Detecting cyber-physical threats in an autonomous robotic vehicle using Bayesian Networks

Anatolij Bezemskij, George Loukas, Diane Gan, Richard J. Anthony
Department of Computing & Information Systems
University Of Greenwich
London, United Kingdom
Email: {a.bezemskij, g.loukas, d.gan, r.j.anthony}@gre.ac.uk

***Abstract*— Robotic vehicles and especially autonomous robotic vehicles can be attractive targets for attacks that cross the cyber-physical divide, that is cyber attacks or sensory channel attacks affecting the ability to navigate or complete a mission. Detection of such threats is typically limited to knowledge-based and vehicle-specific methods, which are applicable to only specific known attacks, or methods that require computation power that is prohibitive for resource-constrained vehicles. Here, we present a method based on Bayesian Networks that can not only tell whether an autonomous vehicle is under attack, but also whether the attack has originated from the cyber or the physical domain. We demonstrate the feasibility of the approach on an autonomous robotic vehicle built in accordance with the Generic Vehicle Architecture specification and equipped with a variety of popular communication and sensing technologies. The results of experiments involving command injection, rogue node and magnetic interference attacks show that the approach is promising.**

I. Introduction

Due to their dependence on sensing, communication and artificial intelligence, cyber-physical systems, such as cars, drones and unmanned vehicles are attractive targets for attacks that cross the cyber-physical divide [1], [2], [3], from forcing a car to veer off road, to hijacking a drone or overwhelming a driverless car's lidar sensors. Here, we use the terminology introduced in [2], where a cyber-physical attack is a security breach in cyber space that has an adverse effect in physical space, and vice-versa, a physical-cyber attack is a security breach in physical space with adverse effect in cyber space. Detecting such threats is challenging, especially for resource-constrained systems, where highly accurate intrusion detection algorithms cannot be run on board and continuously. Here, we present an intrusion detection approach that is based on Bayesian Networks, and is able to determine not only whether there is an attack, but also from what domain it has originated (cyber or physical).

II. Related Work

The security of cyber-physical systems and especially of vehicles is a relatively new area of study. Relevant research has focused primarily on proof-of-concept attacks [4] on the integrity of sensing and actuation or the availability of communications. In most cases, the proposed defence is limited to survivability and resilience through redundancy [5] or prevention through authentication and encrypted communication [6]. However, this is an overly optimistic approach, as attacks, especially zero-day attacks, do get through these defences and so need to be detected. The focus here is on intrusion detection techniques designed specifically for mobile cyber-physical systems and robotic vehicles. Depending on its architecture and application, a robotic vehicle may be able to benefit from communication with other agents or may need to rely solely on its own sensing capabilities and monitoring processes.

There is on-going research on the development of an intrusion detection system specifically for the in-vehicle network. Waszecki et al. [7] have proposed monitoring internal network traffic using a simple Leaky Bucket approach. They have applied this approach to a single CAN bus feature, which is the frame arrival time. In this manner, although the particular approach is very limited and is not easily transferable to other aspects of a vehicle's operation, it has indeed demonstrated that a relatively simple and lightweight method is capable of detecting some malicious activity on the bus.

Taking into account the system's resource restrictions, other approaches may still apply, such as the work by Kang et al. [8], who have proposed the use of a Deep Neural Network (DNN) for monitoring the CAN bus network and detecting malicious activity. However, DNNs are computationally heavy as they require a lot of processing power to teach the neurons using the data, and with resource constraint systems this approach would be unlikely to be integrated. In contrast, the work by Vuong et al. [9], [10], [11] using the relatively lightweight approach of decision trees trained on existing attacks is highly practical, but can only work for known attacks.

