Will the *real* Operating Domain please stand up?

Philippa Ryan
Centre for Assuring Autonomy
University of York
York, United Kingdom
Email: philippa.ryan@york.ac.uk

John McDermid
Centre for Assuring Autonomy
University of York
York, United Kingdom
Email: john.mcdermid@york.ac.uk

***Abstract*—Every safety case should describe the deployment domain and environmental constraints within which the system is expected to operate. Recently many AI and autonomous system safety standards have proposed the use of a detailed formal description of the Operating Domain for autonomous safety critical systems. This OD is based on human understanding of variations and expected limitations, and is used to shape the data collection, testing, validation, verification and operational deployment of the system. For example, we assume an autonomous car will be driving on specific road layouts, with localised markings/signage, weather, and shared with a set of other defined road users. However, a Machine Learning (ML) components OD (e.g., that of a Deep Neural Network) is fundamentally different to ours, and is based on numerical data arrays. Our position is that over-reliance on a human-centred OD to shape V&V will lead to false confidence, and safety issues being missed. Instead we propose effort is spent reverse engineering the ML's view of the world, to better understand the OD's gaps, areas of uncertainty and hence derive strategies for how to mitigate related hazards.**

I. Introduction

The standard definition of a safety case is typically a variant of "a structured argument, supported by evidence, intended to justify that a system is acceptably safe for a specific application in a specific operating environment" [1]. One of the key elements is what the *specific operating environment* will be. Safety engineers will make assumptions about what could happen within that environment, as well as (usually) providing limited safety guarantees outside of it. For example, a particular system may be operated safely within a particular temperature range, but with limitations during high humidity. Physical components are engineered to work in that range, and simple sensor checks can determine when the safe ranges and limitations are breached.

This approach is justified where humans are involved in design and operation of the system, and where system components can be designed to work in the operating environment, and because a human operator will have a broadly similar understanding of it as the designers and can detect and act when the systems falls outside the defined operating environment. However, for AI-enabled Autonomous Systems (AS) neither of these conditions hold. AI, such as Machine Learning (ML) doesn't understand the context within which it is operating. Additionally, even when following robust engineering practice for AI/ML, there is a very high degree of uncertainty in the performance of black-box AI components such as ML in the domain they are designed for [2].

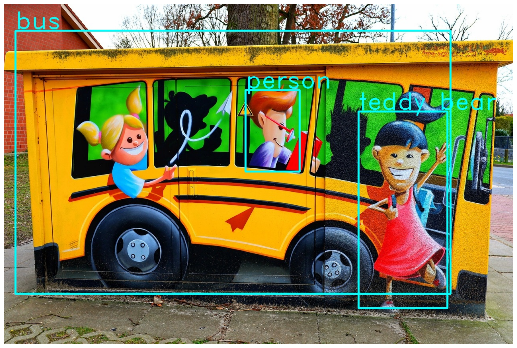

Fig. 1. Street art mis-identified as a bus by Yolo v5[6] (Picture courtesy of pexels.com)

Nevertheless, many AS standards (e.g., [3], [4]) are converging towards the development of a rigorously defined Operating Design Domain (ODD) within which the ML is *designed* to operate safely. SAE J3016 [4] defines the ODD as "Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics". [5] describes this in more broad terms as the Operating Domain Model (ODM), stressing the need to determine when we have fallen outside of the ODM and continue safely where possible. In all cases the ODM/ODD description will constrain the data for the training regime for ML components, and help to define the parameters for verification.

ML's lack of "understanding" of the ODM/ODD it has been trained to operate in leads to amusing situations (see Figure 1 where street art on a building has been misclassified as a bus), but also contributes to deadly consequences such as the fatal collision with a pedestrian in Tempe, Arizona shows [7]. The automated driving system failed to recognise the pedestrian pushing a bicycle, repeatedly changing classification and losing trajectory information, and failing to warn the distracted safety driver. Further, formalising and tightening the ODM/ODD, widening testing for more edge cases and even training with increasing amounts of data doesn't address the fundamental misalignment between the human described ODD and the ML's actual ODD. At best it will plug a few more holes. We may reduce some uncertainty, but the cost

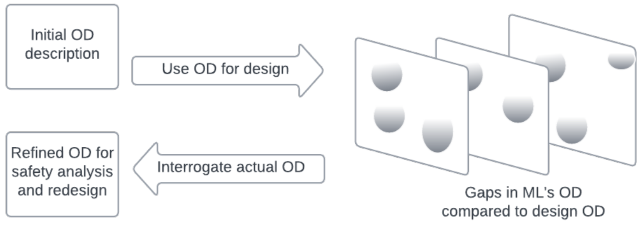
Initial OD description
Use OD for design
Refined OD for safety analysis and redesign
Interrogate actual OD
Gaps in ML's OD compared to design OD

Fig. 2. OD interrogation concept

and effort of doing so may not be proportional to the benefits. Worse, we may gain a false sense of confidence from doing so.

