Toward Trustworthy and Responsible AI Compliance for Practical Drone Deployments

Abdul-Rasheed Ottun[1], Akintola Adeyinka[1], Zhigang Yin[1] Mohan Liyanage[1], Mehrdad Asadi[5], Michell Boerger[2], Pan Hui[3,4], Sasu Tarkoma[3], Nikolay Tcholtchev[2], Petteri Nurmi[3], and Huber Flores[1]
[1]Institute of Computer Science, University of Tartu, Estonia
[2]Fraunhofer Institute for Open Communication Systems, Berlin, Germany
[3]Department of Computer Science, University of Helsinki, Finland
[4]Department of Computer Science and Engineering, HKUST, Hong Kong
[5]Department of Computer Science and Engineering, Vrije Universiteit, Brussel Brussel
ottun@ut.ee

***Abstract*—The maturity of autonomous vehicles (AVs) has reached a point where they can be used for tasks such as medicine and food delivery and environmental monitoring in cities. These operations rely on the integration of powerful and robust AI models into AVs for safety, as any error in the decisions of the AVs can cause damage to citizens and infrastructure. Our paper contributes a vision for trustworthy city-scale deployments of AVs, highlighting key requirements and challenges for the city-scale deployment. We analyse the complexity of using explainable AI (XAI) methods to monitor vehicle behaviour by inducing changes in AI model behaviour with data poisoning attacks. Our results show that XAI methods can detect such attacks, and the combination of multiple XAI methods can improve the robustness of the estimation. However, further research is needed to improve the XAI methods to better and more robustly identify the root cause of attacks.**

***Index Terms*—Autonomous Vehicles; Explainable AI(XAI); Data poison attack**

I. Introduction

The integration of AI into autonomous vehicles is critical for enabling operations that require minimal or no human intervention. Indeed, AI is essential for autonomous navigation, trajectory estimation, collision avoidance, and localization [1] to name but some examples. As these techniques have matured, vehicle applications that automate our daily life activities are becoming a reality, e.g., delivery of food or medicine and applications that harness vehicles for environmental purposes, e.g., air quality monitoring [2], litter detection and separation [3]; and water pollution monitoring [4]. The emergence of these application domains has started to pave the way toward city-scale deployments of autonomous vehicles, yet there still are challenges that need to be overcome before these deployments can become a reality.

City-scale deployments of autonomous vehicles, such as ground vehicles, cars, aerial drones, or even aquatic drones, are only possible if the operations of the vehicles are *trustworthy*, i.e., they can be guaranteed to operate safely without causing harm to the citizens, the urban infrastructure, or the environment. Ensuring trustworthiness, however, is a complex challenge as it requires testing a wide range of environments and dependencies between system components [5]. Traditionally trustworthiness has been analyzed using formal verification methods, but these are only suitable for systems that expose their internal logic. Modern autonomous vehicles largely rely on deep learning models which are black-box in nature, making it difficult to verify the decisions that are made [6]. The emergence of federated learning and other paradigms that perform continuous updates on the model becomes even more complicated, as continuous updates can significantly change the behaviour of the AI models.

Explainable AI (XAI) methods have recently emerged as a mechanism that can help to understand the behaviour of AI models and the factors affecting them. For example, XAI methods can be used to evaluate the quality of the data used for training [7] or to assess the suitability of the internal model structure [8]. While these methods provide a foundation for analyzing the trustworthiness of the AI models integrated onto vehicles, they alone are insufficient as it is difficult to distinguish between multiple factors that have a similar effect on the model. For example, a decrease in model performance may result from a malfunction in the vehicle, a targeted attack, or a situational data bias. XAI methods also require heavy instrumentation of the model, and they tend to rely on offline analysis of the model, making these methods unsuitable for ongoing deployments where threats need to be detected immediately to ensure the safe operation of the vehicles.

We contribute *a vision for trustworthy city-scale deployments of autonomous vehicles*. In our vision, illustrated in Figure 1, autonomous vehicles are widely used for societal and other functions and that they are instrumented with methods that enable diagnosing the AI and detecting possible threats that can jeopardise their perception of their operating environment. We reflect on the current research landscape to identify gaps and open challenges to establish a roadmap that serves as a catalyst for research. We also experimentally demonstrate the importance of trustworthiness by conducting experiments on the effects of data poisoning on autonomous functionality. Finally, we conduct a small-scale field test in ground vehicle based litter monitoring to analyze the performance of different XAI methods in identifying data poisoning attacks to gain practical insights that pave the way to city-scale deployments

of autonomous vehicles. The results demonstrate that adverse effects on the model can be identified but that the root cause may not be easy to uncover. This means that integrating diagnostics into vehicles can help to identify potential periods of misbehaviour, but at the same time further research is needed to extend current XAI methods to be able to identify the root causes of problems. We conclude the article by highlighting further challenges and future directions for research.

II. Key Challenges and Requirements

Realising the vision of large-scale deployments of trustworthy autonomous vehicles is currently infeasible as there are technical and technological challenges that need to be addressed. We next highlight some important challenges and reflect on state-of-the-art to identify research gaps.

