Probabilistic Modelling for Design and Verification of Trustworthy Autonomous Systems

Franca Corradini

IDSIA USI-SUPSI, Lugano, Switzerland

Abstract

Thanks to the recent technological achievements in Artificial Intelligence (AI) and robotics, autonomous systems have been improved to perform increasingly complex tasks in open and uncontrolled environments. Given the uncertainties embedded in machine learning systems and in the environment where they operate, traditional model-based evaluation techniques are not applicable in the design and verification stages. This poses several challenges especially in terms of safety assessment. In the course of the doctoral studies, probabilistic modelling approaches will be investigated to cope with uncertainties and to ensure measurable trustworthiness in autonomous systems. The reference application for the development of the design methodology and the experimental proof-of-concept is the ones of drone-supported autonomous wheelchairs, with a focus on the smart-sensing subsystems. Such application will be developed within a European funded project named REXASI-PRO, where an innovative solution for enhancing mobility and independence for people with disabilities is provided. Deployment of those systems in real-world scenarios imposes strict safety requirements. Probabilistic models can be used to capture uncertainties and variations in the environment and sensory system, enabling the system to change and adapt accordingly. The REXASI-PRO project will address the modelling methodology, tools, reference architecture, design and implementation guidelines. The PhD research will follow project objectives and milestones, including demonstration in relevant indoor and outdoor navigation scenarios. More specifically, a methodology based on Bayesian Network models will be developed and demonstrated to achieve measurable trust and pave the way to quantitative safety assessment of autonomous systems.

Keywords

Trustworthy AI, Autonomous systems, Probabilistic modelling

1. Context and motivation

In the last decades, autonomous systems have seen a growing development in several fields, including automotive [1], navigation, aerospace, industry [2], and military [3] applications. In many cases, those systems are aimed at carrying out operations that were impossible or critical to perform for human workers. Autonomous systems have been applied mostly in environments where uncertain events and disturbances are either absent or largely limited, or where there is supervision by human operators to some extent. Thanks to the recent technological achievements in AI and robotics, autonomous systems have been improved to perform increasingly complex tasks such as driving vehicles in complex, open and uncontrolled

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

franca.corradini@idsia.ch (F. Corradini)

0009-0009-5867-9478 (F. Corradini)

environments, even without human supervision. However, due to the possible criticality of those applications, new vital requirements have been introduced to set next research challenges. A new vocabulary has been recently introduced to address all the necessary aspects in the design and evaluation of those systems, not only from a technical perspective, but also in terms of ethical and legal implications, including fairness and accountability. The "Ethics Guidelines for Trustworthy Artificial Intelligence" [4], presented by the High-Level Expert Group on AI set up by the European Commission, states that trustworthy AI should be:

i. Lawful, to ensure that all laws and regulations are applied and respected;
ii. Ethical, to adhere to moral principles and values;
iii. Robust, to avoid any unintended damage and safety issues.

This represents a new challenge for the *Validation and Verification* ($V\&V$) of Intelligent Systems. The uncertain nature of these systems, related to their ability to adapt in response of external or internal disturbances as well to their capacity of taking choice in autonomy, limit the use of traditional evaluation techniques during $V\&V$ process. Furthermore when machine learning techniques are included in the autonomous system, better performances are in general achieved by increasing machine learning complexity at the expense of explainability [5]. In fact, most deep learning models are considered as "black-boxes" compared to traditional control algorithms and models. Therefore, it is very difficult to use traditional $V\&V$ methods, rather novel and diverse methodologies should be adopted [6]. Considering the complexity characterising these systems, it is to be expected that multiple and different verification techniques may be necessary at different stages of the $V\&V$ process [7].

Several are the initiatives for new collaborative research activity to improve the trustworthiness of autonomous systems with a focus on verifiability (denoting the quality or state of being capable of being verified, confirmed, or substantiated). The UKRI Trustworthy Autonomous Systems (TAS) Hub[1] is a coordination, community-building, and engagement hub that carry on a program interlinked projects addressing issues related to TAS. Between the several projects, the "Verifiability Node"[2] aims to carry out foundational research to enable the possibility of having a verified autonomy store. This is realised providing an heterogeneous collection of verification approaches, together with the semantic foundations to design and justify combinations of these heterogeneous concepts and techniques and analysing the verification issues that emerge across autonomous systems during their application [7].

From the recent systematic review on Testing, Validation, and Verification of Robotic and Autonomous Systems of reference [8], partially supported by the UKRI TAS, arises the problem of a gap in quantitative modelling languages that can capture the complex and heterogeneous nature of robotic and autonomous systems. This is an interesting research opportunity and an important topic for future developments for autonomous systems verifiability.

