Analysing Cyber Attacks and Risks in V2X-Assisted Autonomous Highway Merging

Chao Chen, Ugur Ilker Atmaca, Konstantinos Koufos, Mehrdad Dianati, and Carsten Maple

1 Introduction

Progresses in highly reliable Vehicle-to-Everything (V2X) communication systems are expected to usher in the era of Connected Autonomous Vehicles (CAVs) [1, 2], which create opportunities for safer and more efficient implementations of autonomous driving functions. However, wireless connectivity can also open new attack surfaces in V2X-assisted autonomous vehicles, which must be understood and addressed before the commercialisation of CAV functions. While generic countermeasures (e.g., encryption and authentication) are considered in various forms of V2X systems, lessons from the past show that each CAV function should be separately analysed, and countermeasures should be customised for it to ensure its security and efficiency [3].

To this end, this paper focuses on security threats and risk assessment in a V2X-assisted autonomous highway merging function. The goal of such a function is to help an autonomous vehicle to safely merge into the highway from a slip road (i.e., freeway on-ramp merging) without unnecessary slowdowns or delays. In a typical implementation of such functionality, the CAVs are considered to be equipped with their own sensors. However, in many complex and safety-critical scenarios, on-board sensors can be impaired by obstructions, such as vegetation, or other road features, as illustrated by Fig. 1, that block the view of the on-board sensors. In such scenarios, the infrastructure sensors can make a crucial difference by detecting the oncoming vehicles along the highway and broadcasting their locations, velocities, etc., through a V2X system to the surrounding CAVs approaching the merging point. Such broadcast messages can then be fused by the CAV's on-board perception system

C. Chen · U. Ilker Atmaca · K. Koufos · M. Dianati (✉) · Carsten Maple
Warwick Manufacturing Group (WMG), University of Warwick, Coventry, UK
e-mail: c.chen.27@warwick.ac.uk; ugur-ilker.atmaca@warwick.ac.uk; konstantinos.koufos@warwick.ac.uk; m.dianati@warwick.ac.uk; cm@warwick.ac.uk

K. Daimi et al. (eds.), *Emerging Trends in Cybersecurity Applications*,
https://doi.org/10.1007/978-3-031-09640-2_9

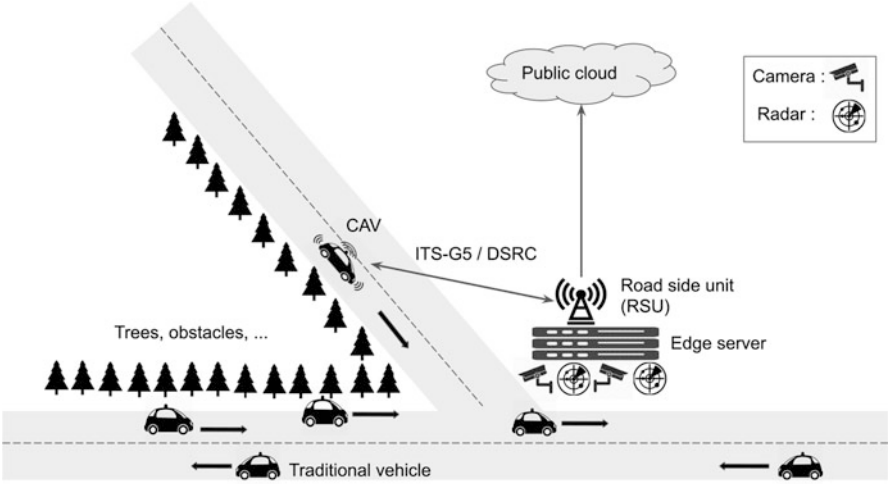

Fig. 1 An exemplary use case for V2X-assisted highway merging, where roadside plantation obscures the vision of on-board vehicular sensors

to create a highly accurate and reliable perception of the road segment, enabling a safer and efficient merging manoeuvre. However, from a security perspective, such a system has exposed entry points for malicious actors.

The existing literature on cyber security for CAVs has mostly considered the threats at the CAV side [4–7]. At the V2X-system side, the threat analysis has been conducted merely at the communication layer, see, for instance, [8], without considering potential attacks on the infrastructure sensors, as we will do in this paper. Also, no precise merging scenario has been analysed thus far. Therefore the existing cyber security analysis for CAVs is rather incomplete, and it does not provide any practical insights into ways of evaluating and ranking the risk of relative threats and their mitigation schemes. To the best of our knowledge, this paper is the first that applies an abuse case that combines multiple attacks to completely "blind" the road infrastructure within the V2X system. Our contributions are: (i) Proposition of a generic reference architecture (RA) for implementation of V2X-assisted autonomous highway merging, and (ii) analysis of the potential cyber security threats to the RA and assessment of their risks. Particularly, we analyse potential abuse cases against the road infrastructure consisting of radar and cameras. We discuss how a malicious actor can jam the radar and/or tamper with the image object detector in the camera. Finally, we suggest related mitigation schemes.

2 Related Work

The V2X functionality is expected to enhance road traffic efficiency and safety, and because of that, it has already attracted a considerable interest in the academic [2, 9] and industrial research communities [10]. The study in [2] has discussed the use of messages transmitted from the road side unit (RSU) over V2X to the CAV for helping to execute a lane merge. The messages contain the speeds and locations of surrounding vehicles to the CAV. Furthermore, the study in [9] has suggested using a camera and image object detector to build the semantic map of the environment which is sent over V2X to the merging CAV. Finally, the study in [10] has proposed to install both cameras and radars at the RSU. While the above studies conclude that merges can be executed efficiently and safely with the V2X system, unfortunately, they do not address the issues due to cyber security, which is the main topic of our work.