Data bias and drift detection: vehicles that are deployed should continue learning over time to improve their operation. As the vehicles should operate in everyday situations, the data that they capture is prone to contain privacy sensitive information (e.g., face, speech or car registration plates) and hence this task is best accomplished using privacy-preserving techniques, such as federated learning [7]. Optimally this process should integrate data from vehicles operating in different parts of the environment as this helps improving the generality and robustness of the AI models. Unfortunately, the model is vulnerable to biases in the data and it can be a target of attacks that affect the vehicle's operations. As a result, there is a need for methods that can quantify the resilience of the model updates and detect abnormal or erroneous updates before they affect the vehicle's functionality. A key challenge is to separate between non-intentional malfunctions (e.g., camera failure), intrinsic data biases, and targeted attacks. Existing methods largely target one type of issue (e.g., drift or poisoning) without being able to separate between the different causes. Another challenge is to ensure the methods can operate at different temporal scale, i.e., can identify problems even when biased or erroneous data is aggregated with valid data and when the erroneous data arrives gradually.

Continuous model verification: Besides detecting issues in the data, trustworthy operations require analyzing and verifying the decisions the vehicles make [9]. For large-scale deployments (e.g., a city), the analysis needs to happen continuously and on-site as otherwise the effort needed for verification limits the scale of the deployments. Indeed, model diagnosis requires accessing the internal structure of the model, which typically requires instrumenting the source code of the vehicle, halting operations, and accessing the internal logic of the vehicle. This process usually requires taking the vehicle to a lab as accessing the internal logic requires bypassing internal security features. While XAI methods offer a partial solution for continuous verification, they similarly need access to data and the model structure. Hence, they cannot be adopted as a general solution. A partial solution is to integrate the XAI methods directly as part of the security features (e.g., as part of trusted execution environments), but also this poses its own challenges as the security features often limit available resources.

Model interpretability and resources: The performance of AI models is intrinsically linked to the resources and components integrated on the vehicles [5]. Over time, these components need maintenance, or may be upgraded to improve the operations of the vehicle. These operations can affect the model and result in unexpected behavior. For example, integrating a higher resolution camera on the vehicle affects the dimensionality of the input data and may contain more detail than previously. This can require replacing the model or at least re-training it. In terms of interpretability, this requires linking model diagnostics with physical components of the vehicle and being able to analyze and interpret the effects individual physical components have on the model's decisions. Note that these changes do not necessarily affect the input data. For example, vehicles can operate using different payloads which affects the weight and resource consumption of the vehicle. This requires integrating physical configuration directly into the model diagnosis. This is essential also for detecting unsafe operation, e.g., detecting unsafe payloads. Current methods are insufficient as they are unable to link model behaviour with physical characteristics of the operating environment. For example, while changes in input data can be detected with current model diagnosis methods, they are unable to detect whether these changes are a result of the physical configuration or external interference.

Network-group diagnosis: Effective large-scale operations of autonomous vehicles are likely to require cooperation between the vehicles as this is essential for reducing resource drain and ensuring optimal performance. Effective coordination results in dependencies between the AI models deployed on the different vehicles and understanding potential errors or threats requires analysing the combined logic of all vehicles working in tandem, e.g., swarm intelligence [10]. Current XAI and other model diagnosis techniques are tailored to analysing individual models and hence they can only be used if the vehicles have a global model that integrates the decision logic of all vehicles together. Note that this task is more complex than analysing the performance of individual vehicles as attacks or errors can affect only some of the vehicles, yet have an influence on all of the vehicles by compromising the coordination of the vehicles through the network [11]. Understanding the effects on coordination requires improved diagnosis mechanisms to analyze network formation groups, individual parts of the network (slices), as well as models of how targeted errors can affect the performance of vehicle collaboration.

III. Motivating Experiment: The Impact of Attacks on Autonomous Vehicles

Model diagnostics is essential not only for offering a mechanism to analyze and understand the behaviour of AI models but also for mitigating risks of external attacks that do not require access to the model itself. We next highlight the need for model diagnostics by drawing on an example from computer vision based object detection to demonstrate how external data poisoning attacks can result in abnormal model behaviour and potentially even break the AI performance.