Our position is that effort could be more effectively targeted at interrogating the ML to determine its actual OD. Then we can identify appropriate system safety mitigations to manage those gaps and produce more effective out of OD detectors. This process is summarised in Figure 2. We first use the human defined OD for design of the ML. Then we interrogate the ML model to understand it's actual OD, finding the gaps represented on the right (which may vary over time and data inputs), to produce a refined OD description, which can be used for further safety analysis.

II. Possible approaches and research challenges

If we accept this position then the immediate research question is *how* do (can) we interrogate the ML to determine the actual OD? We then pose two further research questions. First, how can we use the ML's representation of its OD to identify appropriate monitoring and mitigation strategies? Second, how do we ensure that the process is agile enough to support rapid change cycles common for ML?

A. Interrogation

There is a substantial amount of research in querying ML resilience, out of distribution detectors and so on. We propose to build on this body of work to interrogate the ML and produce an OD description which is representative of how the ML really ”understands” the world. This can be used to support safety analysis where it can be determined if there is a system level mitigation that can be used to counter gaps/misunderstanding, and/or whether more data/training can improve the situation. This is an area where we believe there is a key difference between our proposal and current practice - we would only propose additional training cycles if the gap poses a severe enough risk and it cannot be mitigated in a more effective way.

A substantial challenge is that the approach will require some translation between the ML's internal model to an approximation of the real-world useful for safety analysis.

B. Monitoring the real OD

Assuming we have a better definition of the ML's actual OD, we may need to monitor for situations which are outside of it. Again, this is expected to be challenging depending on the nature of the gap and the need to translate data from real-world sensors. There will be additional temporal challenges to consider, such as whether a single transient or even relatively frequent drop from the OD may be tolerable. It would also be desirable to make this monitoring continue the interrogation process and identify further gaps. This is not too far from existing practice, but again we emphasise the need to monitor for the real OD, and not just assume that situations that are difficult for humans (such as fog or rain) will necessarily pose the same challenge for the ML, e.g., in image classification.

C. Supporting rapid change

One final challenge we cannot ignore is the nature of the ML design process and frequent training cycles. Even if we reduce the amount of retraining, when we make an update due to an intolerable gap in the OD, if we discover a substantial change in the ML's OD then our system safety approach will also need updating, including changes to the design. Ideally, we must constrain the impact on the real OD as much as possible to limit safety assurance impact [8].

III. Conclusions

In this paper we have proposed that we reverse engineer the MLs understanding of the world to develop a more realistic, and less ”human-centred” approach to defining and refining an OD. At present we are unsure how feasible such an approach would be, but we believe the potential benefits could be substantial and research in this area could improve current practices to develop and assure AI-based safety critical-AS.

Acknowledgments

The authors would like to thank the Centre for Assuring Autonomy, a partnership between Lloyd's Register Foundation and the University of York for their support.

References

[1] S. C. S. C. Assurance Case Working Group, “GSN Community Standard Version 3,” 2021.

[2] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, “Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective,” *Artificial Intelligence*, vol. 279, p. 103201, 2020.

[3] EASA, “EASA Artificial Intelligence Concept Paper Issue 2,” 2024.

[4] SAE, “SAE J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles,” 2021.

[5] J. A. McDermid, R. Calinescu, I. Habli, R. Hawkins, Y. Jia, J. Molloy, M. Osborne, C. Paterson, Z. Porter, and P. Ryan Conmy, “The safety of autonomy: A systematic approach,” *Computer*, vol. 57, no. 4, pp. 16–25, 2024.

[6] Misc., “https://github.com/ultralytics/yolov5,” accessed June 2024.

[7] National Transportation Safety Board, “Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018, NTSB/HAR-19/03,” 2019. [Online]. Available: https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf

[8] C. Picardi, R. D. Hawkins, C. Paterson, and I. Habli, “Transfer assurance for machine learning in autonomous systems,” *Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2023)*, February 2023, © 2023 The Authors. [Online]. Available: https://eprints.whiterose.ac.uk/196682/