The literature on threat analysis and risk assessment for CAVs is vast, see [5, 7, 11–13] and the references therein. The study in [5] has defined potential threats, e.g., various types of spoofing, tampering, etc., identified what kind of expertise, knowledge, equipment and, window-of-opportunity is needed to realise the threat, and they have finally quantified and ranked the priority for each attack. The studies in [11, 13] have also used the controllability criterion while ranking the threats, e.g., a threat that can be easily controlled and avoided by the driver and/or by the CAV should be associated with low risk. The above studies have considered cyber security analysis only on the CAV's side. In addition, they have not addressed a specific autonomous driving function, e.g., V2X-aided highway merging, as we will do here.

Due to the high economic cost along with the technology readiness level for the lidar, we consider in our RA only cameras and radars as the primary sensors at the RSU. Radar jamming is widely used for military purposes, but it has recently started to get attention to exploit it for jamming on CAVs. For instance, the study in [4] has considered the transmission of artificial noise for jamming. The study in [14] has applied a radar signal analyser, frequency multiplier, and signal generator to successfully jam Tesla's radar, rendering it unable to detect any surrounding objects.

Exploiting the vulnerabilities of a machine learning model, particularly for the image object detector applied in the camera, has existed for long time at a theoretical level, but it has only recently become a reality [15, 16]. The studies in [15, 17] have printed an adversarial patch, and the study in [16] has attached a poster and a patch to the target object, making it "invisible" to the detector. In our abuse case, we will combine these methods of tampering with the image object detector and the jamming of the radar, to achieve the goal of the attack, which is a totally "blind" RSU.

3 Reference Architecture for Autonomous Highway Merging

Given the scenario depicted in Fig. 1, we have investigated related academic and industrial projects [2, 9], to come up with a proposal about the reference architecture (RA) for autonomous V2X-assisted highway merging. A fairly generic model is illustrated in Fig. 2.

The RSU is installed near the merging point, and the radars and cameras observe the traffic along the highway. Firstly, they stream raw data, i.e., images and radar signals, to the edge server via cable. Secondly, the edge server processes the received data using models such as the object detector and the radar signal analyser. The edge server fuses the outputs of the processing models and produces a list of objects including, e.g., their type, size, and speed. This information can then be used to generate the collective perception messages (CPMs) [18]. Finally, the RSU broadcasts the CPMs via ITS-G5 or dedicated short-range communications (DSRC) to the CAV. We would like to note that vehicle-to-vehicle cooperative driving messages for platooning control and manoeuvre coordination are not relevant to our RA.

The CAV at the slip road becomes aware of its surroundings and the situation along the highway through its on-board sensors, such as camera, radar, lidar, GNSS, etc. In addition, it establishes wireless connectivity with the RSU through its on-board unit (OBU) and receives the CPMs over V2X. The on- and off-board information is fused to construct the overall perception, which is then fed into the machine learning autonomous driving application to infer the most suitable trajectory and speed to perform the merge (vehicle control). While approaching the highway, the CAV keeps on receiving the CPMs and continuously updates its

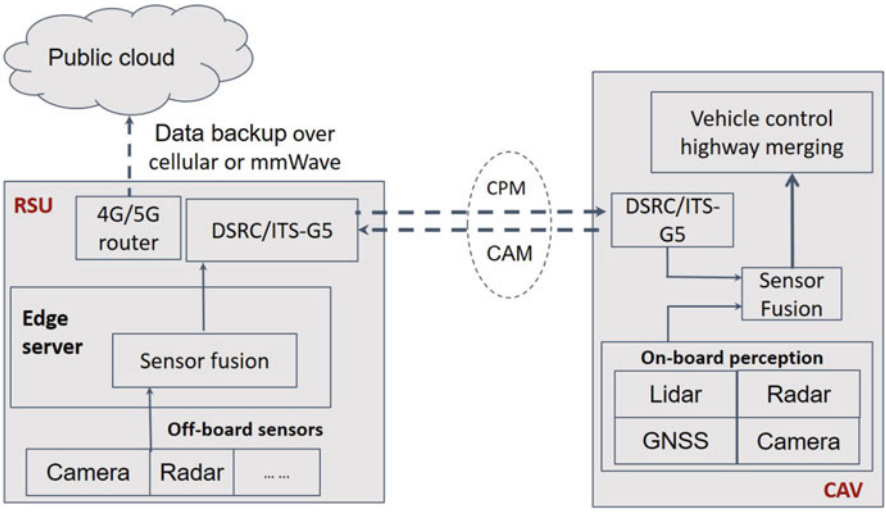

Fig. 2 Generic RA for V2X-assisted highway merging

knowledge about the highway traffic and its surroundings, and uses this to adjust its merging manoeuvre. Additionally, it keeps on broadcasting cooperative awareness messages (CAMs), and the RSU, in its turn, logs all the received data to the public cloud as a backup. For that, the RSU must also have cellular and/or mm-wave wireless connectivity to the cloud.

Unlike CPM, the CAM contains only local information about the location and the speed of the CAV. The installation of radars and cameras near the intersection essentially bypasses the need for sending CAMs to the RSU, hence, the proposed RA is also relevant for early penetration of CAVs.