A further step was made by Theissler et al.[12], where multiple methodologies were combined to form a hybrid for detecting known and unknown attacks in automotive systems using an ensemble-based anomaly detection approach. They have used four Two-Class classifiers (Mixture of Gaussians, Naive Bayes, Random Forest and Support Vector Machines) and four One-Class classifiers (Extreme-Value, Mahalanobis, One-Class Support Vector Machine and a Support Vector Data Description). All these classifiers in combination have shown excellent results for known faults, as well as for

unknown faults. A One-Class Support Vector Machine based approach has proven useful in cyber threat identification of autonomous avionic systems [13], where researchers were able to detect Teardrop, Fuzzing, Port Scan and ARP scan attacks. Loukas et al. [14] have shown that very accurate, but also computationally heavy approaches, such as deep learning can also be used if offloaded to a more powerful infrastructure, as long as the network is sufficiently reliable. This can both reduce detection latency and perhaps more importantly also reduce energy consumption, but of course, it has the major drawback that it depends on the availability of an offloading infrastructure, which is impractical in many application areas of robotic vehicles.

In general, most detection approaches for vehicles are explicitly or implicitly system-specific. There is a need for an approach that can be applied across a variety of vehicles and can adapt by learning what is normal for that vehicle and detecting deviations, so that it is also applicable to unknown/future threats. In addition, existing detection approaches can only tell whether there is an attack in progress, not what domain it has originated from (cyber or physical). Here, we present an approach that addresses both gaps in the landscape of related research.

III. Methodology

A. Experiment Setup

For the purpose of this research, we have developed a modular robotic vehicle (Figure 1) from the ground up, as described in detail in [15], [16]. Here, we limit the discussion of the testbed to a high-level overview.

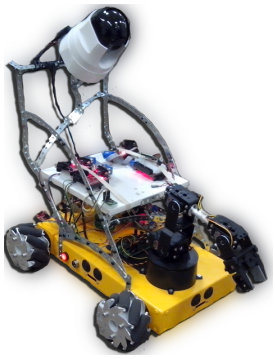

Fig. 1. Robotic vehicle testbed

The robotic vehicle was developed with practicability in mind. We have used the Generic Vehicle Architecture (GVA) [17] to keep the testbed modular, scalable and representative of existing vehicular systems. It is equipped with a variety of sensors, actuators and communication protocols widely used in the industry, including CAN, RS-485, WiFi and ZigBee. Communication between the operator and the testbed is carried out using ZigBee, while the audio/video feed is transmitted over a WiFi connection. This topology can be seen in Figure 2. Internal communication is carried out using the CAN bus, while RS-485 and I2C are used to communicate through the gateways. The monitored features

TABLE I
Description of the data sources

Physical Features		Cyber Features	
Name	Abbr.	Name	Abbr.
Battery Voltage	DS2	Packet Arrival Time	DS1
Compass Bearing	DS3	Action Indicator	DS11
Pitch	DS4	Sequence Number	DS12
Roll	DS5	CAN Packet Rate	DS13
Front Distance	DS6		
Back Distance	DS7		
Left Distance	DS8		
Right Distance	DS9		
Temperature	DS10		
Motor #1	DS14		
Motor #2	DS15		
Motor #3	DS16		
Motor #4	DS17		

and their abbreviations are illustrated in Table I. Note that by cyber features we refer to features corresponding to data processing and transmission (DS1 and DS11-DS13), and by physical features to the ones corresponding to actuation and sensing (DS2-DS10 and DS14-DS17). The data from the data sources is being collected and aggregated every 1 s. This is due to the fact that the sampling rate for the different features ranges between 30 Hz and 0.5 Hz and a communication protocol optimisation of the particular ZigBee hardware implementation we used limits it to a new transmission every 1 s.

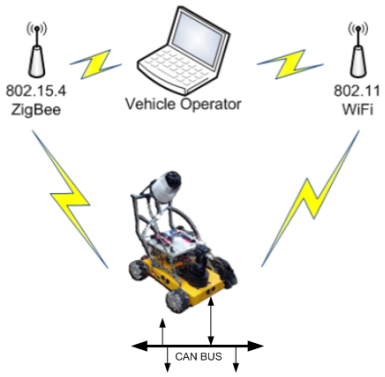

Fig. 2. Attack vectors of the attack scenarios

The autonomy of the robotic vehicle allows the system to undertake several missions. The results discussed here have been derived from a routine mission scenario, where the robotic vehicle has a target which is to reach a destination, on the ground, with stochastic elements that divert the robotic vehicle testbed.