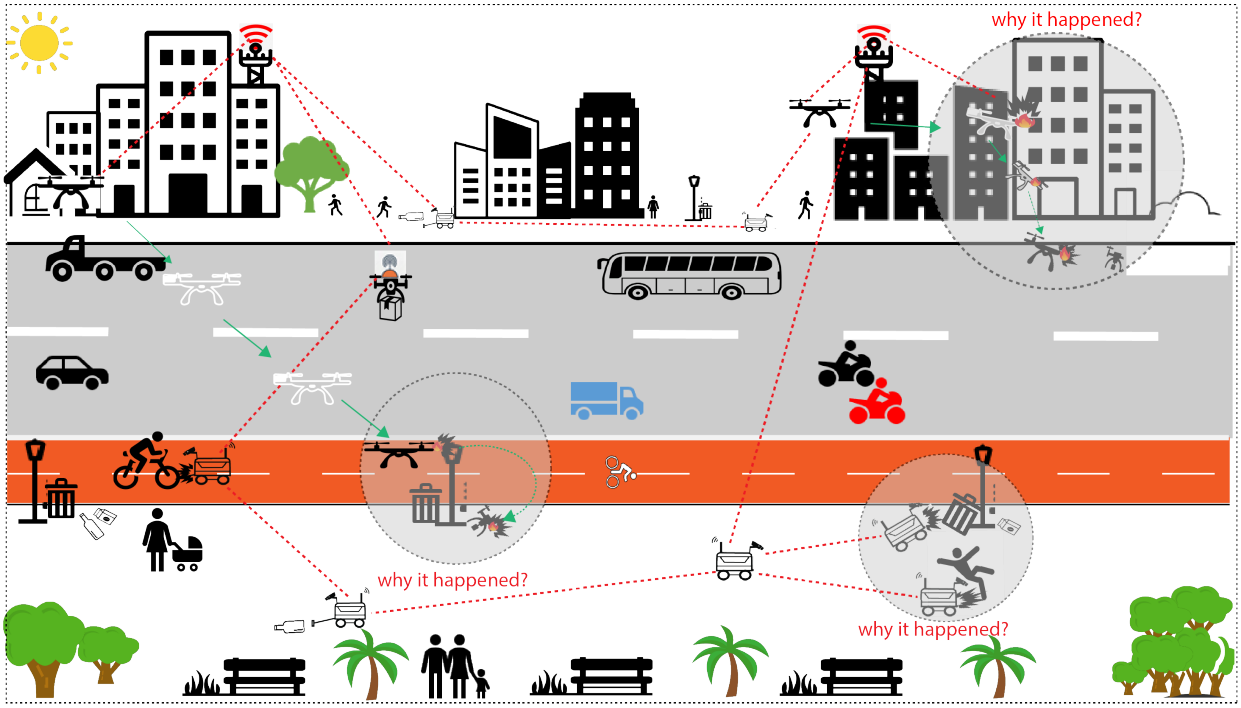


Fig. 1: City-scale deployment of autonomous vehicles and how vehicles can misoperate in urban settings.

Threat Model: We consider a generic threat model where the attacker attempts to case the AI model used in the autonomous vehicle to fail. This can be a targeted attack that results in a specific misbehaviour, e.g., causing accidents by making the navigation support to fail to recognize pedestrians or cars, or an attack that simply causes the AI to malfunction, e.g., a sponge attack that drains the vehicle's resources or a ransomware attack that prevents normal operations. The motivation for the attack can be causing damage or harm to the city or the citizens, financial incentive, or desire for fame.

Application Scenario: We consider litter recognition on autonomous vehicles as a representative example of operations that rely on AI. The application operates on thermal images which are analyzed in real-time to identify different litter objects and to determine their material. Specifically, the vehicle analyses the dissipation of sunlight-induced thermal radiation that is captured by a thermal camera integrated onto the vehicle [3]. Attacks against the model can break the operations of the vehicles or drain their resources. More serious attacks naturally would target navigation, obstacle detection, or other function that could directly result in harm to citizens or damage to the environment. Our use case presents a benign case that allows illustrating the risks of attacks without risking the citizens or the environment, and our findings are applicable to other AI applications that rely on computer vision.

Experimental Setup: Figure 2 shows our testbed and illustrates the use of thermal dissipation to analyze the thermal dissipation fingerprints of materials. We consider three common litter objects with different materials in our experiment: (A) Plastic bottle, (B) Face mask and (C) Cardboard cup. The vehicle records video footage of disposed litter which is pre-processed and analyzed to identify litter [3]. To attack

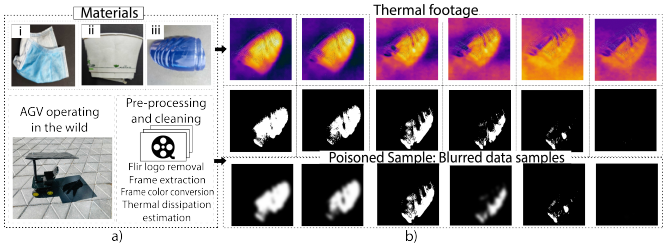


Fig. 2: Autonomous ground vehicle for litter identification using sunlight and thermal imaging, a) Prototype sensing litter in the wild and litter materials used in the experiment; b) Analysis of thermal samples and estimation of thermal dissipation time.

the model, we poison the data using blurring and steganography [7], [12]. While the basic attacks themselves are relatively innocuous on their own, they can be used to install backdoor triggers [12], to drain vehicle resources [13], or otherwise create unexpected behaviours. Note that these attacks do not require access to the vision system of the vehicle as they can simply manipulate objects in the environment, or use additional devices, such as lasers, to manipulate what the devices capture [14].

Results: We first calculate the thermal dissipation time of each material without attacks: plastic bottle 62.5s, cardboard cup 72.5s, and face mask 82s. The relative differences match those reported in [3] for the same materials. The absolute values differ due to the different intensity of thermal source, size of the material and the total exposure time. To analyze poisoning, we consider two levels of poisoning: 10% (low) and 40% (high). Higher values than 40% result in poisoning

taking over the model. For blurring, the dissipation times after poisoning are 51.5s (plastic bottle), 51.1s (cardboard cup), and 38.4s (face mask) for 10% poisoning, and 49.3s, 22.9s, and 40.5s when 40% is poisoned, respectively. The relative differences in the thermal dissipation values thus change significantly, breaking the AI model that is used for detecting litter materials. We also observed a clear increase in the processing time of the vehicle, thus resulting in higher resource drain. In the case of steganography, the thermal dissipation times were not influenced, i.e., the model is resilient against this attack. This highlights how the model response may vary depending on the attack type.

