ARE DIGITAL TWINS SUITABLE TO DRIVE SAFE AI?

Alberto Carlevaro, Giacomo De Bernardi, Marta Lenatti, Sara Narteni, Marco Muselli, Alessia Paglialonga, Fabrizio Dabbene, *and* Maurizio Mongelli
CNR-IEIIT
email: maurizio.mongelli@cnr.it

The document deals with V&V of AI for autonomous vehicles that need to move and perform tasks in crowded environments. The adoption and extension of digital twin technology is discussed to pave the way towards AI certification.

Keywords: *smart mobility, autonomy, V&V, safe AI, SOTIF*

1. Introduction

The document is based upon and motivated by an ongoing H-EU project (REXASI-PRO, "REliable & eXplAinable Swarm Intelligence for People with Reduced mObility"[1]), in which AI is certified for autonomous wheelchairs for elderly and fragile people in indoor environments, such as stations (railway, airport) or museums (Fig. 1). Inherent applications involve complex interactions between humans and robots, such as: robots carrying material or autonomous stretchers in hospitals, material handling tasks in logistics, smart manufacturing, robot-aided physiotherapy or post trauma recovery. The rationale is to study how to develop a digital twin to expand the capacity of trials of the ecosystems like REXASI-PRO. According to traditional VV, the aim is to enumerate all safety issues, hazards and countermeasures, paving the ground for broader verification and certification applications (e.g., preventing collision of autonomous wheelchair through additional cameras and drones). The problem is however even more subtle; the certification needs the re-design of traditional VV when AI is in the loop.

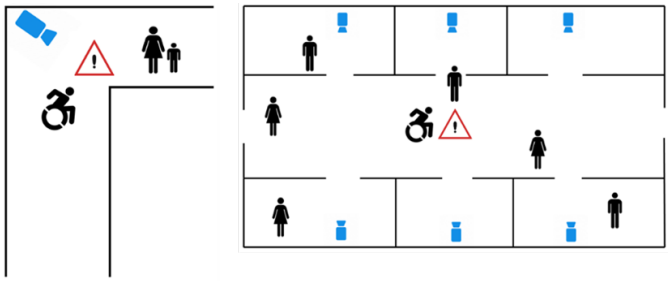

Figure 1: REXASI-PRO scope: autonomous wheelchair should move in the crowd fast and safely.

2. Why safe AI

The recent success of AI over traditional model-based methods is mainly due to its flexibility ("train, plug and play"). The wheelchair moves according to a neural control, trained through a collection of field data (Fig. 2). The database for training is based upon several registrations of passages of the wheelchair, guided by an operator, within a sufficient series of crowd states. The operator registers the most relevant passages with respect to safe and comfortable distance maintenance, as well as acceptable speed to reach the desired destinations.

[1] https://rexasi-pro.spindoxlabs.com/

The power of the approach is evident: just data recording and mapping the inherent movements into a neural network. The mapping between comfort, speed and collision avoidance is still something to be discovered from the trained neural network. More than this, the presumed "sufficiency" of crowd states highlights the problem of uncertainty. Is the trained neural network compatible with any other possible crowd scenario and corresponding field data acquisition? Under which operating conditions could it give rise to unexpected control spikes that make the actuation dangerous (close to a collision)? This leads to the need for a digital twin that incorporates the actuation engine and simulation of field data. AI itself then becomes an integrated tool in the digital twin as explained later on.

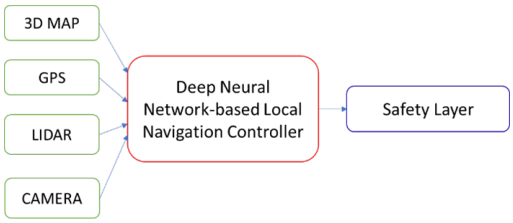

Figure 2: REXASI-PRO basic functional chain: sensing, neural controller, safety layer.

3. SOTIF

Although ISO 26262:2018 series of standards remains the foundation for providing safe hardware, safe software, and safe systems in the automotive industry, the high complexity and the real nature of AI creates a paradigm shift from safety assurance perspective, as safety failures may happen even in the absence of component failures. That is, hazards may result from the functional insufficiency and limitations of the technology in assuring the autonomous function, even in the absence of canonical hardware or software faults. This new safety challenge is presently referred to as safety of the intended functionality (SOTIF)[2], and is defined as the absence of unreasonable risk due to hazards resulting from functional insufficiency of the intended functionality or from misuse by users [1]. The final result is a static vision of possible dangerous situations (Fig. 1 is just one example).