All the aforementioned aspects delineate the context of the doctoral studies subject of the current proposal and are extremely important when addressing real-world adoption of novel technologies leveraging on AI and machine learning, whose failure can have severe consequences on human health. This is the case of autonomous wheelchairs that are meant to support motion-

[1] https://www.tas.ac.uk
[2] https://verifiability.org

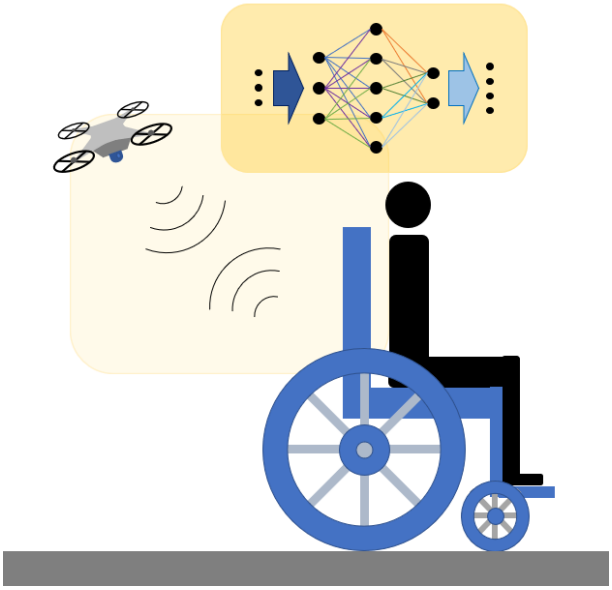

Figure 1: Safety-monitored wheelchair-drone system.

impaired persons in safe door-to-door navigation. We will address this case-study in next section.

2. The REXASI-PRO Project

The work of the doctoral studies subject of the current proposal will be carried out in the context of the recently started European project named REXASI-PRO[3] (*Reliable and Explainable Swarm Intelligence for People with Reduced Mobility*). This project has the challenging objective of implementing a trustworthy swarm intelligence based on the cooperation of autonomous wheelchairs and drones. The aim is to improve the independent and safe mobility for wheelchair-bound persons. A schematic illustration of the system is depicted in Figure 1. The idea of the project is to develop a novel framework in which security, safety, ethics, and explainability are entangled to create a trustworthy collaboration among wheelchairs and flying robots to allow a seamless door-to-door experience for people with reduced mobility.

Among the several topics discussed in the project, the work of the doctoral studies of the current proposal focuses on the realisation of a trustable environmental sensing. The use of a reliable sensors system is an essential aspect in the context of social robotic navigation, which is crucial to guarantee robustness against uncertainties, internal malfunctions and external disturbances. For the case-study, a smart-sensing subsystem will be designed to provide trusted event detection by following a model-based approach where trustworthiness is enforced during the whole system life-cycle. Common causes of failures are reduced by applying the principle of "no single point of failure" and by using strategies that rely on technology diversity. To prove

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

the trustworthiness of the system, a model-based evaluation procedure will be used, in which verification for the sensing subsystem is performed at both design-time and run-time with the aim to fulfil requirements related to *Safety Integrity Levels* (SIL).

3. Research questions, hypothesis and objectives

The main issue that the doctoral studies of the current proposal has to address is the investigation and implementation of methods and models for TAS based on the design and run-time evaluation of trustable sensing through quantitative probabilistic methods that can be applied to assess system safety in presence of machine learning and environmental uncertainties.

Specifically, the main research questions (RQs) to address can be formulated as follows:

RQ1: "How can we use probabilistic approaches to support the design of safe autonomous systems by taking into account all the relevant uncertainties?"

RQ2: "How can we use probabilistic models for design-time and run-time verification in order to quantitatively evaluate system safety by using available knowledge about current system status, machine learning performance, and environmental conditions?"

RQ3: "Which strategies can be applied to the autonomous system to achieve an adequate level of safety at design-time and run-time and to ensure its reliability?"

Although the nature of these questions is more general, a special focus will be set on smart-sensing subsystems within safety-critical TAS.