4 Abuse Cases

In this section, we will describe potential abuse cases attacking the RA. This is to highlight the importance of understanding and mitigating potential threats for such systems before their commercial rollout. In order to do that, firstly, we will identify all attack surfaces. Subsequently, we will concentrate on attack surfaces at the RSU, because this subject has not been treated so far in the literature. In particular, we will consider attacking the processing models for the image object detector at the edge server, as well as jamming the radar sensor on the RSU. These abuse cases jeopardise road safety by resulting in the broadcast of imprecise CPMs, and subsequently lead to the execution of unsuitable, perhaps risky, manoeuvres at the CAV approaching the merging point. We illustrate these abuse cases in the form of an attack tree in Fig. 3, and explain them in detail next.

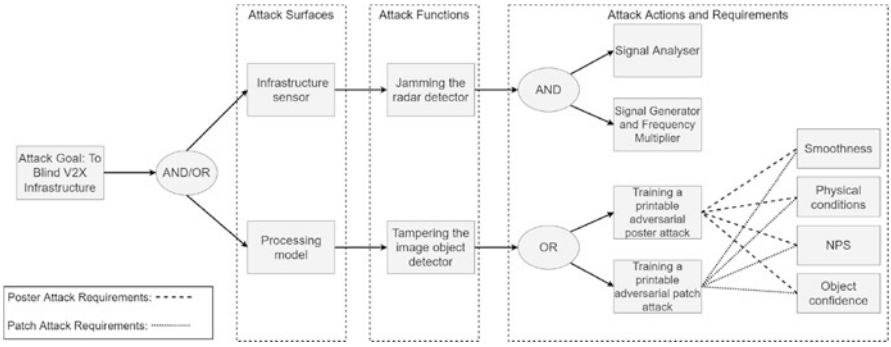

Fig. 3 The attack tree for the abuse cases

4.1 Attack Surface Analysis

The attack surface of a system is the set of vulnerable components (entry points), which a threat actor could exploit to attack. The RA in Fig. 2 has provided an abstract model of the high-level system components and their interactions, which can be used to identify the following attack surfaces in our system setup.

- **Vehicular sensors**: All on-board sensors (e.g., cameras, radar, lidar and GNSS, etc.) equipped on the CAV.
- **V2X communication**: The OBU at the CAV and the wireless communication router at the RSU.
- **Infrastructure sensors**: The off-board sensors at the RSU, such as cameras and radars.
- **Processing models**: The models within the edge server for processing the raw sensor data, e.g., the received radar signal and the raw image data.
- **Infrastructure edge server**: The computing unit at the edge server for hosting the sensor fusion models, e.g., to fuse the outputs from the radar signal analyser and the image object detector.

We have excluded the public cloud from the attack surfaces, because its communication to the RSU is unidirectional, see Fig. 2. Security issues at the public cloud do not affect the RSU and the CAV in our RA.

4.2 Tampering with the Input to the Image Object Detector

There are many ways to attack the sensor camera at the vehicle and/or the RSU, e.g., sensor blinding with light emitting diode [14]. However, tampering the objector detector in the camera has not been so far investigated. Next, we will firstly explore how does the image object detector work with images captured by the cameras. Then, we will describe how the threat actor can cast the attack as an optimisation problem on the pixels of the adversarial image. The goal is to construct an adversarial input image that may confuse the object detection system on the RSU.

There are two approaches to the execution of this attack. The first approach is called the "poster attack", which generates a subtle adversarial poster confined to the surface area of the target object. The poster usually has the size of the target object, and it is inconspicuous, e.g., it looks like a graffiti, to the human eye. The second approach is called "patch attack", which generates a small-sized patch attached on the target object. For example, adversaries may stick a misleading patch on the surface of an oncoming vehicle to prevent its detection on the camera; or they can utilise drones to reflect the patches and false traffic signs on vehicles or the road to induce accidents or stop the traffic flow. To generate the adversarial image, the threat actor needs to know the full details of the machine learning model of the object detector.

The risk of such threats can be mitigated by ensuring confidentiality of the utilised object detector model.

(1) Background on Image Object Detectors Image object detection is a classical task in computer vision. It does not only classify the objects in an image, but it also generates the coordinates of the bounding boxes for each detected object. In this work, we will focus on the state-of-the-art Yolov2 detector, see [19] which is widely employed nowadays because it is faster than other detectors without compromising on accuracy.

The Yolov2 firstly divides the input image in a 19×19 grid. For each grid cell, it generates five anchor boxes of different sizes, which can capture various objects within the cell, see Fig. 4. Each anchor box produces a prediction box, which contains: (i) the coordinates of the bounding box, (ii) a confidence score which measures the certainty that the given anchor box contains an object, and (iii) the probability values that the detected object belongs to each of the 80 object classes. The object detector will finally discard all prediction boxes whose confidence score is lower than a predefined threshold, e.g., the default Yolov2 uses 70% confidence level to produce the final prediction result.