B. Cyber-Physical Attacks

Here, we consider attacks of low to medium complexity, originating from both the cyber and physical domain (Figure 3):

- Cyber-Physical: False Data Injection, replaying sensor setup data packets on a ZigBee network. The assumption is that the attacker was able to collect commu-

nication data externally at an earlier stage, e.g. at the production stage.
- Cyber-Physical: Rogue Node, replaying packets on a CAN bus. The rogue node is assumed to have been planted through a supply chain attack. The aim here is to amplify the traffic transmitted on the CAN bus.
- Physical-Cyber (sensory channel attack): Magnetic disruption of compass readings, with the aim to impede the navigation capabilities of the autonomous vehicle.

C. Heuristic Binary Classification

The data that we feed into the Bayesian Network system presented here is the output of a heuristic binary classification mechanism described in [15], [16]. The particular mechanism uses an anomaly detection approach based on the defined signature characteristics which are described in Table II.

TABLE II
SIGNATURE CHARACTERISTICS

Value Type	Characteristic
Raw	Minimum Maximum
Exponential Smoothing	Minimum Maximum Lowest Difference Highest Difference
Deviation	Standard Deviation (Std)
Spike Regions	0.5*Std - 1.0*Std 1.0*Std - 1.5*Std 1.5*Std - 2.0*Std Over 2.0*Std

The mechanism itself learns the signature characteristics from the data being transmitted to the CAN bus. The key feature is that it uses a simple form of generalisation of the sensor into the data source format thus producing a behaviour signature for the sensor ignoring specific sensor context, such as distance, temperature or bearing. It extracts the metadata from the numeric data stream. This data set is transformed into a data behaviour format which is defined as a set of signature characteristics. The example of such metadata extraction is to observe a frequency of occurrences of the data samples within the deviation region (for example the region between 0.5 and 1.0 of the standard deviation). These occurrences are observed during a specific time slot accompanied by the learnt range between the minimum and the maximum from the data samples. These attributes act as the boundaries of an anomaly filter i.e. any incoming data samples violating these boundaries will raise an anomaly on one of the characteristics. The heuristic binary classifier uses raw data values as an input producing a binary output in a signature format as described in Table II.

D. Bayesian Network Implementation

The heuristic method presented above has been implemented as a prototype in a restricted resource environment and achieves reasonably high detection rates. However, it does not have the ability to tell anything more about the nature of a threat beyond its existence. To address this, we have added a complementary mechanism using Bayesian Networks, which have been previously used in evaluating cyber-threats in smart grids[18] or evaluating cyber-security risks in nuclear instrumentation and control systems[19]. Here, we use them to determine the domain from which a threat originates.

Bayesian networks can be used in statistical analysis providing a probability of events based on certain evidence. We have already described that the heuristic binary classification method is capable of identifying normal behaviour using a sensor agnostic approach, i.e. not taking into account the sensor context information. The methodology uses a self-learning approach to generalise the sensor data into signatures and use these signatures as the data source's unique description that demonstrates a sensor's specifics. We have used a Bayesian network based approach because it is capable of working with discreet data, that can be in any generic form. In addition, Bayesian networks are able to infer the unknown variables which are useful in a situation where an intrusion detection mechanism has to make a decision based on the request of the operator or any other on-demand request. There is obviously a vast range of other popular models that can be used, such as Neural Networks, Decision Trees or Random Forest, but they cannot infer the unknown variables within a reasonable precision range. This capability provides multiple ways on how to use such a model. As it can learn and create relationships between the nodes, no expert knowledge is required to identify the conditional probabilities of various events. Here, we have used the statistical analysis environment **R**, which is an open-source statistical analysis environment with publicly available libraries for Bayesian networks [20] provided by the **bnlearn** library. The process of threat domain identification is illustrated in Figure 4. An initial process starts by training a heuristic binary classifier (1) that will learn the behaviour of all data sources. Thereafter, the output of the heuristic binary classifier is fed to the Bayesian Network as an input to train the Bayesian Network model (2). When the training phase has been completed, the sensor readings are classified using the heuristic binary classifier (3) and then the output is used to query the Bayesian Network on the probability of the Normality, Cyber Threat or Physical Threat given the evidence (4). This methodology becomes a hybrid that is using both unsupervised and supervised learning requiring prior knowledge of a threat situation. When the learning phase has been completed, the raw data from the data sources is fed to the heuristic binary classifier to transform the data in an anomaly signature format. The output is then used as a query argument to the Bayesian network.

