Exploration of Explainable AI in Context of Human-Machine Interface for the Assistive Driving System

Zenon Chaczko[1,3] [0000−0002−2816−7510],
Ilya Thai-Chyzhykau[1] [0000−0001−6499−2115],
Marek Kulbacki[2,3] [0000−0003−4609−106X],
Grzegorz Gudzbeler[4] [0000−0002−9169−5543], and
Peter Wajs-Chaczko[5] [0000−0003−2079−746X]

[1] University of Technology Sydney, Faculty of Engineering and IT, NSW, Australia
[2] Polish-Japanese Academy of Information Technology, R&D Center, Warsaw, Poland
[3] DIVE IN AI, Wroclaw, Poland
[4] Faculty of Political Science and International Studies University of Warsaw, Warsaw, Poland
[5] Macquarie University NSW, Australia

zenon.chaczko@uts.edu.au, ilya.thai-chyzhykau@uts.edu.au, mk@pja.edu.pl,
grzegorz.gudzbeler@gmail.com, peter.Wajs-Chaczko@students.mq.edu.au

Abstract. This paper presents the application and issues related to explainable AI in context of a driving assistive system. One of the key functions of the assistive system is to signal potential risks or hazards to the driver in order to allow for prompt actions and timely attention to possible problems occurring on the road. The decision making of an AI component needs to be explainable in order to minimise the time it takes for a driver to decide on whether any action is necessary to avoid the risk of collision or crash. In the explored cases, the autonomous system does not act as a "replacement" for the human driver, instead, its role is to assist the driver to respond to challenging driving situations, possibly difficult manoeuvres or complex road scenarios. The proposed solution validates the XAI approach for the design of a safety and security system that is able to identify and highlight potential risk in autonomous vehicles.

Keywords: Explainable AI · HMI · Convolutional Neural Network · Assistive System for Vehicles

1 Background

The Artificial Intelligence (AI) component of many modern systems is often perceived as a black box [2], where data goes in, and a prediction goes out. The inner mechanisms of the module are often not well understood even by the designers of the solution. This, however, poses a serious problem in the real-world solutions, where a human is unable to qualify the prediction accurately without additional and time-consuming investigation. As the AI mechanism (module) is represents an enclosed "black box" subsystem, it is not easy to determine the real factors that influenced the computational component to provide result A over result B. Looking at how users perceive an explanation, it can be seen that an explanation is influenced by the cognitive bias, contrasting events that occurred, social beliefs and often may refer to cause rather than a statistical measure [4, 6, 10]. In most implemented algorithms, the statistical measure is often the qualifying value to choose event A over B, regardless of the "cause" of the event. This occurs since computer algorithm machine is unable to process the entire context of an event, and produce an explanation based on the cause. Rather, the algorithm looks for patterns that it has "learned" and provides an output based on the possible results set by humans. Hence, if the computer program is unable to provide "reasoning" to create a satisfactory explanation, then how can an Explainable Artificial Intelligence (XAI) [2, 7, 12] exist? The underlying research project investigates an approach into how an intelligent system can, not only make a sound prediction [3], but also explain or reason the outcome to a human user. In doing so, enhancing the Human-Machine Interaction [1] and promoting trust [16]. The investigation looked into boundary (edge) values case scenarios of autonomous vehicles, taking into account the time constraints of human users to react, and the fact that an explanation has to be easy to understand, provide a sound level of explanation and can be processed by the human operator as quickly as possible. These factors influence the overall design. The main goal of the designed prototype is to have a system that would detect, analyse and provide explanation of the predicted risks within the edge scene of an autonomous vehicle [3]. This approach provides a sufficiently wide range of base cases, where the system could use the AI module for risk detection and provide a satisfactory level of reasoning. The design of the project was separated into two primary parts:

- Training a Convolutional Neural Network (CNN)
- Create a system that delivers CNN output and compatible reasoning

1.1 Implemented Convolutional Neural Network

Given the "real time" constraints of the project, CNN by itself is unable to provide quick reasoning for the prediction it makes. The predictions made by the CNN are based on the patterns the model was trained on [8]. Considering the need for explanation, a design decision was made to train a model to detect specific entities, referred to as Vulnerable Road Users (VRUs). This approach added an extra layer of explainability to the overall system, as the CNN's primary goal was to detect certain entities within the scene, but not do the actual risk analysis. Chosen architecture was also able to localize the detected object, thus providing further reasoning for the decision made.