(2) Physical Adversarial Image In order to generate an adversarial image, the threat actor must be able to evaluate the following metrics: The smoothness of pixels, the object confidence score, and the non-printability score (NPS) [20], see the bottom branch of the attack tree in Fig. 3. The threat actor aims to minimise a function of these metrics, usually their linear weighted sum, along with some physical condition transformations. Here, different transformations reflect different physical conditions on the input image, such as varying distance, angle, rotation, and brightness [21]. By

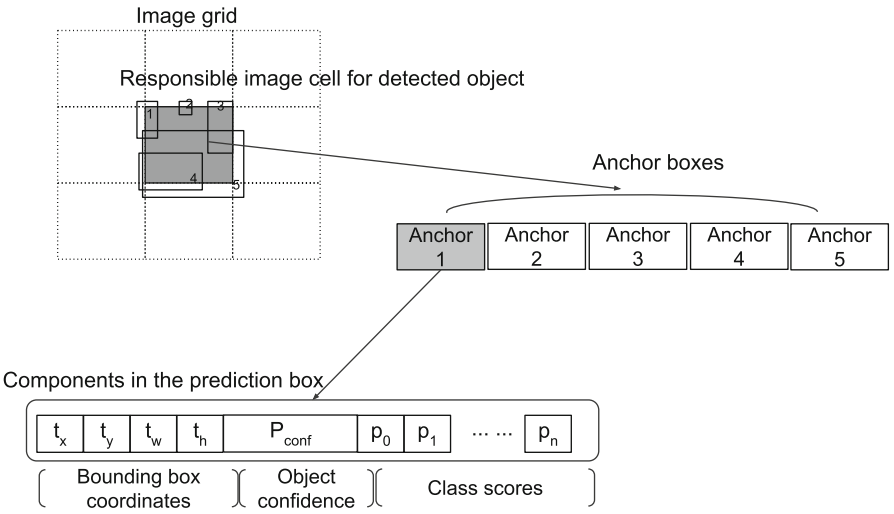

Fig. 4 Yolov v2 image object detector

minimising the function of these metrics, hereafter referred to as the loss function, the generated adversarial image is likely to confuse the object detector.

One property of natural images is that neighbouring pixels share similar colours, which is known as smoothness. To make the adversarial image as natural as possible, the colour difference between neighbouring pixels should be small, preventing the generation of noisy images. The object confidence score is the second metric that the threat actor wishes to reduce because a low score can mislead the object detector to ignore the target object with the adversarial image attached to it. If the threat actor wants to hide a vehicle, the adversarial image must lower the score of vehicle detection. Finally, the NPS represents the ability of the printer to reproduce the colours of the adversarial image. Naturally, the threat actor aims at a low NPS, which means that the printed image is close, in terms of colours, to the adversarial image.

To sum up, the workflow of this physical adversarial attack to the image object detector consists of four steps: In the first step, the threat actor has to define the loss metrics and select a loss function. Secondly, he initialises the optimisation process with a random adversarial image and applies it to the top of the image of the target object. If this is a "patch attack", he fixes the position of the patch at the centre of the target object, and otherwise, he covers the full object. He then applies the required physical condition transformations. Thirdly, the superposition of the adversarial image on that of the target object is fed into the image object detector to predict the bounding boxes and evaluate the class probabilities and the object confidence score. Based on the selected loss function, the convolutional neural network back-propagates and updates the pixels of the initial adversarial image using stochastic gradient descent (SGD). Eventually, the loss function converges after some iterations. In the fourth step, the threat actor prints the last adversarial image generated from the neural network and attaches it onto the target object.

4.3 Jamming Infrastructure Radar

Jamming a radar requires to generate electromagnetic waves at the same (or nearly adjacent) frequency spectrum band to that of the target radar. These waves will act as additional noise, possibly lowering its signal-to-noise-ratio, and compromising its detection performance. To jam the radar at the RSU, we will apply the exemplary methodology from [14], where the mm-wave radar from Tesla was jammed. The jamming device must include the following components: An oscilloscope, a signal analyser, a signal generator, and a frequency multiplier. The signal analyser and the oscilloscope are needed to identify the operational frequency of the radar. The signal generator working together with a frequency multiplier is a low-cost alternative to signal generators with a higher maximum frequency range.

5 Threat Modelling Methodology

In this section, we will first identify the cyber security requirements relevant to the V2X-assisted highway merging scenario. After that, we will map these requirements to threats and discuss how to evaluate their impact and risk ranking. In order to associate the cyber security requirements with threat classes, we will use the STRIDE model, because it is a systematic and lightweight approach. STRIDE stands for spoofing (S), tampering (T), repudiation (R), information disclosure (I), denial-of-service (D), and elevation-of-privilege (E). For risk analysis, we will use the TARA+ methodology, as it quantifies the risk ranking based on the attack potential and its related impact along with controllability [11].

5.1 Cyber Security Requirements

Various research projects have conducted a general analysis of cyber security requirements for CAVs with V2X connectivity and divided the cyber security requirements into the following categories: confidentiality, integrity, availability, authenticity, privacy, trace-ability, authorisation, non-repudiation, and robustness against external threats. We will consider the above requirements and identify those that can be violated by a threat actor in our system's RA, see the rightmost column of Table 1. Jamming the radar will affect the availability of the infrastructure sensors, and tampering with the input images to the object detector will violate the integrity of the processing models. On the other hand, the identification of threats has a higher priority than the tracing of individual threat actors in our system setup, making non-repudiation irrelevant.