In order to address the challenges posed by those RQs, a multi-agent, multi-modal and self-adaptive sensing system is proposed to achieve trusted event detection, where sensors outputs are combined to give a common result for the measured variables. In the case of event detection, one possible approach is based on *voting*, where the presence of a certain target is determined by the majority of detectors whose outputs match. Furthermore, by analysing and tracking detectors' performance over time, it is possible to score their reputation and exclude from voting those who are no more reputable and that could negatively affect the outcome of the decisions. With a change in its internal state, the sensing system can consider a subsystem of the initial set of detectors to maintain a certain level of reputation. In this way, the system is able to self-adapt when internal faults occur or when exogenous environmental events cause performance degradation.

Considering the multi-agent structure of the system, sensors characterised by different technologies and therefore affected by different types of internal or external faults are used. The assumption about the *diversity* of sensing technology/mechanism is essential to exclude correlations between them and common-mode faults. Indeed, the multi-sensor and multi-modal approach implies enough *redundancy* to evaluate the information we are interested in. It allows to increase system robustness against the malfunction of some of its components and, to some extent, to reduce the costs by using cheaper components. Moreover, technology redundancy and diversity is a necessary feature to improve resilience against environmental disturbances. As highlighted in reference [9] "*Diversity* should be taken advantage of in order to prevent vulnerabilities to become single points of failure".

The properties of the described system, such as self-adaptation, allow to deal with dynamic environmental uncertainties. Since the system changes over time, it is not acceptable to apply

traditional validation methods, which involved a single validation step at the end of the system design. One possibility to cope with the uncertain nature of our system is to adopt probabilistic approaches based on graphical models. *Bayesian networks* (BNs) [10] can be used due to their suitability to represent complex causal relationships between system components, and to visually describe interdependencies in an easily interpretable way. BNs extensions such as *Dynamic Bayesian Networks* [11] and *Time-Varying Dynamic Bayesian Networks* [12] are also useful to manage time-varying and dynamic aspects [13].

The BN approach can be linked to the *voting* approach, as described in references [14] and [13]. In the former, BNs are used to evaluate the effect of a *"k-out-of-m"*, voting approach on the performance of different sensor clusters chosen from a group of five sensors with different technologies. Dependencies among technologies are also discussed, showing how they worsen the results. In reference [13], the same concept is used for a self-adaptive system: a case-study in the domain of vehicle detection is used to demonstrate the approach, based on sensor detection performance measured in a previous study.

Based on the described approaches, we will leverage on the state-of-the-art in multi-modal sensing, and we will employ inherently explainable probabilistic methods based on BN models to dynamically evaluate sensing trustworthiness at run-time. To that aim, we will keep alive design-time models, and explore paradigms such as digital twins and autonomic computing, e.g., Monitor-Analyse-Plan-Execute over a shared Knowledge. The final objective will be to address safety integrity requirements and to set up appropriate model templates for the static and dynamic verification of critical subsystems within TAS. The complexity of the threat detection use cases in cooperative navigation scenarios that are included in REXASI-PRO will allow to set up appropriate proof-of-concepts to develop and benchmark novel techniques for probabilistic SIL evaluation within TAS.

4. Research approach, methods, and rationale for testing the research hypothesis

The activities of the doctoral studies presented in this proposal are planned as follows:

i. PhD objectives, research questions and draft activity plan are defined in line with research project objectives and timeline, but at more general and cross-domain level.
ii. Preliminary study on relevant methodologies and tools for machine learning, autonomic computing, digital twins, and probabilistic safety analysis through Bayesian Networks and their extensions is performed to build the necessary background knowledge and modelling skills.
iii. Systematic Literature Review (SLR) is performed on quantitative methods for the design and safety analysis of machine learning systems by using a sound SLR methodology and reputable sources.
iv. Theoretical definitions, sensor characterisation, and system/environment uncertainty classification underlying the methods and models discussed in Section 3 are provided.
v. Reference methodology development, REXASI-PRO architecture integration, and model implementation are performed with the aid of existing libraries and templates of BN and other Probabilistic Graphical Models as discussed in activity ii

vi. Tests with synthetic data and performance evaluation are carried out in order to validate the approach by using appropriate real-world simulators and reference benchmarks.
vii. Methodology and models are applied to real-world use-case scenarios from REXASI-PRO project (autonomous wheel chairs and drones), and thus tested and demonstrated through a proof-of-concept in laboratory environment at TRL (Technology Readiness Level) 3-4, by using project indoor and outdoor navigation and sensing data-sets.

5. Results and contributions to date;

The first thre activities mentioned in the previous section have recently started and are thus ongoing.

An extended abstract with the author of the current proposal as first author has been recently accepted as a poster contribution to the First International Symposium on Trustworthy Autonomous Systems (TAS'23). In the work, the model described in Section 3 is presented with a focus on trustable sensing systems within TAS.