IV. XAI as Model Diagnostics

Explainable AI (XAI) methods provide a natural starting point for integrating model diagnostics on the vehicles and to overcome the effects of attacks. We next analyze different the potential of different XAI methods to detect targeted poisoning attacks, focusing on understanding the benefits and disadvantages of different methods. We first examine the quantifiable values provides by XAI methods in benign case, after which we analyze them against poisoned data. We separately analyze the full image and a processed image where the background is removed to better understand how different processing techniques affect the behaviour of XAI methods.

Experiment Setup: We perform the experiment using the TrashNet litter classification dataset which consists of 2527 litter images [15]. We rely on this dataset as it contains a large amount of real-world images which makes it possible to analyze different environments and contexts for litter classification. As AI model we consider a convolutional deep learning model (CNN) as this has been shown to achieve good performance on the dataset [15]. Images are resampled to 300×300 to have consistent input dimensionality. We augment the training data using horizontal and vertical flipping and rescaling. We train the dataset using 2276 images with batch size of 32 for each epoch iteration. The remaining images are used for testing and we separately consider a collection of 10 poisoned and non-poisoned images for illustrating the performance of XAI methods. Our experiment was conducted on the Google Colab platform using the latest version of the Keras library (2.8.0) with TensorFlow (v2.8.2).

XAI methods: We consider three model-agnostic XAI methods that can be applied for any type of AI model: LIME (Local the Interpretable Model-agnostic Explanation), SHAP (Shapley Addictive Explanations) and Occlusion sensitivity. As these methods are model-agnostic, they do not require any information about the CNN gradients to analyze model behaviour. These methods are also perturbation-based, which means that they manipulate the input (i.e., image pixels) to extract details that can be linked with the predictions. LIME creates an interpretable representation of the litter image by segmenting the image pixels, based on similarity, into superpixels [16]. SHAP explains the CNN's prediction by attributing importance values to the features that contributed to a prediction. This is accomplished using a binary vector of simplified inputs by perturbing the pixels (i.e. input space)

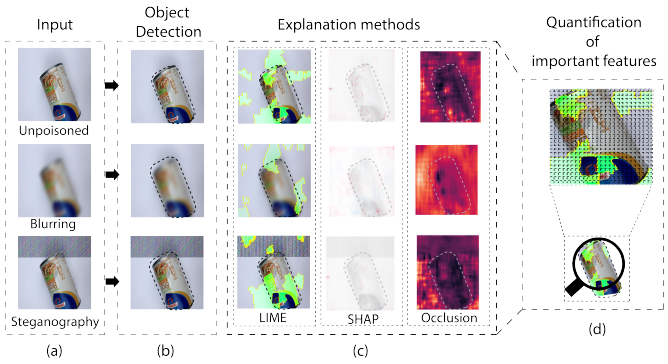


Fig. 3: Pipeline to analyze objects using XAI, a) Data samples (poisoned and unpoisoned), b) Object detection, c) XAI methods output over samples (LIME, SHAP and Occlusion sensitivity) and d) Object extraction.

of the litter image into superpixels (features) that represent the image [17]. The features that contribute to the probability of the prediction are highlighted in one color (red) and the features that decreases the probability of the predicted class in another color (blue). Finally, occlusion sensitivity explains predictions by using a sensitivity heat map to observe the impact a perturbation mask has on the neural network predictions. Specifically, an occlusion mask (i.e small gray square patch) is placed on the image and changes in the prediction probabilities are observed [18]. As stated, the XAI methods are applied separately for images where the background is removed (i.e., only the litter object) and for the original input image. The process for extracting the object is shown in Figure 3 and works by applying a dynamic patch (determined using object detection) on the image to isolate it. From the final output of the XAI methods, we calculate a pixel percentage metric that captures the importance of a pixel.

Samples and Poisoning: We considered six litter categories: glass, paper, cardboard, trash, metal and plastic. For poisoning we consider two attacks, blurring and steganography, as described in the previous section. Blurring can make autonomous vehicles to misidentify targets in urban areas, e.g., crossing signals and pedestrian sides. Steganography introduces extra information in the binary information of the images, which can become resources intensive for the autonomous vehicle as more processing power is required to extract relevant information (similar to a sponge attack). We systematically assess how the level of poisoning affects the results by poisoning the data in 10% increments from 10% to 40%.

V. Results

Model Performance under Poisoning: The performance to classify litter of our CNN is 0.7 when no data is poisoned, but this performance is reduced as data is gradually poisoned. After blurring attack, the model accuracy is reduced to 0.61 (10% poisoned); 0.53 (20% poisoned); 0.53 (30% poisoned) and 0.60 (40% poisoned). Similarly, after steganography, the model accuracy is reduced to 0.52 (10% poisoned); 0.52 (20% poisoned); 0.62 (30% poisoned) and 0.67 (40% poisoned).

In both cases, we observe a clear drop in accuracy. Unlike our earlier experiment, the performance drop is higher for data poisoned with steganography than with blurring. This difference in results is simply due to differences in the sensors (RGB vs thermal camera) and the processing pipeline and highlights how the effectiveness of the attack is influenced by the task and the specifics of the AI that is being used. The performance drop resulting from poisoning depends on how much the attack affects the patterns in the data. In general, once larger amounts of the data become poisoned, the inference process starts to be dominated by the poisoned patterns whereas in smaller amounts they result in distortions that can confuse the model. This pattern is observed with both attacks with the sole exception being blurring at 10% rate. The reason for this exception stems from small amount of blurring failing to distort the patterns of the litter object.

