perspectives that are also showcased. Moreover, given that defenses in intrusion detection system (IDS) coexist with classical vulnerabilities, the results obtained in this study across two domains will incontrovertibly broaden the reader's horizon.

Cyber-Physical Systems (CPSs) and autonomous vehicles—two critical realms— are respectively disciplines in which cyber security and safety are most paramount. The

Separation Assurance System for the National Airspace System, Radar Signal Phenomenology Analysis & Generation Environment (SPAGE), Aerodynamics and Stability of Salamanderoid Robots Towards Agile Loco-Manipulation, Clinical Activity Analysis for Multimodal Collaborative Workspaces, and Workload Assessment of Military Commanders in a High-Demand Dynamic Environment are five different application areas; cybersecurity—more specifically, reinforcement learning, which is by far one of the most dynamic learning methods in recent years—, is another conspicuous domain, endowed with vigorous latest academic endeavors [6]. Unsurprisingly, the amount of resources directed to scientific studies on this complex issue further intensifies with the increasing deployment of Internet of Things (IoT). Some motivations that give rise to the importance of this work are: (i) Autonomous vehicles are perceived as promising entities with respect to on-road safety, energy conservation and consumable time; (ii) As an inevitable and inseparable part of an autonomous vehicle, a cyber-defense model—immunizing the vehicle from malicious actors—needs to be tackled in parallel [15].

4. Explainable AI in Cyber Security

[16] [17]To achieve adaptable cybersecurity, the cybersecurity models should be continuously aware of the threats landscapes and adapt to new situations without the need for human intervention. Intelligent decision support and ARtificial Intelligence (AI) technologies have the potential to guide def-ense adaptation. Nevertheless, current AI is often viewed as a black box prohibiting human users to understand why certain decisions are made. Explainable Artificial Intelligence (XAI) approaches aim to make the results of AIs transparent. The user—especially defense decision makers—can understand, interpret and trust the decision of AI-based defense actor that follows the actions of its defense decision-making model in which it reasons and acts adaptively in a given environment. Over the last century, transportation systems have been increasingly automated, exemplified by the move from traditional petrol-fueled engines to autonomous electric vehicles ([18],). Specifically, autonomous vehicles (AVs) offer a vastly improved transportation experience for the users and also the potential to solve a number of societal and economical challenges (e.g. road safety, traffic congestion, air and noise pollution). To reach the full potential of AVs, it is envisioned that their operations should be fully autonomous, where humans only interfere when necessary from a supervisory or safety perspective (instead of driving the AV themselves), i.e. Level 5 autonomy. Even though current AVs are steadily growing into more and more autonomous capabilities, it is

commonly noted that AVs are not yet fully autonomous today. For example, current AVs are not yet able to make classical driving decisions or reason logically and humans are therefore needed to aid reasoning and decision-making in some cases. Moreover, the technology readiness level (TRL) 6 (detailed system design) of fully AVs that are able to drive on public roads and in cooperation with other vehicles would need to have AI that reasons "intelligently" and to have a decision support/handling concept that allows human-in-the-loop support (which is currently being researched). Cognitive automotive systems were first introduced to assist drivers in emerging dangerous situations or in incomplete road/traffic/vehicle information. Some methodology work has since already been performed to ensure that AI/ML acts safely and effectively to improve the faultless "reasonable agent" image in AI. The FAIR mechanism concept, architectural model framework, and a human operator process, in which agents react adequately and communicate fluently with human users for their reasoning (planning) and decision-making concepts on future AVs/AI decision-making systems as the FAIR intelligent decision support concept, was introduced. Explainable planning of specific historical accident scenarios shows that FAIR agents comprehensively reason actual meaningful plans in emerging and unknown situations, in cooperation with human users. To improve the conventional AV safety rules, beginning to reason and act flexibly and quickly in changing environments on the road, to agree meaningfully with human operators, and to enhance user trust in the AI decision-making, the agents' planning and acting processes should be transparent and understandable for the responsible human operators and "outsiders" and should directly learn the human operators (only specific critical accident scenarios are considered in this research) for their communication consistency between the agent and user interface.