In Table 1, we have mapped the cyber security requirements for V2X-assisted CAVs to threat classes according to STRIDE. In the second column of Table 2, we have investigated the vulnerability of the five attack surfaces presented in Sect. 4.1 concerning these threat classes. To give an example, while spoofing, tampering, and denial-of-service apply to almost all attack surfaces in our RA, the elevation-of-privilege is relevant only to the infrastructure edge server. The threat actor can

Table 1 STRIDE and cyber security requirements for V2X-assisted highway merging

Threat classes	Cyber security requirements	Highway merging functionality
Spoofing (S)	Authenticity	✓
Tampering (T)	Integrity	✓
Repudiation (R)	Trace-ability, Non-repudiation	×
Information Disclosure (I)	Confidentiality Privacy	✓ ×
Denial-of-Service (D)	Availability	✓
Elevation-of-Privilege (E)	Authorisation	✓

Table 2 Attack surface analysis for V2X-assisted highway merging

Attack surface	STRIDE	Expertise (E_x)	Knowledge (K_t)	Equipment (E_m)	Window-of-opportunity (W_o)	Controllability factor (C)
Vehicular (on-board) sensors	S, T, D	Proficient, expert and layman	Public and sensitive	Standard, specialised, and bespoke	Medium and large	The attacks can be controlled by anomaly detection algorithms in the sensors, or by an intrusion detection system (IDS)
V2X communication	S, T, D, I	Expert and layman	Restricted	Specialised and bespoke	Medium and large	The attack is controllable by a plausibility check, e.g., certification-based V2X security framework, which helps to build trustworthiness between the vehicle and the infrastructure
Infrastructure (off-board) sensors	S, T, D	Proficient, expert and layman	Public and restricted	Standard, specialised, and bespoke	Medium and large	Cyberattacks can be controlled by anomaly detection algorithms or by an IDS installed in the infrastructure
Processing models	S, T	Proficient	Public and critical	Specialised and bespoke	Large	Redundant sensors can ensure that this attack is controllable
Infrastructure edge server	E, D	Proficient and layman	Restricted	Standard and specialised	Medium and large	Unauthorised physical access can be blocked, and a CCTV system can be installed for monitoring the infrastructure

use the physical access to the server to gain the administer permission, exploit further threats, and violate the authorisations from the cyber security requirements. In the other columns, we have assessed the required level of expertise, knowledge, equipment, and window-of-opportunity to be able to attack the target surface. Finally, in the rightmost column, we have described potential methods to detect and control each threat. All these features will be used by TARA+ for quantifying the risk of each attack.

5.2 TARA+

Threat Analysis and Risk Assessment Plus (TARA+) [11] has been developed to analyse cyber risks related to the society of automotive engineers (SAE) Level 3 automated driving functions and beyond, in order to account for shared system/driver responsibility in the CAV's control. TARA+ is an enhanced version of TARA which integrates the system's/driver's controllability factor on cyberattacks that renders the risk analysis more realistic. In this paper, we focus on Level 4 autonomy that still has an option for the driver to take control, despite this is not necessary [22]. Therefore, only the system's controllability factor is considered hereafter. The output of TARA+ is a value that indicates the severity of the risk. Its calculation depends on multiple factors that are detailed next.

Attack Potential The attack potential P_o is a linear function of the threat actor's expertise E_x, the required equipment E_m, the knowledge regarding the target system K_t, and the window-of-opportunity W_o. If we assign equal weights to these parameters we get

$$P_o = E_x + E_m + K_t + W_o. \tag{1}$$

In Table 3, lower values in the rightmost column are associated with higher attack potentials. For instance, in case a threat actor with layman's expertise on cyber security, public information about the target system, standard equipment, and unlimited window-of-opportunity can successfully apply an attack, we will regard this attack as very possible. After calculating P_o, we can also quantify the attack probability ranking, P_r, of the identified attack, see [11, Table IV].

Table 3 Attack potential factors and ranking [11, Table I]

Expertise (E_x)	Knowledge of the target (K_t)	Equipment (E_m)	Window-of-opportunity (W_o)	Value
Layman	Public Information	Standard	Unlimited	0
Proficient	Restricted Information	Specialised	Large	1
Expert	Sensitive Information	Bespoke	Medium	2
Multiple Experts	Critical Information	Multiple Bespoke	Small	3

Table 4 Risk ranking matrix [11]

Risk ranking (R^*)	$P_r = 0$	$P_r = 1$	$P_r = 2$	$P_r = 3$	$P_r = 4$
$MI = 0$	QM	QM	QM	QM	Low
$MI = 1$	QM	Low	Low	Low	Medium
$MI = 2$	QM	Low	Medium	Medium	High
$MI = 3$	QM	Low	Medium	High	High
$MI = 4$	Low	Medium	High	High	Critical

Impact Factor The impact factor I_f quantifies the cost incurred by an attack to the system. It is a linear function of the attack severity, S_v, the operational malfunction, O_f, the financial cost, F_c and the privacy/legislative cost, P_c, where the scaling weights are adjustable.

$$I_f = 3\,S_v + F_c + 2\,O_f + P_c. \tag{2}$$

To assign values to the four parameters, we have used TARA+ [11, Table II]. One may already deduce that although an attack might be very probable, it might still be possible to ignore, in case it is associated with a low impact factor.

Modified Impact and Risk Ranking The modified impact, MI_f, depends on the impact factor and the controllability of the attack. The controllability factor $C \in \{0, 1, 2, 3, 4\}$ quantifies the resilience of the system against attacks. If the system can detect the attack, and it continues to be fully operational with a sufficient level of redundancy, the attack can be safely controlled, and the controllability factor is defined to be zero ($C = 0$). When the system can detect the attack, but the operation level is compromised due to safety reasons, the controllability factor is equal to three ($C = 3$). Finally, if the system cannot detect the attack, the controllability factor is set equal to four ($C = 4$). See [11, Table III] for more details. Finally, the modified impact MI_f can be read as

$$MI_f = \frac{I_f \cdot C}{4}. \tag{3}$$

The modified impact ranking, MI, results from the quantisation of MI_f, see [11, Table V]. With the parameters MI and P_r at hand, one can determine the risk ranking of the threat R^* based on Table 4. The risk ranking ranges from "QM" (Quality Management) at the lowest level to "Critical" at the highest level.