Analysis of XAI Methods: We next analyze the effectiveness of XAI methods to identify poisoning by considering 10 randomly chosen poisoned samples from each litter category and report the accuracy of estimating the correct class for each sample. Table I summarizes the results for the different XAI methods. The effect of poisoning depends on the litter category and the extent of poisoning. Paper and cardboard objects with regular shape are easiest for the XAI methods, whereas classes containing irregular shapes (metal, plastic, trash) showing highest variation in results. As with the results for the CNN model, in some cases a higher level of poisoning can result in smaller drop – or in some cases even in an increase – in performance. This pattern is more common for steganography as the poisoned data starts to dominate the inference process once a higher fraction of the data is poisoned. While XAI methods can only help recognize poisoning without directly enhancing the performance of the classifiers, they can indirectly offer insights that can help to improve the classifiers. For example, samples that are identified as poisoned can be used to develop data augmentation techniques which can be incorporated into the model training process to improve robustness of the classification models. To illustrate this point, blurring already is a commonly used data augmentation technique for improving the training of AI models. From our experiments, we also visually observed that the attacks tend to affect more the background and thus processing techniques that separate the foreground object from background are likely to improve performance.

Diagnosing Objects with XAI: Lastly, we examine the effect of data poisoning over the important features of the object when it is isolated from the background. As the metric we consider the coefficient of variation of the poisoned pixels, which depicts the ratio of the standard deviation to the mean. The higher the value of the coefficient, the higher the dispersion and thus the better the method is at identifying poisoned data. Figure 4 shows the results for the 10 test samples of each class. For the blurring attack, the average values of the XAI methods are 0.35 (LIME), 0.17 (SHAP) and 0.3 (Occlusion). For data poisoned with steganography, the corresponding values are 0.22 (LIME), 0.10 (SHAP) and 0.26 (Occlusion). One-way ANOVA between the three XAI methods indicates statistical significance, (F(2,1794)=118.4, p-value < 0.001), indicating that there are differences in the applicability of the different XAI methods. The higher average values of LIME and Occlusion indicate that they generally are better at identifying poisoned data. SHAP performs well for metal objects which are the most irregular, but struggles with other categories. We also used one-way ANOVA test to verify that the difference in variation across classes is significant across all XAI methods, poisoning attacks, and levels of poisoning (F(5,1791)= 14.76, p-value < 0.001). Across all XAI methods, the coefficients of variation are larger for steganography than for blurring indicating that XAI methods can also provide clues about the nature of the error. We also investigated the effect between attack type and data poison level. Two-way ANOVA test between attack type and data poisoning level indicates significant effect (F(1,4)=3.396, p-value < 0.01), i.e., the coefficients of variation depend not only on the attack type but also the extent of poisoning. Taken together, these results show that XAI methods help to identify the important features of the image, even after data is poisoned but their effectiveness is affected by the object, the type of attack, and the extent of poisoning generated by the attack. In any case, even when the objects can be separated and analyzed, this requires more processing and more elaborate processing pipelines which drains the resources of the vehicle faster and limits their operations.

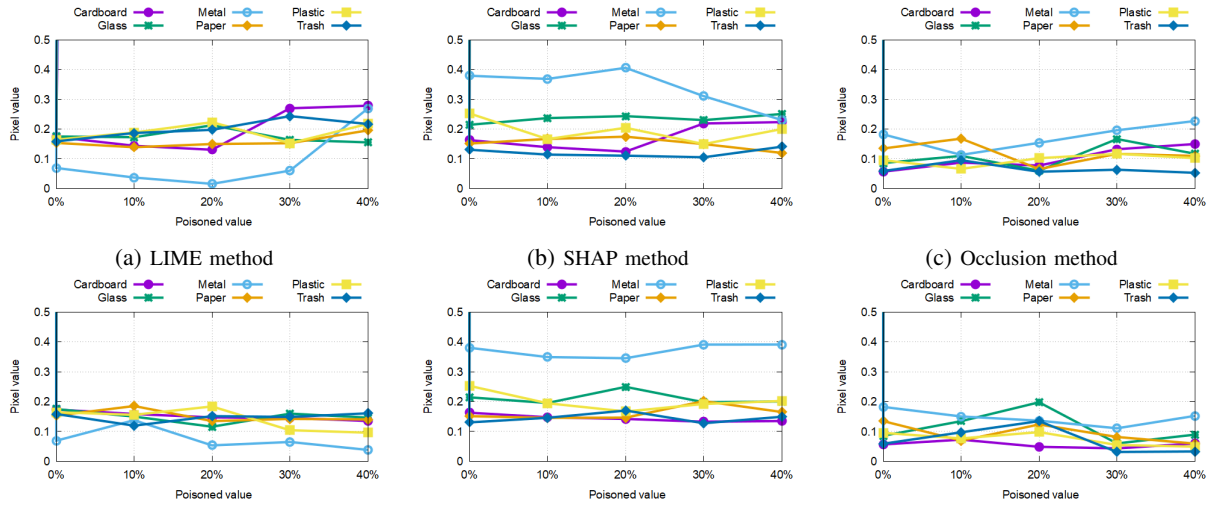
VI. Towards AI Robustness: Reducing Causes and Failures

AI regulations stipulate a range of desirable properties for ensuring AI trustworthiness within societal contexts. These properties encompass properties such as robustness, safety, privacy, fairness, accountability, and explainability. However, the rigorous evaluation, analysis, and validation of these properties to confirm their manifestation during deployment is challenging and complex. Compliance with the requirements for trustworthy AI involves trade-offs due to the intricate nature of prioritizing and balancing these multifaceted properties. Among these properties, AI robustness is a very crucial regulatory stipulation. It mandates that AI systems exhibit resilience against diverse challenges and adversarial conditions. Evaluating and validating AI robustness in light of the trade-offs is no doubt tedious. We explore some approaches, methods and technologies for fortifying robustness in AI during deployment in the paragraphs that follow.