The first step is to identify the relationships between the entities in the data. Researchers have published a variety of algorithms for identification of relationships between entities for Bayesian Networks, which have their strengths and weaknesses. We have used the Hill-Climbing algorithm to construct a Direct Acyclic Graph (DAG), which creates all connections so that the graph does not have cycles or disconnected entities in the end. Other provided algorithms

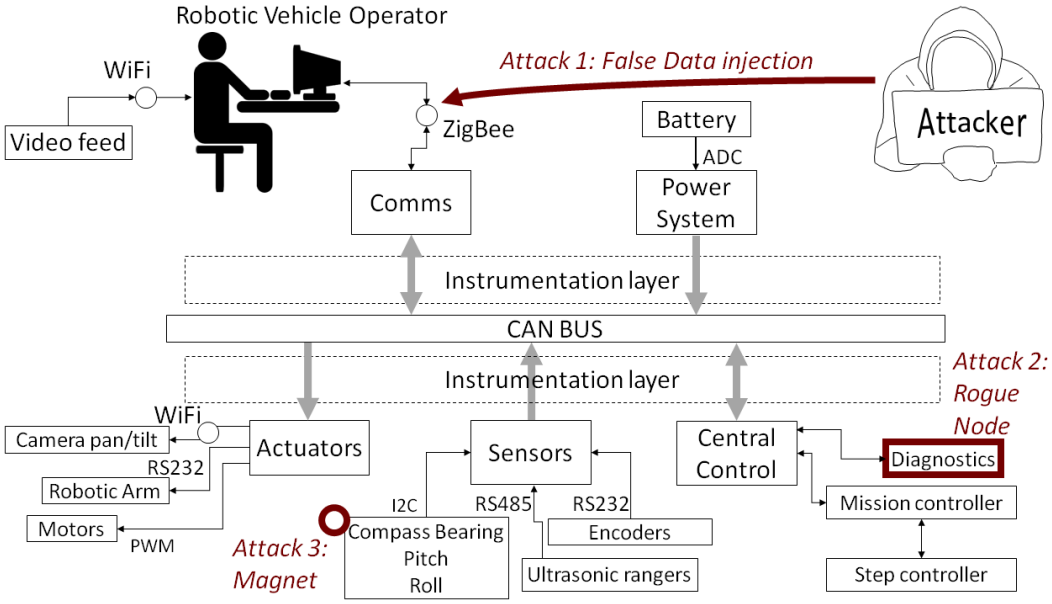

Fig. 3. Attack vectors of the attack scenarios

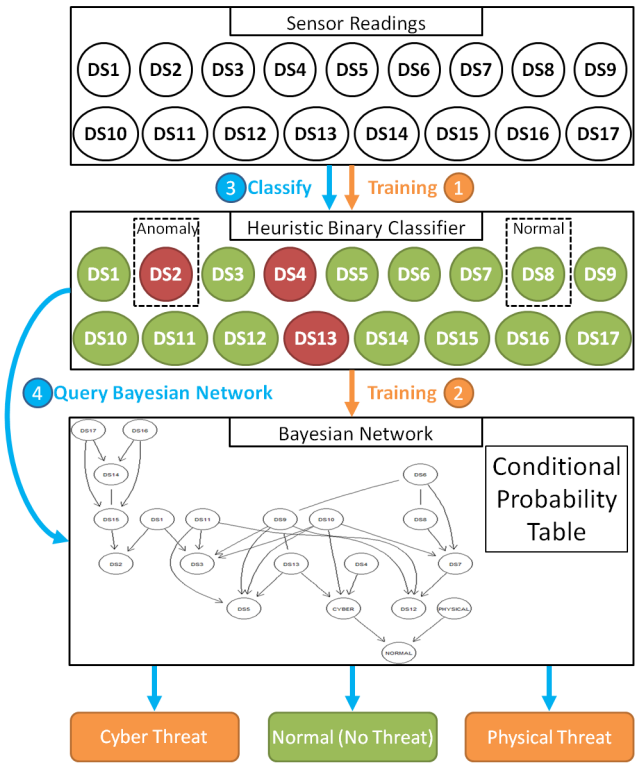

Fig. 4. Raw sensor data is fed to the Heuristic binary classifier where data sources are classified as having normal behaviour (green) or anomalous (red), then the Bayesian Network is using the output of the heuristic binary classifier as an input

were used, including Incremental Association Markov Blanket, Max-Min Parents & Children, Tabu Search and other algorithms that are provided by the **bnlearn** library packages. The Hill-Climbing algorithm demonstrated that it is capable of generating a closed DAG taking into account all entities that are given to the Bayesian Network. The weakness of the Bayesian network is that it uses a Supervised Learning approach, in that the data set needs to have data for the events that are being queried. In this work, we use a data set with **Normal**, **Cyber Attack** and **Physical Attack** data. These are the events that will be queried, given the evidence, which is the data source heuristic binary classification output. We used a 70/30 training/testing dataset split.