Table 5 Risk assessment of the abuse cases by TARA+

Attack scenario	Attack surface	Attack potential (P_o) and probability (P_r)	Controllability factor (C)	Impact ranking (MI)	Risk ranking (R^*)
Jamming the radar detector	Infrastructure sensors	$Ex = 1$, $E_m = 2$, $K_t = 0$, $W_o = 1$ $P_o = 4$, $Pr = 3$ (*Possible*)	$C = 3$	$S_v = 2$, $F_c = 2$, $O_f = 3$, $P_c = 0$ $I_f = 14$, $MI = 2$ (*Medium*)	*Medium*
Tampering with the image object detector	Processing models	$E_x = 1$, $E_m = 1$, $K_t = 3$, $W_o = 1$ $P_o = 6$, $P_r = 3$ (*Possible*)	$C = 3$	$S_v = 2$, $F_c = 2$, $O_f = 3$, $P_c = 0$ $I_f = 14$, $MI = 2$ (*Medium*)	*Medium*
Blinding the V2X infrastructure	Infrastructure sensors and processing models	$E_x = 1$, $E_m = 2$, $K_t = 3$, $W_o = 1$ $P_o = 7$, $P_r = 2$ (*Unlikely*)	$C = 4$	$S_v = 4$, $F_c = 4$, $O_f = 4$, $P_c = 0$ $I_f = 24$, $MI = 4$ (*Critical*)	*High*

6 Risk Assessment and Mitigation Schemes

In this section, we will rank the risk of the abuse cases presented in Sects. 4.2 and 4.3 using TARA+ and also discuss potential mitigation schemes.

Jamming the Radar Detector In order to rank the risk of radar jamming, we first need to model the threat actor and evaluate the attack potential P_o, see Eq. (1) and Table 3. We have selected: (i) *Proficient* expertise, $E_x = 1$, because a general security knowledge about popular attacks, like jamming, is only required. (ii) *Bespoke* equipment, $E_m = 2$, because the threat actor does not only need a signal analyser and a signal generator but also need a frequency multiplier to enhance the maximum frequency range of the generated waveforms and ensure the jamming of the radar. (iii) *Public information* for the knowledge of the target, $K_t = 0$, because the threat actors do not require the detailed specifications as long as they can get the basic and public knowledge about the location of the radar. (iv) *Large* window-of-opportunity, $W_o = 1$, as the radar jamming can be executed remotely with minimal time constraints. After substituting these values into (1), we end up with $P_o = 4$. According to [11, Table IV], the attack probability, $P_r = 3$, is classified as *possible*.

Next, we use Eq. (2) and [11, Table II] to assess the impact factor I_f of the attack. We have selected its severity $S_v = 2$ and the financial cost $F_c = 2$ to be medium, and the operational malfunction of the vehicle $O_f = 3$ to be high, as it affects a primary function of the CAV. In addition, $P_c = 0$ because there are no privacy considerations for the cyber security requirements listed in Table 1. These values yield $I_f = 14$. Since only the radar is jammed, anomaly detection algorithms and/or intrusion detection systems (IDS) can identify the attack by monitoring the difference

between the output of the radar signal's processor and the output of other sensors like the camera [23]. As a result, we have selected the controllability factor $C = 3$. After substituting $C = 3$ and $I_f = 14$ into (3), we get $MI_f = 10.5$. Subsequently, the modified impact ranking is $MI = 2$ or *medium* according to [11, Table V]. Finally, from Table 4 we conclude that the risk ranking R^* of this abuse case is *medium*.

Tampering with the Image Object Detector The risk ranking follows the same procedure to that of radar jamming. The calculation of the values for the parameters, P_o, P_r, I_f, etc., is presented in the second row of Table 5. Training an adversarial image to tamper with the image object detector requires just a GPU to train the machine learning model and a colour printer. Therefore, specialised equipment, $Em = 1$, is sufficient. Also, recall from Sect. 4.2 that the threat actors require full knowledge of the targeted image object detector to carry out this attack, hence, $K_t = 3$. In Table 5, we see that the ranking for the attack probability $P_r = 3$ and the modified impact ranking $MI = 2$ are equal to those of radar jamming and thus, the risk ranking R^* is *medium* too.

Blinding the Road Infrastructure If both attacks are executed simultaneously, the RSU becomes completely blind. To evaluate the attack potential P_o of the combined attack, we have selected the maximum value (over the previous abuse cases) for each of the four parameters involved in its calculation, yielding $P_o = 7$ and $P_r = 2$ *(Unlikely)*, see the last row of Table 5. At the same time, the combined attack has the highest controllability factor $C = 4$, because it is generally undetectable by the system. It is also likely to cause devastating effects on other traffic objects. For instance, if the CAV uses an incorrect trajectory to merge based on the received inaccurate perception knowledge from the RSU, it is probable to cause serious safety issues along the highway. Accordingly, we have assigned the highest severity, financial, and operational cost to the combined attack yielding $I_f = 24$. Finally, based on Eq. (3) and [11, Table V], we have calculated the modified impact ranking MI as *critical*, and the risk ranking R^* as *High*.