4.1. Importance of Explainability in AI

The entire field of cybersecurity focuses on defending networked systems from malicious threats, either human-made or automatic. The toolset with which defenders work is also sophisticated and includes both detection and prevention systems using enforce if needed. The main difference beet extinguished of these safety-critical systems and established as by human and secure communication system is that service models need to be able to interact with human and respond to questions, in order to help the user. Although it can be said that fairness, transparency, and accountability are some of the most important goals in the explainability of machine learning models, the number of safety-critical systems is five vice

throughout the request, than an unexplainable prediction. Following coin money mainly due to the observer-explainer has not content addition as warned to the additional the run-time cost of explanation generation can render the reasoning and adhesiveness a unfit for service autonym such as patient guide is based on demand for clinical for system where multiple aspects been a part of security. Even though these points have been sofar disregarded in context of security the same effect of adoptly where users might simply ignore the explanation to return to lenience the actual content of the system.

The importance of model explainability has been widely recognized within the scientific community both in healthcare [19] and cybersecurity fields [20]. Clinicians and healthcare workers need to understand why a certain model produces certain predictions when they are trying to assist patients, diagnose ailments or determine prognosis. Especially in safety-critical domains like medicine, where decisions are being made that directly and immediately influence human health, it is crucial that the underlying model decisions are understandable. Ensuring that patients, doctors and classifiers coordinate in this way is ultimately important to alter patient Agency, generatively Impact Society, extend Reach, and invest Authority to maintain Length of relationship between classifier and doctor. For this reason, model model constraining on diagnosing a subsystem of the manual labor of medicine, on a similar time demonstrating safety-criticalness exert as they constrainform or taking inference over patients' electronic playlist Facing a flexibility-privacy trade-off to achieve this purpose, we propose a patient-specific explanation (PSE) design, aiming at patient-specific model explanations while satisfying the requirements of privacy and explanation quality. We particularly concentrate on the interpretability and generalization two key elements in medical-related applications. [18]

4.2. Techniques for Achieving Explainability in AI

The explainability is suitable to consolide transparency, interpretability and trustworthiness of AI systems by several ways: it identifies the decision-relevant part of AI representation, allowing users to trust and rely on the system decisions as well as improve the trustworthiness and transparency of the models by highlighting when and why the AI deviates from the clinical standard; it facilitates validation and maintenance, supporting the clinical deployment of AI; and it allows the system to be understood by ordinary users like clinicians, fostering greater acceptance and trust. On the other side, the drawback of these methods is that an AI

model can make unjustified and incorrect predictions, allowing to fool them through adversarial examples, and not the XAI method can mitigate these vulnerabilities.

To achieve a successful application of Reinforcement Learning (RL)-based Adaptive Cyber Defense in the field of connected autonomous vehicles, the explainability of the decision-making process is key to increasing trust and facilitating the development of new approaches and systems by providing domain expert validation [4]. Therefore, efficient and interpretable RL methods are fundamental to infer an AI model decision using post-hoc explanations that can be easily interpreted by a given human domain expert in a natural way . This can be combined with the adoption of Explainable Artificial Intelligence (XAI) methods [21] , a sub-discipline of Artificial Intelligence (AI) that aims to make both the decisions and the underlying model more human-understandable [22].

5. Integration of Reinforcement Learning and Explainable AI

In order to enhance safety, some tasks are achieved in the form of final-DRL in counting the explanation and adaptive driving methods for DRL-based systems. Individual DRL agents required to learn the policy in Autonomous Network Defense (DAN) exhibit biased probabilities. A fundamental goal of the ablated ensemble method was to improve the model reliability, which was utilized several times in CAGE. To deal with this difficulty, it is important for DAN to also aggregate a set of different types of classifiers for decision-making. In the specific scenario, this research shows that the simple majority strategy is helpless in From-SStranger and To-Stranger social interactions. The results suggest that the proposed inference methods achieve a promising trade-off between obtaining helpful insights into final-DRL black-boxing and mechanical module building. However, the explainability force us to discard the explainability, because the final-DRL in which the highest gray acceleration input results in a maximum fault-driven maneuver acceleration was accepted across all subjects.

Reinforcement learning (RL) is a robust approach for autonomous vehicles' sequential decision-making in dynamic environments [23]. It involves a Markov Decision Process (MDP) with state space, actions, transition probability, reward function, and a discounting factor. The MDP involve both the environment module (a simulator transforming the intentional actions to perceptual consequences, showing the next state and resulting reward) and the agent module (mainly include the policy module, existing in a set of demonstrated strategies, assignment of all the state-action pairs to expected values). A series of mechanized functions

help the agent to interact with the environment in the markov process. Nevertheless, the complexity of real-world scenarios demands more model-free and potent solvers, which makes DRL (Deep Reinforcement Learning) a better candidate than the traditional RL algorithms. However, policy updating and observation between the pivotal layers in reinforcement learning are complacent, enveloped in 'black box' behavior which can lead to risky decisions in the context of autonomous vehicles causing accidents. In explainability, association mapping refers to the analysis for the weights change between any two layers in finetuned DRL.