Digital Twin: Leveraging digital twins (DTS) can enable the comprehensive observation and monitoring of AI behaviour to enhance AI robustness. Beyond mere surveillance, DTs can be utilized for logging and assessing the security of the underlying models running in AI systems. Utilizing information gleaned from DTs, we gain real-time insights into AI behavior, enabling early detection of potential issues and proactive intervention. Furthermore, empowering DTs with interactive feedback mechanisms facilitates efficient tuning and troubleshooting, further bolstering AI resilience.

	LIME					SHAP					Occlusion Sensitivity				
Poisoning Level	**0%**	**10%**	**20%**	**30%**	**40%**	**0%**	**10%**	**20%**	**30%**	**40%**	**0%**	**10%**	**20%**	**30%**	**40%**
Poisoning type	**Blurring**														
Cardboard	1	0.9	0.9	0.8	0.9	1	0.8	0.8	0.8	0.9	1	0.8	0.9	0.8	0.9
Glass	1	0.7	0.7	0.8	0.6	0.9	0.8	0.8	0.8	0.6	1	0.8	0.8	0.8	0.6
Metal	0.7	0.8	0.7	0.8	0.4	0.6	0.8	0.7	0.8	0.4	0.7	0.7	0.7	0.8	0.4
Paper	0.9	0.9	0.7	0.7	1	0.9	0.9	0.9	0.7	0.9	0.9	0.9	0.9	0.7	1
Plastic	0.8	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7
Trash	0.9	0.6	0.6	0.8	0.7	0.9	0.6	0.6	0.8	0.6	0.9	0.6	0.6	0.8	0.7
Average	**0.9**	**0.8**	**0.7**	**0.8**	**0.7**	**0.9**	**0.8**	**0.7**	**0.8**	**0.7**	**0.9**	**0.8**	**0.8**	**0.8**	**0.7**
Poisoning type	**Steganography**														
Cardboard	1	1	1	0.8	0.8	1	1	1	0.8	0.8	1	1	1	0.8	0.8
Glass	1	0.7	0.6	1	1	1	0.6	0.6	1	0.9	1	0.7	0.6	1	0.9
Metal	0.7	0.6	0.6	0.7	0.7	0.7	0.6	0.6	0.7	0.8	0.7	0.6	0.6	0.7	0.8
Paper	0.9	1	1	1	1	0.9	1	1	0.9	0.9	0.9	1	1	0.9	0.9
Plastic	0.8	0.7	0.7	0.7	0.8	0.8	0.7	0.7	0.7	0.8	0.8	0.7	0.7	0.7	0.8
Trash	0.9	0.8	0.6	0.9	1	0.9	0.6	0.6	0.9	0.9	0.9	0.7	0.6	0.9	0.9
Average	**0.9**	**0.8**	**0.8**	**0.9**	**0.9**	**0.9**	**0.8**	**0.8**	**0.8**	**0.9**	**0.9**	**0.8**	**0.8**	**0.8**	**0.9**

TABLE I: Individual performance of XAI methods on selected poisoned and unpoisoned samples.



(a) LIME method (b) SHAP method (c) Occlusion method

(d) LIME method (e) SHAP method (f) Occlusion method

Fig. 4: Object analysis with each XAI method as data is poisoned with, (a-c) Blurring and (d-f) Steganography

Formal verification: Formal verification of AI properties offers a rigorous approach to ensure AI behaves as intended. This method is particularly valuable for safety and robustness, key pillars of trustworthy AI deployment. By formally validating the satisfiability of expected behavior throughout its operation, potential vulnerabilities and deviations can be identified and mitigated. Additionally, deploying techniques like non-linear activation functions allows for in-depth analysis of neuronal activation, information propagation, and overall network expressiveness. This enhanced understanding facilitates rigorous validation of the internal dynamics governing AI behavior, strengthening its reliability and interpretability [19], [20].