4. Black swans

The problem is however far from being solved through a clear methodology and subsequent guidelines. As stated by the standard, the most critical challenge is to understand the so-called "black swans", which means discovering hazards disregarded by the static risk analysis. Figure 3 helps understand. The risk associated with the introduction of autonomous functions is linked not only to the probability of incorrect execution of the safety-related functions (areas 1 and 4), but to the intrinsic uncertainty in system performance, in predicted (area 2) and unpredicted operating conditions (area 3). Methodologies for the systematic search of black swans constitute an unexplored area of research and this is where the digital twin comes into play.

Categorization of real-life driving scenarios	Known	Unknown
Safe	Area 1 Nominal behavior	Area 4 System robustness
Potentially hazardous	Area 2 Identified system limitations	Area 3 "Black swans"

Figure 3: SOTIF categorization of scenarios.

[2]https://www.iso.org/standard/70939.html

5. SOTIF and digital twin

The scheme in Fig. 4 helps summarize all the steps involved in the certification process. Static hazards are translated into specific requests to the simulation engine in order to investigate the realization of dangerous and critical events (e.g., collision/comfort of trajectory) under the simulated fragment of the operative design domain (ODD, e.g., no cameras, corridor corners, high speed and large crowd density). Finding the desired classes (safety versus criticalities) with respect to the dynamicity of the system (e.g., speed versus crowd density) is the final goal to pursue. The rationale of this investigation relies on the fact that static hazards may just simply suggest the categories of simulation runs, but cannot anticipate the exact mapping of the parameters leading to safety versus criticalities. It is rather necessary to perform several runs around the hypothesis made by SOTIF and extract knowledge in an automatic way. The approach follows explainable and reliable machine learning[3] [2] in order to improve the coverage of the scenarios considered, thus avoiding brute force simulation. The core of the methods is XAI as it may drive iteration with the experts in the field who, in turn, can understand AI reasoning and merge natural and artificial intelligence [3]. As to further details on functional architecture, risk analysis and preliminary results, the interested reader is invited to take a look at the extended version of the document available in the link below[4].

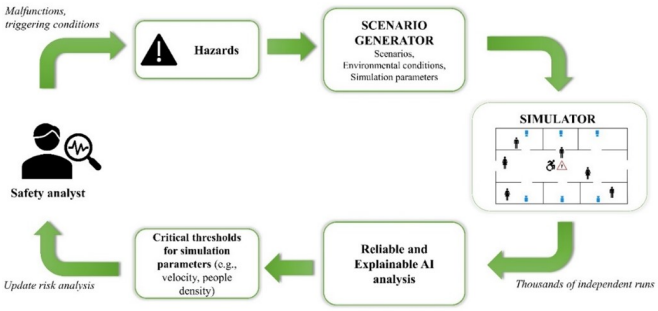

Figure 4: SOTIF and digital twin, empowered by AI.

ACKNOWLEDGEMENT

This work was supported in part by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, Grant agreement ID: 101070028.

REFERENCES

1. Kaiser, B. An integrative solution towards sotif and av safety, *IQPC SOTIF Conference*, (2019).

2. Narteni, S., Carlevaro, A., Dabbene, F., Muselli, M. and Mongelli, M. Confiderai: Conformal interpretable-by-design score function for explainable and reliable artificial intelligence, *Conformal and Probabilistic Prediction with Applications*, pp. 485–487, PMLR, (2023).

3. Lenatti, M., Carlevaro, A., Guergachi, A., Keshavjee, K., Mongelli, M. and Paglialonga, A. A novel method to derive personalized minimum viable recommendations for type 2 diabetes prevention based on counterfactual explanations, *Plos one*, **17** (11), e0272825, (2022).

[3] eXplainable AI (XAI): a machine learning (ML) model that is understandable by humans (e.g., expressed by rules of the if-then-else type). It lies into the transparency requirement of trustworthiness. Reliable AI (RAI): a ML model that it robust (often said resilient) to oscillations and attacks to its inputs in operation and keeps the model error under control.

[4] https://tinyurl.com/3nztxd7v