In addition, a SLR on quantitative $V\&V$ methods for machine learning systems is ongoing as a result of the first phases mentioned in Section 4.

6. Expected next steps and final contribution to knowledge.

Two reports are planned as part of the REXASI-PRO project:

i. The first one (deadline May 2024) has a focus on design methodology for trustable sensing
ii. The second one (deadline December 2024) focuses on verification of trustable sensing.

Two papers on trustable sensing are expected to be published based on those reports.
At the end of the project and of the PhD, the following results will be achieved:

i. An experimental proof-of-concept (TRL 3), where the models are tested and validated in a simulated environment;
ii. A lab-validated technology (TRL 4), where tests and validation are carried out in real-world use-cases and scenarios.

References

[1] R. Bishop, A survey of intelligent vehicle applications worldwide, in: Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No.00TH8511), 2000, pp. 25–30. doi:10.1109/IVS.2000.898313.

[2] M. Müller, T. Müller, B. A. Talkhestani, P. Marks, N. Jazdi, M. Weyrich, Industrial autonomous systems: a survey on definitions, characteristics and abilities, at - Automatisierungstechnik 69 (2021) 3–13. URL: https://doi.org/10.1515/auto-2020-0131. doi:doi:10.1515/auto-2020-0131.

[3] Q. Ha, L. Yen, C. Balaguer, Robotic autonomous systems for earthmoving in military applications, Automation in Construction 107 (2019) 102934. URL: https://www.sciencedirect.com/science/article/pii/S0926580518309932. doi:https://doi.org/10.1016/j.autcon.2019.102934.

[4] A. HLEG, Ethics guidelines for trustworthy artificial intelligence, High-Level Expert Group on Artificial Intelligence 8 (2019).

[5] A. B. Arrieta, N. Díaz-Rodríguez, J. del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, Information Fusion 58 (2020). URL: https://hal.science/hal-02381211. doi:10.1016/j.inffus.2019.12.012.

[6] C. Ebert, M. Weyrich, Validation of autonomous systems, IEEE Software 36 (2019) 15–23. doi:10.1109/MS.2019.2921037.

[7] M. R. Mousavi, A. Cavalcanti, M. Fisher, L. Dennis, R. Hierons, B. Kaddouh, E. L.-C. Law, R. Richardson, J. O. Ringer, I. Tyukin, J. Woodcock, Trustworthy autonomous systems through verifiability, Computer 56 (2023) 40–47. doi:10.1109/MC.2022.3192206.

[8] H. Araujo, M. R. Mousavi, M. Varshosaz, Testing, validation, and verification of robotic and autonomous systems: A systematic review, ACM Trans. Softw. Eng. Methodol. 32 (2023). URL: https://doi.org/10.1145/3542945. doi:10.1145/3542945.

[9] J.-C. Laprie, From dependability to resilience, in: 38th IEEE/IFIP Int. Conf. On dependable systems and networks, 2008, pp. G8–G9.

[10] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, Adaptive computation and machine learning, MIT Press, 2009. URL: https://books.google.co.in/books?id=7dzpHCHzNQ4C.

[11] K. P. Murphy, Dynamic bayesian networks: Representation, inference and learning, dissertation, PhD thesis, UC Berkley, Dept. Comp. Sci (2002).

[12] Z. Wang, E. E. Kuruoğlu, X. Yang, Y. Xu, T. S. Huang, Time varying dynamic bayesian network for nonstationary events modeling and online inference, IEEE Transactions on Signal Processing 59 (2011) 1553 – 1568. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-79952665170&doi=10.11092fTSP.2010.2103071&partnerID=40&md5=cda0ead67147dfc635fbfd7bebe49153. doi:10.1109/TSP.2010.2103071, cited by: 42; All Open Access, Green Open Access.

[13] F. Flammini, S. Marrone, R. Nardone, M. Caporuscio, M. D'Angelo, Safety integrity through self-adaptation for multi-sensor event detection: Methodology and case-study, Future Generation Computer Systems 112 (2020) 965–981. URL: https://

www.sciencedirect.com/science/article/pii/S0167739X19333734. doi:https://doi.org/10.1016/j.future.2020.06.036.

[14] F. Flammini, Model-based analysis of 'k out of m' correlation techniques for diverse redundant detectors, International Journal of Performability Engineering 9 (2013) 551. URL: http://www.ijpe-online.com/EN/abstract/article_2887.shtml. doi:10.23940/ijpe.13.5.p551.mag.