IV. Performance Evaluation

Here, we use Receiver Operator Characteric (ROC) curves to evaluate the performance of the approach in terms of true positive and false positive rates, and specifically the Area Under Curve (AUC) metric, which is a standard approach in classification comparison. Figure 5 contains a variety of events that the Bayesian network is being queried for. The evaluated cases are the detection of Normal Behaviour, Cyber Threat Behaviour and Physical threat behaviour. For reference, we also include the case where detection would be random (the (0,0)(1,1) line, with AUC score of 0.5).

We observe very high accuracy of detection of cyber attacks, and a little lower accuracy for normal states and physical attacks. This is due to mission behaviour which is producing a large amount of noise which is cancelling out the attacks themselves. However, it is still performing well and can produce a high probability identification of a threat domain from the learnt data set. We have also experimented with a variety of data sources looking at the cyber features and physical features separately.

Figure 6 demonstrates the performance of threat detection using only cyber features. It is noticeable that the performance of cyber threat identification has not changed,

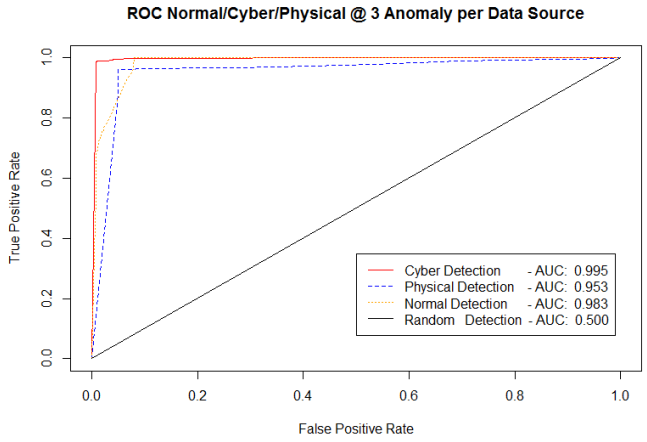

Fig. 5. Bayesian network performance identifying cyber-physical domain threat using both physical and cyber features