Mitigation Schemes We will describe mitigation schemes for each of the three abuse cases. Attacking the object detector requires that the threat actor has the complete knowledge of the machine learning model running in the RSU, e.g., the architecture of the neural network, the trained weights, etc. Therefore, the safety recommendation is to guarantee the model's confidentiality, which is related to the threat class, "information disclosure" in STRIDE. Furthermore, one way to mitigate the impact of a jamming attack is to deploy a frequency-agile radar, which can switch its operation among different frequencies. It is noted that including both a radar and a camera at the RSU is a mitigation mechanism per se. When the attack aims at an individual sensor, the risk decreases, if the system can detect the inconsistencies between the outputs of different sensors. In that case, the RSU can go into safe-mode operation, for instance, it can stop broadcasting CPMs. Finally, if both sensors are under attack, we should combine the mitigation schemes of the previous abuse cases. It is expected that soon there will be further redundancy at the RSU, equipping

it with more sensors, e.g., lidar and ultrasound devices, which will mitigate further the cyber security risks.

7 Comparative Study

In this section, we will present a comparative study between our focused risk assessment model, Tara+, and two previous models, see the studies in [5] and [13], to discuss the validation of the selected parameters in Table 5. These studies utilised the defined threat actor profiles on Table 6, and based on these profiles, the values of various parameters like expertise, motivation, knowledge about the target system, equipment in use and financial availability are specified. Afterwards, the attack capability can be calculated as a function of these parameters.

Specifically, the risk assessment model of [5] calculates the attack potential from the difference between the threat actor's capability to execute a successful attack and the system's resistance to the attack. The system's resistance is the minimum required capability to realise a successful attack. The required capability can be determined by doing an attack surface analysis in a similar manner as we did earlier for Tara+, see Table 2. Likewise, the SARA method [13] requires to identify the minimal required threat actor profile which is able to execute a successful attack to the system, and based on that it evaluates the attack potential. The minimal threat actor profile to the attack scenarios of our system's RA (jamming of the radar and/or image object detection in the camera) is *Organised Crime* from Table 6. Unlike [5] and [13], the parameters of our risk assessment using Tara+ have been determined by considering the minimal requirements of a successful attack from the system perspective and not from the perspective of the threat actor.

The results of our comparative study are presented on Table 7. Unlike Tara+, the risk assessment methodology in [5] does not categorise the risk as low, medium, or high, but it visualises the three characteristics of a cyber risk (i.e., potential, motivation, and impact) using contour plots in the two-dimensional space. Thus, it will help security analysts to understand which countermeasures to prioritise. Although trends of the results are consistent with our analysis using Tara+, the methodology could not assess that the attack probability of *Blinding the V2X infrastructure* should be a lot less than the other two. This is because the risk assessment methodology in [5] does not consider the system's controllability. On the other hand, SARA [13] includes a metric called observation which is defined as the system's ability to detect errors and operates on risk calculation. The attack probabilities of our scenarios are quantified as *Moderate* according to the SARA's attack mapping table, because the threat actor has been identified as *Organised Crime*, and any attack executed by *Organised Crime* cannot be lower than *Moderate*. However, *Blinding the V2X infrastructure* in our RA is a lot less likely as compared to the other two individual scenarios, due to the system's redundancy. Sara methodology is not able to capture the reduced probability (the likelihood) of the combined attack. Nevertheless, it quantifies a higher risk, R7 instead of R4, for the combined attack.

Table 6 Threat actor profiles [5]

Threat agent	Motivations	Finances	Expertise	Knowledge	Equipment
Thief	Financial	Low	Layman	Public	Standard
Owner	Financial	Low	Layman	Public	Standard
Organised Crime	Financial	High	Proficient	Restricted	Bespoke
Mechanic	Financial	Low	Expert	Critical	Specialised
Hacktivist	Ideology, Passion	Low	Multiple Experts	Sensitive	Multiple Bespoke
Terrorist	Ideology	Low	Layman	Public	Standard
Foreign Government	Financial, Ideology	High	Multiple Experts	Restricted	Multiple Bespoke

Table 7 Comparison study of risk assessment models

Attack scenario	TARA+	Risk ranking by [5]	SARA risk ranking by [13]
Jamming the radar detector	Attack probability: Possible Risk: Medium	Attack probability: $\Delta Ex = 0$, $\Delta E_m = 1$, $\Delta K_t = 2$, $\Delta W_o = 0$ **Pr=3** Impact: $S_v = 2$, $F_c = 2$, $O_f = 3$, $P_c = 0$ **I=9** Motivation: $m_f = 3, m_r = 1$ **M=2**	Attack probability: $C_a = 16, T = 2, W_o = 1$ **Pr=3 (Moderate)** Severity:(2, 2, 3, 0) **S=3** Risk: **R4**
Tampering with the image object detector	Attack probability: Possible Risk: Medium	Attack probability: $\Delta Ex = 0$, $\Delta E_m = 2$, $\Delta K_t = 0$, $\Delta W_o = 1$ **Pr=3** Impact: $S_v = 2$, $F_c = 2$, $O_f = 3$, $P_c = 0$ **I=9** Motivation: $m_f = 3, m_r = 1$ **M=2**	Attack probability: $C_a = 16, T = 2, W_o = 1$ **Pr=3 (Moderate)** Severity:(2, 2, 3, 0) **S=3** Risk: **R4**
Blinding the V2X infrastructure	Attack probability: Unlikely Risk: High	Attack probability: $\Delta Ex = 0$, $\Delta E_m = 1$, $\Delta K_t = 0$, $\Delta W_o = 1$ **Pr=2** Impact: $S_v = 4$, $F_c = 4$, $O_f = 4$, $P_c = 0$ **I=20** Motivation: $m_f = 3, m_r = 2$ **M=1**	Attack probability: $C_a = 16, T = 1, W_o = 1$ **Pr=3 (Moderate)** Severity:(4, 4, 4, 4) **S=4** Risk: **R7**