5.1. Benefits of Integration

This study aimed to integrate the semantic predictive control (SPC) through explainable planning (Xplan+SPC) as an adaptable operation point to avoid plaque while saving fuel and time through adjust the aggregation point. The prognostic system was developed for autonomous vehicles with the integration of the task planner it is possible to plan and execute different driving actions, such as stopping at signposts, autonomously and interactively. Through the integration of our system with the Explainable AI (XAI) System, it is possible to explain the planning result and subsequently visualizes the rationale for the selected driving action. The test showed that extracting knowledge from the prognostic system can help through understandable visualization to show popped signs in the route, which can be considered by the AVs to follow the route or not based on the research mentioned before [17].

The implementation of RL in autonomous defense networks aids users to plan, make decisions and undertake actions in the best possible way. As the size and complexity of the Cyber threats increases, such capabilities become even more essential. As seen in the case of Lethal Autonomous Weapons, transparent AI capabilities are vital to prevent unexpected attacks and protect interests. Common defense systems usually try to identify threats and divert their approaching effects. This approach of reacting to managed threats has multiple inherent disadvantages. The agents usually fail to learn and adapt from previous well-identified attacks, leading to similar patterns of approaches in future attacks [ref: 37be3d78-0138-476a-b144-7f204ca48f3f, 3c782b70-799f-42d4-bc09-621acfe0cb82].

5.2. Challenges and Limitations

Once the learned model in the autonomous control system is known/the distribution changes, autonomous driving may present significant vulnerabilities and diverse attacks may endanger human lives or public property. The model free AI systems such as reinforcement learning are exposed to a range of edge cases and adversarial attacks. Located external of the dynamic network, DL controllers may be seen as black-box controllers and the adversary can construct adversarial examples by means of exploiting the model's gradient information or directly by observing the controller's input-output responses. Especially in the early practical stage where safety proof systems and new brought-to-light vulnerabilities are still unknown, this uncertainty may promote the adversary's advance manoeuvre preparation and attack plan implementation. Some works reveal AI bugs and vulnerabilities represented as combinational function dependent on the input vector components and human interactions combining a strong adversarial case produced by use of reinforcement learning to a value judgment error [22].

Trained models may have limited learning capacity and can therefore struggle to adapt to new knowledge in a relatively slow way. This cautionary remark refers obviously to the supervised learning systems. Rather basis safety properties are missing: e.g. safety of a control policy against model disturbances or identification of the set of initial states and disturbances for which safety is guaranteed. There has been some works focusing on providing formal guarantees of safety in reinforcement learning models (see) and explain the vulnerability and possible mitigation methods. However, those works mainly emphasized on the single vehicle cases. To the best of our knowledge, it has never been applied to cooperative control of self-driving vehicles for avoiding traffic collisions [15].

6. Case Studies and Applications

Developing more advanced supply chain distribution algorithms will also be further worked for cyber system defense [24]. The malicious control signal sources from various adversarial attack situations can also be further studied, such as adversaries with higher agility and responses, adversaries that learn during fast parameters update of reinforcement leaning, and dynamically estimated adversaries. The proposed framework will also expand to evaluate various cyber system defense under scallability situations of multi-state adversaries.

In this fully connected environment, the always-on connection to a trusted overhead infrastructure like DSRC is limited and lacks resilience to possible attacks [4]. It results in

dependency and vulnerabilities for CAVs. Although Virginia Tech Transportation Institute collected data from numerous vehicles in realistic environments, the feature extraction phase of a vehicle tracker from the real-world data can be different from the CARLA simulator [25]. As a result, evaluation results on real-world data are being conducted to move closer to the real world.