The chosen CNN was based on the Singe Shot Multi-Box Detector (SSD) architecture (Fig. 1), which scans the image once, and utilizes bounding box proposals to detect entities with an image [13]. VGG16 was adopted as the backbone feature extractor for SSD given its good performance and good accuracy in non-high-performance equipment.

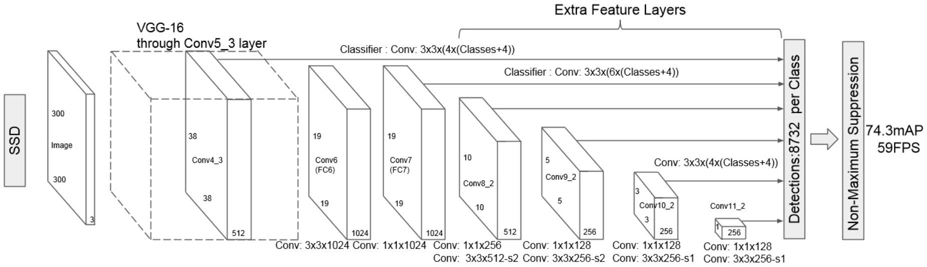

Fig. 1. SSD Network Architecture ([13])

CNN was trained on manually collected dataset comprised of people talking and looking at mobile phones (otherwise referred to as VRUs), cyclists and general pedestrians.

Table 1. Dataset distribution

Label Name	Total	Training / Test	Validation
VRU_Pedestrians	234	180	54
Cyclist	150	135	15
People	200	150	50

Low volume of dataset posed a great challenge for training of CNN. However, through training techniques such as transfer learning and further hyper-parameters tuning, the final accuracy of model was around 73 points. For an initial prototype and given the time and resource constraints of the project, this result was deemed "good enough".

2 System Design

As the accuracy of the CNN model was low, the second part of the project, was designed to integrate precautionary measures to further analyse the risk and actions required. The below diagram (Fig. 2) depicts relations between entities integrated in the system, as well as the communication sequence between the modules. Unlike early attempts of out-of-the-loop automatic control and performance evaluation of such systems [5], various functions of the proposed Assistive Driving System (ADS) operate within a continuous loop process that includes the input video streams, identifies objects, performs analysis of risks, and generates actions according to the identified risks in context of HMI [1].

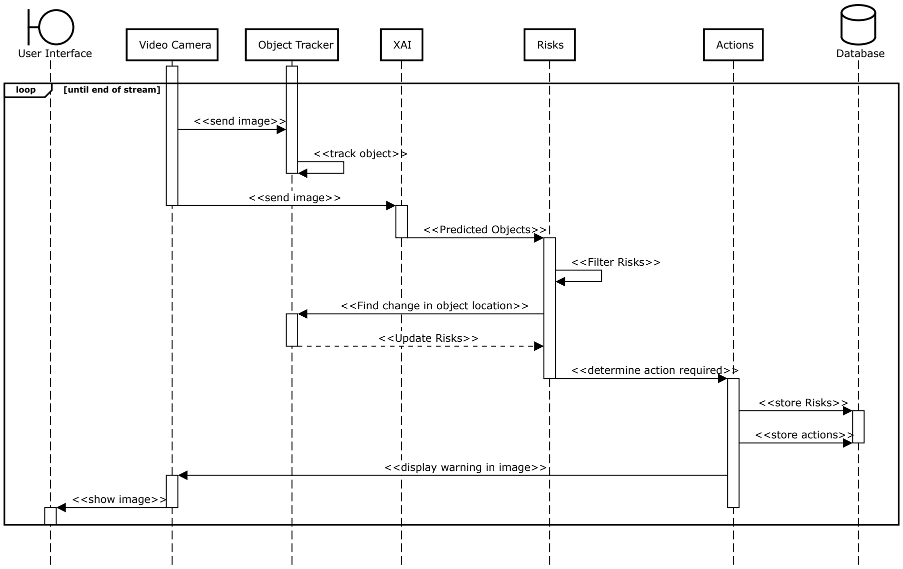

Fig. 2. Communication sequences, modules and entities of the system

The intention behind “Risks” module was to analyse the identified object in the context of risk of collision. Each risk was calculated within the context of a single entity, however, further risk analysis is possible where all entities are considered as a group, thus provisioning a way to calculate contextual overall environment risk. This further calculation and analysis was not within the scope of the project. Risk of collision was calculated based on factors including:

- Total speed (subject + vehicle speed)
- Risk factor based on subject location
- Distance to subject (calculated using “similar triangles” property and defined by equation below)

$$Distance\,to\,object = \frac{Real\,Object\,Height * Focal\,length}{Object\,Height\,in\,Image} \tag{1}$$