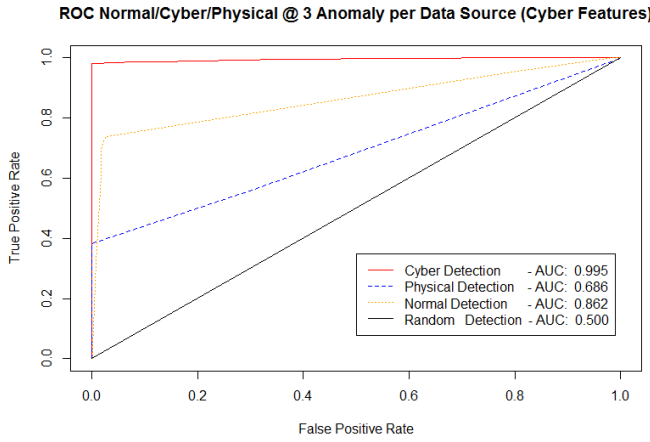

Fig. 6. Bayesian network performance identifying cyber-physical domain threat using only cyber features

but, as expected, detection of physical domain attacks and of normal behaviour have decreased. This shows that it is beneficial to monitor not only cyber features, as in conventional computing systems, but also using physical features. Physical threat detection shows promising results as it is capable of identifying physical threats with a lower confidence level, based only on the cyber features. This means that it is necessary to monitor the physical features as they are affected by the malicious activity of the system.

Figure 7 demonstrates the capability of this methodology to produce accurate probabilities by only monitoring the physical features. The detection rate produces reasonably high performance for cyber-physical domain threat detection. However, we observe an interesting fact if we compare Figure 7 with Figure 5 and focus on Physical domain detection. AUC performance is slightly lower than that produced in Figure 7. This shows that there is a potential situation when the combinations of various domain features may act as noisy evidence when the probability is calculated for a specific

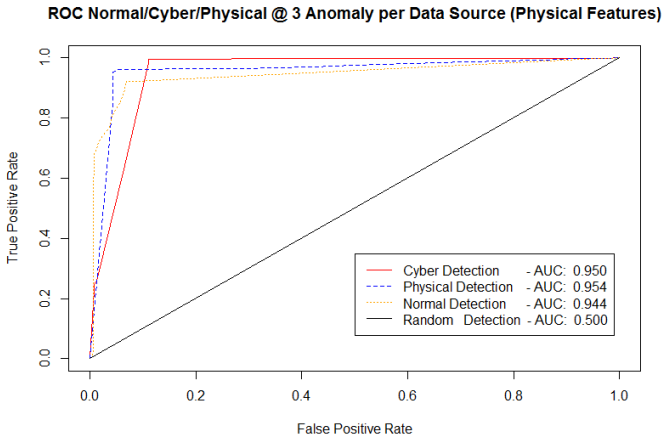

Fig. 7. Bayesian network performance identifying cyber-physical domain threat using only physical features

event, probing cyber and physical threat domains. However, the difference is relatively small. Using all cyber and physical features generally leads to improved identification of normal behaviour as well as of cyber domain threats.

V. Conclusion

The challenge that we have addressed here is how to determine the domain from which a threat has originated in a cyber-physical system vehicle, such as an autonomous vehicle, which relies on the integrity and availability of a variety of sensing and communication technologies to perform its mission. To address this, we have presented a mechanism based on Bayesian networks which can receive information in real-time from a vehicle's sensors, processing and communication modules and determine whether there is an attack and if so, whether it originates from the cyber or the physical domain. This is particularly useful for attacks that cross the cyber-physical divide, such as a sensory channel attack (here, magnetic interference) affecting the ability to reason with correct data and carry out a mission, and cyber security breaches (e.g. a rogue node or command injection attack) affecting a vehicle's physical behaviour. A limitation of this mechanism that we will address in future work is that it performs effectively only for the relatively low sampling frequency of one sample per second. That is because with a higher sampling frequency, lack of precise synchronisation between the heterogeneous data sources used would mean that the data could be received out of order. Furthermore, we will apply the approach on much larger and much smaller vehicles to evaluate its scalability and applicability across different systems.

Acknowledgments

This research has been funded and supported by the Defence Science and Technology Laboratory. We thank Robert Sayers for his invaluable assistance and feedback.

REFERENCES

[1] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, and S. Savage, "Experimental security analysis of a modern automobile," in *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 2010, pp. 447–462.