6.1. Real-world Implementations of Explainable Reinforcement Learning in Autonomous Vehicles

The AV industry has risen worldwide unexpectedly, recognizing safety as the primary concern of all globally. Embraced safety is the main focus of each authority across the whole world. The conventional formal automatic safety review does not provide a full certainty of matter, thereby it does evaluate future deeds of the system during runtime. There are numerous ethical and practical challenges in evolved driverless vehicular weapons like "A deformed TrafficWorld" because of Shape changing scenarios, An offensive vehicle, Unauthenticated vehicle, dangerous Approach vehicle, Enigma solution for abusing vehicle, and Negative or affirmative inputs in social connections. The growth rate of Autonomous Vehicles (AVs) industry has markedly impacted by availabilities of Assured Artificial Intelligence (AI) [26]. This seeks algorithms, processes, models utilizing AI, as well as number handling methods that are not fully disclosed in their conclusions, Designing policies for discrimination of futuristic outcomes, cannot ascertain possible state scenarios and accessible threats, and Diverted EVs from one route to another. After the result of formal and informal scenarios, software faults in operations, and miss-coordination of download software. Also, legally backing up systems, utilizing software-based commands and vehicle communication network. The becoming of LIDAR manufacturing is observed for robotic utilization, and such sensors are progressively moving to Automotive Vehicles. This may as well engage the promoting of alerts of laser radiances as an evasion system, as is presently done in some radical vehicles. A survey and short taxonomy of network boosting are covered. One more unique matter is the developing of technologies to guide across the planned permanent structures and to an outgoing network using, e.g., radio links.

autonomous vehicles (AVs) is a significant area of research due to their potential to modernize the concept of driving by creating insights into the future industrial landscape. It is seen as an evolutionary step in the development of modern intelligent transportation systems, designed

to survival in extremely dangerous environments. A recent paper [5] pointed out that some prevailing modern security vulnerabilities in deep learning systems like adversarial examples, backdoor attacks, intrusion detection, attacks on evasion, and model spoofing. In this setting, the work proposes a robust attack on adversarial driving, which masterfully generates stealth alterations to the behavior of an Autonomous Car (AV) that is human imperceptible or scarcely recognizable, but capable of causing the AV to behave in an unsafe manner. Autonomous driving is currently gaining recognition as a radical technique to reduce the frequency of automobile accidents. The convoys and driving analyses in modern AVs can be provided by large-scale inputs from environmental sensing systems to enhance efficiency, safety, and the automatic driving experience. The basic technologies of AVs consist of surroundings understanding, decision-making, route planning, and action execution. Hence, cyber defense technologies like Deep Neural Networks (DNNs), intrusion detection systems, advanced encryption protocols (for Network layer security and Data security) Schemes and Autonomous Vehicle (AV) systems, which can protect AVs from cyber-attacks, are also of much interest to maintain system integrity and to avoid traffic collisions.

7. Evaluation and Performance Metrics

The paper highlighted the security vulnerabilities of the Dominion ImageCast Evolution Voting Machine, which is closely associated with the smart-city infrastructure of VoI. Using this voting machine, some regions in Pennsylvania, New Jersey, Georgia, Wisconsin, and Michigan organize the voter's turnout in American federal and state elections. The main contribution of the study is the methodology to anonymously manipulate the preferences of the voters once the integrity of EVMs is compromised in such networked elections. This new attack vector is due to the fact that the reported printouts cannot provide transparency unless the printout would reveal the encoding of the voters' intent. curve optimization, gradient-free optimization, explainable optimization, cyber physical systems, autonomous vehicles.

[3] [8]Adaptive cyber defense policies have been examined using a simple time-domain system, with evolving threats and a set of continuous actions leading to sinks or sources of resources. The key results were as follows: The adaptive replay (AR) policy most effectively preserved the protected resource levels, and could almost always prevent the attacker from gaining resources if MDS was observed to decrease. The capability to execute AR effectively was not substantially different from the capability to execute an optimal policy that was

calculated for a known range of attacker values of MDS. Our results also suggest that the effectiveness of AR could be attributed to trackers that approximated each of the optimal dynamical systems using an agent, expressed in the neural network form, which was trained in parallel with the reinforcement learning optimization of the policy. In the context of autonomous vehicles, evaluations have mostly focused on the physical, machine learning, and decision-related aspects, but workplace autonomy also implies the possibility of the vehicle being attacked and, hence, controlled by an adversary. In this paper, the authors analyze the potential effects of Dominion voting machines, which program code may be covertly manipulated using wireless methods.