“Actions” primary focus is to advise the vehicle operator of identified risks using a number of visual cues and messages. Following the underlying project theme of “explainability”, a big design focus was on figuring out how to warn the vehicle operator of the upcoming risk in the most efficient way. An optimal solution was to use a col-our indicator based on the analysed risk. Risk classification (see Table 2) in the previous module would provide details of the consequence and the likelihood of a collision event. The colour coding match those used on road (i.e., Safety Signs), and thus are easy to understand the meaning of. The localized box around the potential risk would utilise the colour coding, as well as limited meta-data constrained to:

- Warning icon (representative of the identified risk factor)
- Distance to object in meters
- Likelihood of collision value

Table 2. Risk Classification

Consequence						
		Insignificant	Minor	Moderate	Major	Severe
Likelihood	Rare	Low	Low	Low	Medium	High
	Unlikely	Low	Low	Medium	High	High
	Possible	Low	Medium	Medium	High	Extreme
	Likely	Medium	Medium	High	Extreme	Extreme
	Almost Certain	Medium	High	High	Extreme	Extreme

3 Results and Evaluation

The results were split based on the two distinct project phases: CNN design and overall system design. CNN model testing was evaluated based on the accuracy, loss, validation loss and validation accuracy metrics. During the tests, there were issues raised if the model was over-fitting to a rather limited dataset available, as this was a primary risk of the CNN design. As seen in the table below, a number of hyper-parameters were turned at each test. Each test design was done iteratively, as the main aim was to identify the parameters that had the best outcome. The overall system was tested using two approaches. This is mainly due to the fact that “explanation”, by its nature, is qualitative and cannot be quantitatively judged, as what determines it to be a “good” explanation. What humans consider to be ”good” explanation is usually based on the reasoning, personal bias (desires/intentions) and social belief. Hence, the system’s performance was judged on (1) quantitative values of proximity to real time (based on FPS) and the accuracy of prediction; and (2) qualitative values of ease of use (usability) and level of explanation evaluated by human users.

3.1 Quantitative Results

The quantitative tests (Table 3) and real-time proximity benchmarking was done using two different machines, with the calculated average FPS results (Table 4). Using a better performance model of GPU, the system was able to obtain a good level of “Frames Per Second” (FPS). This can be seen as an acceptable result as the output of the video was very close to real time, with minimum lag. However, the accuracy of the model did not perform as well as anticipated. At times, the model picked up “normal pedestrians” as potential risks. Further analysis and evaluation indicted that the AI module required a much larger training data set to achieve an acceptable accuracy for the live system implementation. Also, the test results indicated that the solution had some performance issues evaluating poor environmental conditions scenarios, as in such scenarios it was much harder to identify subjects. To address the described performance issues further experiments will involve a wider range of various difficult environmental conditions and a much larger training dataset.

Table 3. Quantitative Tests

TEST #	weight	Labels	Optimizer	Batch Size	Scheduler
0	vgg16_2_classes (COCO)	3 (risk)	SGD	8	Ir_schedule
1	vgg16_2_classes (COCO)	4 (risk)	SGD	8	Ir_schedule
2	vgg16_2_classes (COCO)	5 (risk)	SGD	8	Ir_schedule(updated to 6)
3	vgg16_2_classes (COCO)	6 (risk)	SGD	8	Ir_schedule(updated to 5)
4	vgg16_2_classes (COCO)	7 (risk)	SGD	8	PolynomialDecay(maxEpochs=20, initAlpha=1e-3, power=5)
5	vgg16_2_classes (COCO)	8 (risk)	SGD	8	PolynomialDecay(maxEpochs=20, initAlpha=1e-3, power=2)
6	vgg16_2_classes (COCO)	9 (risk)	SGD	8	PolynomialDecay(maxEpochs=20, initAlpha=1e-3, power=1)
7	vgg16_2_classes (COCO)	10 (risk)	SGD	8	PolynomialDecay(maxEpochs=20, initAlpha=1e-3, power=3)
8	vgg16_2_classes (COCO)	11 (risk)	Adam (lr=0.001, be)	8	PolynomialDecay(maxEpochs=20, initAlpha=1e-3, power=3)
9	vgg16_2_classes (COCO)	12 (risk)	Adam (lr=0.003, be)	8	PolynomialDecay(maxEpochs=20, initAlpha=1e-3, power=3)
10	vgg16_2_classes (COCO)	13 (risk)	SGD	8	PolynomialDecay(maxEpochs=20, initAlpha=1e-3, power=3)
11	vgg16_2_classes (COCO)	14 (risk)	SGD	8	PolynomialDecay(maxEpochs=20, initAlpha=1e-3, power=2)