Human Oversight: Human agency and oversight are among the primary requirements for making AI trustworthy. Human expertise and judgment can be leveraged throughout the lifecycle of AI from design to deployment to ensure that AI operates within defined boundaries, identifies undesirable outcomes for review, and behaves expectedly during changing contexts after deployment [21]. During an adversarial situation where the underlying models of AI are attacked, a set of guidelines can be proactively implemented by humans as guardrails to prevent abnormal behavior. For instance, as validators or testers can implement safety policies when AI inference is considered unsafe and can conduct several test scenarios for attack detection and verify the robustness and safety of the AI. Similarly, humans can act as analysts to assess the risks and potential impacts of the risk of an AI during deployment [22]

Standardization: Several vulnerabilities that compromise the robustness of AI have been explored [], and preventing them from the outset is paramount. The reliability of AI systems can be fortified against potential failures through the implementation of standardized design frameworks for AI designs. Adopting frameworks like NIST AI 100-2e2023 [23]and MITRE [24] that categorise AI vulnerabilities by context, domain, and applicable mitigations, alongside other adversarial tools and best practices from leading practitioners, for instance, Microsoft AI Security Risk Assessment Framework [25], and Microsoft *Counterfit* [26] can enable developers and AI engineers to address any flaws in AI design early enough. Moreover, transparency in design practices in compliance with standards enables traceability in the development and machine

learning process, which fosters interpretability of the internal workings of AI and its auditability.In addition, employing an explainability-by-design approach during AI development, an approach that caters to diverse audiences' understanding of AI decisions, can contribute to the trustworthiness of AI.

VII. Discussion

Application Domains: While our experiments focused on litter recognition, our results are more broadly applicable to any scenario that involves camera as the sensor to collect data. Indeed, we considered two different camera modalities (thermal and regular) and two common attack models that have been used on deep neural networks (blurring and steganography). These data are widely used on autonomous vehicles, e.g., for detecting pedestrians, signs, and obstacles, to support navigation and positioning, and more broadly to analyze the current operating environment. Naturally our results do not apply to all operations, e.g., to those that involve other type of input data (e.g., lidar or radar) or more targeted attacks (e.g., random spoofing of ultrasonic sensors [27]). Nevertheless, our work serves as an important starting point for analysing and understanding the benefits of XAI techniques in detecting and countering attacks on autonomous vehicles.

Stakeholders: Model diagnostics are particularly important for the companies and organizations that operate vehicles in urban settings. At the same time, municipalities and governmental authorities can require diagnostics to be integrated onto the vehicles before they issue permits as this can help ensure the vehicles are not vulnerable to targeted attacks. We would expect diagnostics to eventually become a legal requirement as well as this allows auditing the vehicle behaviour when accidents or other harmful behaviours occur.

Improvement: As room to improve our work, it would be important to study also other vehicle functionalities, such as navigation, localization, or collision avoidance. As these functions can result in dangerous behaviours, it is essential to design experiments that minimize risks, as well as to obtain the necessary ethical and legal permissions to carry out the experiments. We are also interested in exploring other environmental monitoring use cases, such as air quality [2] or water pollution monitoring [4]. These are examples of domains where vehicles could be used for emission accounting and thus there would be financial incentives to influence the performance of the AI models.

Practical Limitations: Attacks on AI models are not the only aspect that changes the behavior of autonomous vehicles, as hardware failures and software malfunctions can also affect their behavior. This requires methods that can operate on the vehicles and differentiate in real-time between targeted attacks and other errors. Another essential aspect is to integrate the vehicle with rollback mechanisms, return to home protocols, or other procedures that allow them to react to situations where erroneous data or other abnormalities are observed. Exploiting infrastructure backdoors is another potential way to attack vehicles.

End-User Support: Our work has demonstrated how changes in data can affect predictions and how XAI methods can be used to identify which parts of the data contributed to a given prediction. While this can help to detect abnormal behaviors, the output of the XAI methods can be difficult to interpret for end-users that analyze the vehicle behavior (e.g., technicians or vehicle operators). Further work is thus needed to develop tools and mechanisms that can translate the results of the XAI methods into insights that end-users that can use to fix potential problems.

Debugging Models: Optimally the AI models could be debugged in the wild semi-autonomously. This, however, would require dedicated support mechanisms on the vehicle which may be difficult to implement. For example, detecting points of deviation would require maintaining a reference model and retraining the model updates by replaying individual data points. The model could then be analyzed with XAI methods after each iteration to identify potential abnormalities in the data and to support the detection of issues. This process easily becomes highly resource intensive and requires sufficient memory and physical storage on the vehicles. In parallel, there is a need to verify the integrity of the reference models and XAI tools, which requires dedicated mechanisms such as trusted execution environments.

VIII. Summary and Conclusions

Powerful AI models are important for enabling autonomous operations of vehicles, particularly in complex and highly varying environments such as cities. These models need to be accurate and perform robustly as otherwise the vehicles can cause harm to citizens or damage the environment. We presented a research vision of how to enable the AI models to be trustworthy, identifying key challenges and requirements, and arguing that methods for model diagnosis should be integrated to vehicles to verify that the operations are indeed safe. We experimentally demonstrated the importance of model diagnosis by showing how targeted attacks, such as data poisoning, can break the AI models integrated on the vehicles. What makes these attacks particularly problematic is that they do not require access to the device or the AI model, operating solely by manipulating the inputs that it uses. We also demonstrated that explainable AI (XAI) methods provide a foundation for identifying issues but that they are also subject to limitations. In particular, XAI methods are unable to distinguish between targeted attacks and other malfunctions, their performance depends on the sensor modality, the environment, and the characteristics of the input data.

References

[1] Y. Fu *et al.*, "A survey of driving safety with sensing, vehicular communications, and artificial intelligence-based collision avoidance," *IEEE T-ITS*, 2021.

[2] N. H. Motlagh *et al.*, "Toward blue skies: City-scale air pollution monitoring using uavs," *IEEE Consum. Electron. Mag*, 2022.