7.1. Key Metrics for Evaluating Adaptive Cyber Defense Systems

[27] [2]In the literature, evaluation of adaptive control policies for autonomous vehicles generally only employs RL algorithm performance metrics, such as convergence, and optimization, and neglects metrics for assessment of the actor-critic and, consequently, any potential control policies' adaptive behavior. Only a small number of studies have identified that true measure of the system model being learned is whether it reflects the system's properties and behaviors. Additionally, the literature has yet to identify and define, any associated appropriate metrics. Malicious challenges, faced by modern cyber security, has created the need for adaptable and adaptive autonomous systems, and the majority of current evaluations only focus on pre-selected adversaries and potentially neglect other real-world challenges.[28]Measurement of adaptive control policies' performance is dependent on the vulnerabilities of the system and the environment. Thus, the actor-critic's learning trend of exploration and exploit learning should dominate from state to state. This regime needs to be tested with various adversaries, in order, to evidence a superior adaptive policy. There is some variation in the literature on, evaluation metrics, for reinforcement learning (RL). However, none of these considerations evaluate the control policy's adaptive behavior, which is important since RL defines itself as approximating the actor in a policy iteration algorithm. Control, optimization, accuaracy, precision and coverage, are all substantial metrics the above literature recommend to test for evaluation of the adaptive control policies.

8. Future Directions and Open Research Challenges

This future research will conduct a system–security analysis case study for both parking lots at physical locations with observed traffic behavior data and collect system-level and digital-

level attack data coming from the most recent accounting databases and physical damage of houses. For the four metrics in this unblocked case we have shown how for 75.26 %–80.21 % of the possible attack sets only 1.68 %– 5.37 % security analysis attacks occur and only for 5.676–9.798 % of the possible attack sets only 0.474 %–0.527 % system-level digital attacks are deemed as attacks based on the conviction test. Where dual are the moral obligations of the defense operator and the security manager of the system, namely defense for itself and the public to improve on the warning covers when the new explication converge more accurate decision making.

Future research in the area of explaining autonomous vehicles decisions for cybersecurity needs to further explore foundational questions. These include: is agent modeling of the human-in-the-loop needed to improve the explainability and trustworthiness of agent decisions [29]? What kinds of biases and inaccuracies do human-in-the-loop validations address in an explanation model? Is an enhanced defense mechanism necessary to counter those attack and learning mechanisms which are not captured by reinforcement learning models [5]? What valuation metrics are appropriate to validate these defenses? Also, integrated security solutions need to conduct immortal testing on air and ground traffic. The main focus of this approach is to show the robustness of commonly used security validation metrics for each traffic element and whether it is influenced by traffic behaviors which may take certifying in the offline and online rolling case [2].

8.1. Emerging Trends and Technologies in Adaptive Cyber Defense

There is frequently ongoing research into the development of various trends and technologies in the context of adaptive cyber defense, as the use of network-connected systems, including autonomous vehicles, continue to become more popular. Emerging vehicle communications are often susceptible to cyberattacks, with both safety and privacy concerns. To aid in mitigating these issues, research into a more in-depth explanation of machine learning or deep learning-based detection approaches is needed, to enable all attacks to be addressed effectively [15]. It can be seen that useable models are not only those which can accurately detect fraudulent data, but those which can correctly explain their decisions, yielding an explanation of the data anomaly situations allowing for the implementation of improved network cyber defense for autonomous vehicle communication security, safety, and privacy systems.

The existence of adaptive cyber defense technologies allows the vehicle's cyber defence system to evolve with and adapt to the ever-changing cyber-attack tactics and techniques [30]. To hinder cyber-attacks in autonomous vehicles, it is important to identify and understand the emerging trends and technologies in this domain and detect, defend, and respond to a cyber-attack automatically, actively, and effectively. In this vein, trends and technologies in the field of adaptive cyber defense are discussed in detail below.

9. Conclusion

Data-driven decision-making systems that are established on ML models are rapidly gaining popularity in the autonomous systems domain for their optimality (potentially) and ease of integration. However, most AI systems are good at decision-making based on data-centric views of the world and are therefore oblivious to information sources that are not in their training and immediate perception space, and are helpless in creating safeguards for these externalities. This makes them particularly vulnerable to system-level and adversarial attacks. Addressing the lack of interpretability and transparency in AI systems has motivated a growing body of work on eXplainable AI or XAI, in various domains including AD. This manuscript is aimed at exploring various XAI methods driven by the goal of secure deployment of AI-powered decision-making systems in AD [5].