TEST #	Epochs	Steps / epoch	Acc	Loss	val_loss	val_acc
0	20	120	0.67	2.681	3.2149	0.65
1	20	30	0.7359	2.9011	2.8885	0.68
2	20	30	0.7475	2.92	2.8929	0.7008
3	20	24	0.7480	3.0476	2.9868	0.7083
4	20	24	0.7194	3.0663	3.0066	0.6515
5	20	24	0.7198	2.7959	2.7637	0.6705
6	20	24	0.7248	2.8759	2.7143	0.6847
7	20	24	0.7249	2.8508	2.8967	0.6843
8	20	24	0.2277	6.1883	5.9808	0.2253
9	20	24	0.3674	6.0277	5.8737	0.3428
10	20	24	0.6611	nan	invalid loss	
11	20	24	0.7367	2.93	2.8858	0.6874

Table 4. Real Time Proximity Tests

System Spec	Average FPS	Note
Desktop, Intel i5 CPU Nvidia GTX 960 (4GB GPU)	14 – 18	The FPS saw a significant drop using a mobile phone as the input device, however, a potential cause of that is the poor WIFI available
Laptop, Intel i7 CPU Nvidia GTX 1050 TI (4GB GPU)	18 - 22	The laptop contained an in built camera, and seemed to be processing image stream a faster rate due to better Graphical Unit.

3.2 Qualitative Results

The qualitative evaluation and results was based on the feedback provided by a small group of users who were given a video on their mobile phones showing the streets with pedestrians and the assistive system evaluating risk of each pedestrian scenarios (Fig. 3). Prediction and Accuracy tests (Table 5) indicate cases where not all pedestrians at risk were detected and predicted as being at risk, as well as, cases of pedestrians evaluated by users at being at no risk, were indicted and predicted by the system as being a potential risk (Table 5). User feedback approach was used to evaluate and compare risk values based on a positive detection indicted by the system with risk values based on user perceptions (Table 6).

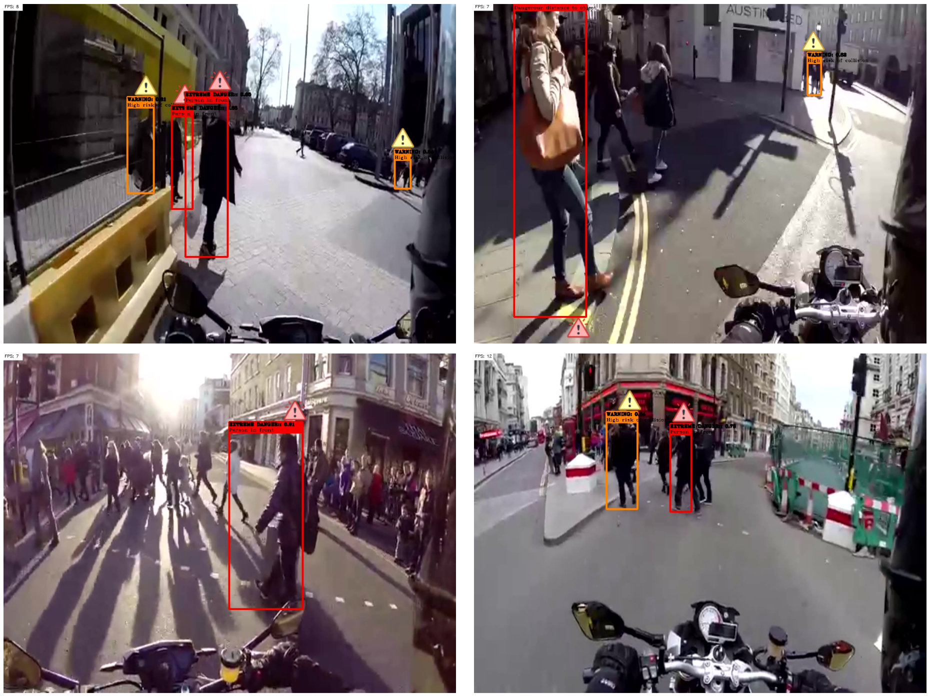

Fig. 3. CNN model testing using video captures

Table 5: Prediction and Accuracy Tests.