[3] Z. Yin, M. Olapade, M. Liyanage, F. Dar, A. Zuniga, N. H. Motlagh, X. Su, S. Tarkoma, P. Hui, P. Nurmi, and H. Flores, "Toward city-scale litter monitoring using autonomous ground vehicles," *IEEE Pervasive Comput.*, 2022.

[4] H. Flores, N. H. Motlagh, A. Zuniga, M. Liyanage, M. Passananti, S. Tarkoma, M. Youssef, and P. Nurmi, "Toward large-scale autonomous marine pollution monitoring," *IEEE IoT Mag*, vol. 4, no. 1, pp. 40–45, 2021.

[5] A. Wojciechowska *et al.*, "Designing drones: Factors and characteristics influencing the perception of flying robots," *Proceedings of IMWUT 2019*, vol. 3, no. 3, pp. 1–19.
[6] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini, "Meaningful explanations of black box ai decision systems," in *AAAI Artif Intell.*, vol. 33, no. 01, 2019, pp. 9780–9784.
[7] A. Taïk, H. Moudoud, and S. Cherkaoui, "Data-quality based scheduling for federated edge learning," in *2021 IEEE LCN*. IEEE, 2021, pp. 17–23.
[8] P. Angelov and E. Soares, "Towards explainable deep neural networks (xdnn)," *Neural Networks*, vol. 130, pp. 185–194, 2020.
[9] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.
[10] M. Anneken, M. Veerappa, and N. Burkart, "Anomaly detection and xai concepts in swarm intelligence," 2021.
[11] M. K. Shehzad *et al.*, "Artificial intelligence for 6g networks: Technology advancement and standardization," *IEEE Vehicular Technology Magazine*, 2022.
[12] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2088–2105, 2020.
[13] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson, "Sponge examples: Energy-latency attacks on neural networks," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 212–231.
[14] Z. Fu, Y. Zhi, S. Ji, and X. Sun, "Remote attacks on drones vision sensors: An empirical study," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 5, pp. 3125–3135, 2021.
[15] R. A. Aral *et al.*, "Classification of trashnet dataset based on deep learning models," in *IEEE BigData*. IEEE, 2018, pp. 2058–2062.
[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *ACM SIGKDD*, 2016, pp. 1135–1144.
[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv Neural Inf Process Syst*, vol. 30, 2017.
[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*. Springer, 2014, pp. 818–833.
[19] J. M. Wing, "Trustworthy ai," *Communications of the ACM*, vol. 64, no. 10, pp. 64–71, 2021.
[20] T. Wu, Y. Dong, Z. Dong, A. Singa, X. Chen, and Y. Zhang, "Testing artificial intelligence system towards safety and robustness: State of the art." *IAENG International Journal of Computer Science*, vol. 47, no. 3, 2020.
[21] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: a review," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–38, 2022.
[22] E. C. J. R. Centre., "Robustness and explainability of artificial intelligence," European Union Commission Joint Research Center, Tech. Rep., 2020.
[23] A. Vassilev, A. Oprea, A. Fordyce, and H. Andersen, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," 2024.
[24] "MITRE ATT&K® Navigator," accessed: February 23, 2024.
[25] Microsoft Azure, "AI Security Risk Assessment," https://github.com/Azure/AI-Security-Risk-Assessment/blob/main/AI_Risk_Assessment_v4.1.4.pdf, accessed: February 23, 2024.
[26] ——, "Azure Counterfit," https://github.com/Azure/counterfit, accessed: February 23, 2024.
[27] W. Xu, C. Yan, W. Jia, X. Ji, and J. Liu, "Analyzing and enhancing the security of ultrasonic sensors for autonomous vehicles," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 5015–5029, 2018.

BIOGRAPHIES

Abdul-Rasheed Ottun is a Doctoral candidate at the University of Tartu. Research: Autonomous vehicles, Federated learning and Explainable AI. Contact him at rasheed.ottun@ut.ee

Akintola Adeyinka is a Doctoral candidate at the University of Tartu. Research: Light sensing, autonomous vehicles and pervasive systems. Contact him at zhigang.yin@ut.ee

Mohan Liyanage is a Lecturer at the University of Tartu. Research: Internet of Things, edge computing and networking. Contact him at mohan.liyanage@ut.ee

Michell Boerger is a Research Scientist for the Fraunhofer Institute for Open Communication Systems. Research: AI, Explainable AI (XAI), Quantum Artificial Intelligence, and Cybersecurity. Contact him at michell.boerger@fokus.fraunhofer.de

Pan Hui is a Professor at HKUST and a Nokia Professor at the University of Helsinki. Research: Mobile computing, opportunistic networks, and AI. Contact him at pan.hui@helsinki.fi

Nikolay Tcholtchev is working for the Fraunhofer Institute for Open Communication Systems. Research: Smart Cities (Open Urban Platforms), Cybersecurity, and AI. Contact him at nikolay.tcholtchev@fokus.fraunhofer.de

Sasu Tarkoma is a Full Professor at the University of Helsinki. Research: AI, data science, and sensing systems. Contact him at sasu.tarkoma@helsinki.fi

Petteri Nurmi is a Professor at the University of Helsinki. Research: Distributed systems, pervasive data science, and sensing systems. Contact him at petteri.nurmi@helsinki.fi

Huber Flores is an Associate Professor at the University of Tartu. Research: Distributed, mobile and pervasive computing systems. Contact him at huber.flores@ut.ee