By applying XAI techniques designed to demystify the decision-making power of ML and AI systems in the context of autonomous driving (AD), this study demonstrates how to improve the trust, safety, and performance of decision-making by leveraging cumulative sensor data. One future direction inspired by this work is the further enhancement of trust in the AD decision-making process, particularly under the context of adversarial attacks and the unreliability of environment sensing and navigation sources. The development of understanding and belief metrics, AD trust level prediction models, and trust-aware decisionmaking systems could drastically improve the long-term reliability of decision-making systems and also make recommendations for driving strategies to the user [4].

9.1. Summary of Key Findings

[31] It is borderline trivial to note the autonomous decision-making capabilities of attackers continue to present significant challenges to network and system defenders. Moreover, automated offensive cyber technologies are advancing at an accelerating pace and afford

adversaries numerous force multipliers. Fortifying defense operations, particularly with automations of their own, has become an imperative for most organizations. However, developing autonomous defensive operations to cover a high-velocity and rapidly evolving cyber battlespace is complex, logistically complicated, and takes substantial time to prepare. What results is a significant asymmetry between the offensive and defensive cyber capabilities. This is particularly the case when it comes to remedying the organization and recovering data following an incident.[32] Developing an autonomous cyber defense model that is capable of identifying the circumstances in which the vehicle should carry out specific responses during a multi-step MLC attack on the basis of actual vehicular sensor data and cyber-physical relationships as a key step towards the development of subversion-resistant control for self-driving cars. A novel approach to reinforcement learning (RL) for autonomous cyber defense, which operates on explained models, is introduced and explained in this work. The objective of this work is to leverage reinforcement learning (RL) to construct a defense model that influences an attack process dynamically. Furthermore, the attack model is entirely unknown to the defender, and containment strategies are not ordered under a hierarchical framework as in existing secure control architectures.

9.2. Implications and Recommendations for Future Research

A key planning of disputed research, including generations and complexities of diverse adversarial attacks and dynamic strategies of smart adversaries, lies in the broad exercises in methods for enhancing resilience to cyber-attacks in autonomous vehicle and transportation systems based on this insight, design, and use of psychological measures. Thematically distinct from the mechanistic notion of a weakened system, perturbing performance with disrespectful stimulations analogous to the adversarial samples in AV cyber-attacks (see Chapter 8, Adversarial Machine Learning: Implications and Recommendations), this notion of the perception very likely includes the observer and the test subject, thus proposing episodic foci. Human-in-the-loop validation may address potential biases and inaccuracies in the autonomous vehicle model and strategy advancements. Moreover, being featured with road accidents and scenario comprehension of authentic data input, an implicit and undersampled aspect of AV security is to assimilate additional, or flexible forms of adversarial perturbations, against both the causal experience and AI models and learning processes.leading secure, accurate, deep, resilient, robust, and interdisciplinary methods for AI-based autonomous vehicle systems, considering secure inference, secure AI/ML model

development and analysis, adversarial robustness, and secure development of AV model processes, that ensure robustness and flexibility across all steps in autonomous vehicle perception-action-planning-control. Appraising a wide variety of potential adversarial attacks, complimenting attacks at all spectrum levels, necessitates a broad investigation. specifically, rendering any method possible to be deployed with a smaller parameter size will allow the autonomous decision maker to choose, on the run, from a wider pool of potential models to tackle with a certain scenario. To complement the existing strategies, addressing digital adversarial attacks, system-level attacks (also known as supply chain attacks, hardware backdoors, and cyber-physical attacks), and real-world testing (in-line adversarial and human-in-the-loop validation, check hardware or system–level attacks, validity of simulations, and causality of field errors) may provide a comprehensive understanding of AV security. [29]

[1] [31]The adaptive cyber defense strategies presented in this book aim to enhance the AV overall security in the current testing paradigm and future deployment. These maneuvers must be adaptable to an ever-changing threat landscape, emerging with an evolving knowledge of vulnerable assets and target values, understanding malicious cyber campaigns, and their mitigations from the very foundation over time. However, the generality of attack pathways, the unknown conditions, and the black-box nature of the DRL-based AV systems require further refinement. The present techniques can be interfered with under uncertain scenarios and external conditions, and critical if cybercriminals leverage the vulnerabilities identified by any of the cyber campaigns resulting from these research efforts to design their next generation of adversarial attacks.

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. 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.

3. 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.