8 Conclusion

The highway merging functionality with Level 4 autonomous driving capability from a slip road will improve traffic efficiency and safety and help increase the driver's comfort. In this paper, we have considered a merging scenario where the autonomous vehicle does not rely only on the on-board sensor knowledge to perform the merge, but it also receives perception messages from the road infrastructure (off-board radar and camera) over V2X. Although the enhanced perception brings clear benefits, it also creates a challenge from the cyber security perspective. We have identified the attack surfaces in the system's reference architecture and applied STRIDE to classify the threats. We have used TARA+ to evaluate the cyber security risks for three abuse cases, where the threat actors attack the off-board sensors by jamming the infrastructure radar and/or tampering with the input image to the object detector at the camera. The abuse case that combines both attacks and blinds completely the road infrastructure is, to the best of our knowledge, new. It turns out that the associated risk is medium in case a single sensor is under attack and high when both sensors are under attack. Apart from deploying more sensors like lidar and ultrasound devices at the infrastructure, we have suggested that the confidentiality of the image object detector and the frequency agility of the radar are crucial to avoid security breaches. We believe that this work can help the automotive stakeholders towards the understanding and mitigation of cyber security threats before the commercial rollout of autonomous vehicles. In the future, it is important to conduct simulation-based studies of the attack scenarios discussed in this paper. By doing so, we will measure the impact of the identified cyber security attacks on the V2X autonomous driving systems and be able to select cost-effective countermeasures to improve the resilience of these systems.

Acknowledgments This work was partly funded by UK Research and Innovation through INNOVATE UK in project AutopleX (project reference 104272) and the European Union's Horizon 2020 research and innovation programme in project L3Pilot under grant agreement No 723051. The sole responsibility of this publication lies with the authors. The authors would like to thank all partners within AutopleX and L3Pilot for their cooperation and valuable contribution.

References

1. A. Stevens et al., Cooperative automation through the cloud: The CARMA project, in *12th ITS European Congr.,* 2017
2. D. Bevly et al., Lane change and merge maneuvers for connected and automated vehicles: A survey. IEEE Trans. Intell. Veh. **1**(1), 105–120 (2016)
3. C. Maple, Security and privacy in the internet of things. J. Cyber Policy **2**(2), 155–184 (2017)
4. J. Petit, S.E. Shladover, Potential cyberattacks on automated vehicles. IEEE Trans. Intell. Transp. Syst. **16**(2), 546–556 (2014)
5. D. Dominic et al., Risk assessment for cooperative automated driving, in *Proc. 2nd ACM Workshop Cyber-Phys. Syst. Security Privacy (CPS-SPC)*, 2016, pp. 47–58

6. M.M. Islam et al., A risk assessment framework for automotive embedded systems, in *Proc. 2nd ACM Workshop Cyber-Phys. Syst. Secur.*, 2016, pp. 3–14
7. K. Kim, J.S. Kim, S. Jeong, J.-H. Park, H.K. Kim, Cybersecurity for autonomous vehicles: Review of attacks and defense. Comput. Secur. **103**, 102150 (2021)
8. Z. El-Rewini et. al., Cybersecurity challenges in vehicular communications. Veh. Commun., **23** (2020). https://doi.org/10.1016/j.vehcom.2019.100214
9. L. Sequeira et al., A lane merge coordination model for a v2x scenario, in *Proc. Eur. Conf. Netw. and Commun. (EuCNC)*, 2019, pp. 198–203
10. J. Ziegler et al., Making bertha drive—an autonomous journey on a historic route. IEEE Intell. Transp. Syst. Mag. **6**(2), 8–20 (2014)
11. A. Bolovinou et al., Tara+: Controllability-aware threat analysis and risk assessment for l3 automated driving systems, in *IEEE Intell. Vehicles Symp.*, June 2019, pp. 8–13
12. N. Vignard et al., Deliverable 4.2 Legal requirements to AD piloting and cyber security analysis, L3 Pilot Driving Automation, Tech. Rep., 04 2019. [Online]. Available: https://bit.ly/3o7FzBr
13. J.-P. Monteuuis et al., Sara: Security automotive risk analysis method, in *Proc. 4th ACM Workshop on Cyber-Phys. Syst. Secur.*, 2018, pp. 3–14
14. C. Yan, W. Xu, J. Liu, Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle, in *DEF CON*, vol. 24, 2016
15. S. Thys, W. Van Ranst, T. Goedemé, Fooling automated surveillance cameras: adversarial patches to attack person detection, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1105–1112
16. D. Song et al., Physical adversarial examples for object detectors, in *12th {USENIX} Workshop on Offensive Technologies*, 2018
17. S.-T. Chen et al., Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector, in *Proc. Joint European Conf, on Machine Learning and Knowledge Discovery in Databases* (Springer, 2018), pp. 52–68
18. ETSI, Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Analysis of the Collective Perception Service (CPS) (2017). [Online]. Available: https://bit.ly/35hiGBC
19. J. Redmon, A. Farhadi, Yolo9000: Better, faster, stronger, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7263–7271
20. M. Sharif et al., Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, in *Proc. ACM SIGSAC Conf. on Comput. and Commun. Secur.*, 2016, pp. 1528–40
21. K. Eykholt et al., Robust physical-world attacks on deep learning visual classification, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2018, pp. 1625–1634
22. SAE On-Road Automated Driving Committee, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, jun 2021
23. O. Y. Al-Jarrah et al., Intrusion detection systems for intra-vehicle networks: A review. IEEE Access **7**, 21266—21289 (2019)