[2] G. Loukas, *Cyber-Physical Attacks: A Growing Invisible Threat*. Butterworth-Heinemann, 2015.

[3] J. Petit and S. E. Shladover, "Potential cyberattacks on automated vehicles," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 16, no. 2, pp. 546–556, 2015.

[4] K. Koscher *et al.*, "Experimental security analysis of a modern automobile," in *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 2010, pp. 447–462.

[5] A. Deshpande, O. Obi, E. Stipidis, and P. Charchalakis, "Integrated vetronics survivability: Requirements for vetronics survivability strategies," in *System Safety, 2011 6th IET International Conference on*. IET, 2011, pp. 1–6.

[6] M. Wolf, A. Weimerskirch, and C. Paar, "Secure in-vehicle communication," in *Embedded Security in Cars*. Springer, 2006, pp. 95–109.

[7] P. Waszecki, P. Mundhenk, S. Steinhorst, M. Lukasiewycz, R. Karri, and S. Chakraborty, "Automotive electrical/electronic architecture security via distributed in-vehicle traffic monitoring," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2017.

[8] M.-J. Kang and J.-W. Kang, "Intrusion detection system using deep neural network for in-vehicle network security," *PloS one*, vol. 11, no. 6, p. e0155781, 2016.

[9] T. Vuong, A. Filippoupolitis, G. Loukas, and D. Gan, "Physical indicators of cyber attacks against a rescue robot," in *IEEE International Conference on Pervasive Computing and Communications*. IEEE, 2014, pp. 338–343.

[10] T. P. Vuong, G. Loukas, D. Gan, and A. Bezemskij, "Decision tree-based detection of denial of service and command injection attacks on robotic vehicles," in *Information Forensics and Security (WIFS), 2015 IEEE International Workshop on*. IEEE, 2015, pp. 1–6.

[11] T. P. Vuong, G. Loukas, and D. Gan, "Performance evaluation of cyber-physical intrusion detection on a robotic vehicle," in *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2106–2113.

[12] A. Theissler, "Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection," *Knowledge-Based Systems*, 2017.

[13] S. G. Casals, P. Owezarski, and G. Descargues, "Generic and autonomous system for airborne networks cyber-threat detection," in *Digital Avionics Systems Conference (DASC), 2013 IEEE/AIAA 32nd*. IEEE, 2013, pp. 4A4–1.

[14] G. Loukas, Y. Yoon, G. Sakellari, T. Vuong, and R. Heartfield, "Computation offloading of a vehicles continuous intrusion detection workload for energy efficiency and performance," *Simulation Modelling Practice and Theory*, 2016.

[15] A. Bezemskij, R. J. Anthony, D. Gan, and G. Loukas, "Threat evaluation based on automatic sensor signal characterisation and anomaly detection," in *Proceedings of The Twelfth International Conference on Autonomic and Autonomous Systems*. IARIA, 2016, pp. 1–7.

[16] A. Bezemskij, G. Loukas, R. J. Anthony, D. Gan *et al.*, "Behaviour-based anomaly detection of cyber-physical attacks on a robotic vehicle," 2016.

[17] F. Bergamaschi, D. Conway-Jones, and N. Peach, "Generic vehicle architecture for the integration and sharing of in-vehicle and extra-vehicle sensors," in *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2010, pp. 76 940B–76 940B.

[18] A. Le, Y. Chen, K. K. Chai, A. Vasenev, and L. Montoya, "Assessing loss event frequencies of smart grid cyber threats: Encoding flexibility into fair using bayesian network approach," in *Smart Grid Inspired Future Technologies: First International Conference, SmartGIFT 2016, Liverpool, UK, May 19-20, 2016, Revised Selected Papers*. Springer, 2017, pp. 43–51.

[19] J. Shin, H. Son, and G. Heo, "Cyber security risk evaluation of a nuclear i&c using bn and et," *Nuclear Engineering and Technology*, 2016.

[20] M. Scutari, "Learning bayesian networks with the bnlearn r package," *arXiv preprint arXiv:0908.3817*, 2009.