Scene	Description	Expected Detections	Positive Detections	False Positive Detections	Note
1	Single Pedestrian on mobile	1	1	0	
2	Multiple pedestrians on mobile	3	3	0	
3	Multiple pedestrians on a phone call	5	4	0	1 pedestrian not detected
4	Multiple pedestrians with mobile and Normal pedestrians	3-mobile 1-normal	3	1	Normal Pedestrian detected as risk
5	Normal Pedestrians	3	0	2	2 pedestrians detected risks
6	Single Cyclist	1	0	0	Cyclist not detected
7	Multiple cyclists	4	3	0	1 cyclist not detected
8	Cyclists and pedestrians on mobile	2-cyclists 6-mobile	2-cyclists 4-mobile	0	2 pedestrians on mobile not detected
9	No subjects in screen	0	0	0	
10	Multiple subjects on screen (lower visibility)	1-cyclist 2-normal pedestrians 1-mobile	0-cyclists 0-mobile	1	1 normal pedestrian detected as “with mobile”
End of Table 5					

Table 6. Risk Values Based on Positive Detections

Scene	Positive Detections	Correct Risk Value	Incorrect Risk Value
1	1	1	
2	3	3	
3	4	2	2 identified as high risk (low risk)
4	3	3	
5	0	0	
6	0	0	
7	3	2	1 identified as low risk (mid-high risk)
8	2 cyclists 4 mobile	3	1 (mobile pedestrian) identified as high risk (low-mid risk)
9	0	0	
10	0 cyclist 0 mobile	0	

4 Conclusion

Discussion on explainability of threats and risks detected by the Assistive Driving System conducted with a group of users, indicated their preference for an explainable AI system. The current implementation of the system is still lacking the required accuracy, as the users often saw a higher risk produced by the system, then they would otherwise visually and cognitively classify themselves. The user group feedback validated the preference of having short and high-quality information (fiducial markers), over long texts. The test users mentioned that in "edge case" situations, where quick decisions are necessary, they would prefer not to read a large amount of text, but rather, have the system identify and locate the risks on the screen. They also noted that too much "information" on the screen could potentially occlude objects and reduce the vision of the driver. This could create a potentially even more dangerous situation.

Acknowledgements This study was supported by the Scientific Research from Technical University of Technology Sydney, School of Electrical and Data Engineering, Macquarie University and DIVE IN AI.

References

1. Abdul, A.M., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.S.: Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018. p. 582 (2018). https://doi.org/10.1145/3173574.3174156
2. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052

3. Badue, C., Guidolini, R., Carneiro, R.V., Azevedo, P., Cardoso, V.B., Forechi, A., Jesus, L.F.R., Berriel, R.F., Paixão, T.M., Mutz, F.W., Oliveira-Santos, T., de Souza, A.F.: Self-driving cars: A survey. CoRR **abs/1901.04407** (2019)
4. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
5. Endsley, M.R., Kiris, E.O.: The out-of-the-loop performance problem and level of control in automation. Human Factors **37**(2), 381–394 (1995). https://doi.org/10.1518/001872095779064555
6. Flint, A., Nourian, A., Koister, J.: xai toolkit: Practical, explainable machine learning. Website (2019), https://www.fico.com/en/latest-thinking/white-paper/xai-toolkit-practical-explainable-machine-learning, (Accessed: 06.10.2019)
7. Goebel, R., Chander, A., Holzinger, K., Lécué, F., Akata, Z., Stumpf, S., Kieseberg, P., Holzinger, A.: Explainable AI: the new 42? In: Machine Learning and Knowledge Extraction - Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27-30, 2018, Proceedings. pp. 295–303 (2018). https://doi.org/10.1007/978-3-319-99740-7_21
8. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G.: Recent advances in convolutional neural networks. CoRR **abs/1512.07108** (2015)
9. Gunning, D.: Darpa's explainable artificial intelligence (XAI) program. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019 (2019). https://doi.org/10.1145/3301275.3308446
10. Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E.R. (eds.): Machine Learning and Knowledge Extraction - Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27-30, 2018, Proceedings, Lecture Notes in Computer Science, vol. 11015. Springer (2018). https://doi.org/10.1007/978-3-319-99740-7
11. Kaul, A.: Explainable ai and its impact on ai adoption. Website (2018), https://www.tractica.com/artificial-intelligence/explainable-ai-and-its-impact-on-ai-adoption/, (Accessed: 06.10.2019)
12. van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior. In: Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA. pp. 900–907 (2004)
13. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I. pp. 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2
14. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
15. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
16. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778