Probabilistic Modelling for Trustworthy Artificial Intelligence in Drone-Supported Autonomous Wheelchairs

Franca Corradini
IDSIA USI-SUPSI
Lugano, Switzerland
franca.corradini@idsia.ch

Francesco Flammini
IDSIA USI-SUPSI
Lugano, Switzerland
francesco.flammini@supsi.ch

Alessandro Antonucci
IDSIA USI-SUPSI
Lugano, Switzerland
alessandro.antonucci@idsia.ch

ABSTRACT

In this work, we address the potential of probabilistic modelling approaches for ensuring trustworthy AI in sensing subsystems within drone-supported autonomous wheelchairs. The combination of drones and autonomous wheelchairs provides an innovative solution for enhancing mobility and independence of motion-impaired people. However, safety is a critical concern when deploying such systems in real-world scenarios. To address this challenge, probabilistic models can be used to capture uncertainty and non-stationarity in the environment and sensory system, enabling the device to make informed decisions while ensuring safe autonomy. The approach is being developed in the context of a recently started European project named REXASI-PRO, which addresses the modelling methodology, tools, reference architecture, design and implementation guidelines. In the project, relevant indoor and outdoor navigation use cases are addressed to demonstrate the effectiveness of the proposed approach in providing trustworthy autonomous wheelchairs in real-world environments.

CCS CONCEPTS

• **Computing methodologies** → **Model verification and validation**; *Modeling methodologies*; Uncertainty quantification.

KEYWORDS

Trustworthy AI, Autonomous Wheelchairs, Probabilistic Modelling.

ACM Reference Format:
Franca Corradini, Francesco Flammini, and Alessandro Antonucci. 2023. Probabilistic Modelling for Trustworthy Artificial Intelligence in Drone-Supported Autonomous Wheelchairs. In *First International Symposium on Trustworthy Autonomous Systems (TAS '23), July 11–12, 2023, Edinburgh, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3597512.3599716

1 INTRODUCTION

In the last decades, autonomous systems have seen a growing development in several areas, including automotive [6], navigation, aerospace, industry [18], and military [11] applications. In many cases, those systems are aimed at carrying out operations that are impossible or critical to perform for human workers. Currently, autonomous systems are mostly applied in environments where uncertain events and disturbances are either absent or largely limited, and they are supervised to some extent by human operators.

Thanks to the recent technological achievements in AI and robotics, autonomous systems have been improved to perform increasingly complicated tasks such as driving vehicles in complex, open and uncontrolled 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" [1], 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.

Trustworthy AI is fundamental for the notion of *Trustworthy Autonomous Systems* (TAS). TAS must be technically robust and therefore they must be evaluated according to the attributes and means of dependable, secure and resilient computing, as defined in the seminal papers [3] and [16]. More recently, in reference [9], the notions of dependability, resilience, and cyber-security have been connected as core concepts within a comprehensive TAS taxonomy.

TAS must be robust to uncertainties in the surrounding environment, secure against threats coming from cyber-space, and capable of safe human-machine interactions [12].

As addressed in reference [17], trustworthiness must be put in relation not only to purely technical aspects, but also to the benefits and wellness of people as a consequence of AS actions.

In general, better performance is achieved by increasing machine learning complexity at the expense of explainability (XAI) [2], which refers to the possibility of explaining the internal behaviour of intelligent systems. 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 evaluation techniques, rather novel and diverse methodologies should be instead adopted [17].

All the aforementioned aspects 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* (AWs) that are meant to support motion-impaired persons in safe door-to-door navigation.

TAS '23, July 11–12, 2023, Edinburgh, United Kingdom

ACM ISBN 979-8-4007-0734-6/23/07.
https://doi.org/10.1145/3597512.3599716

From the invention of the first powered wheelchairs in 1956, there have been several efforts to improve their technology. Modern wheelchairs can be very complex and include sophisticated components, such as navigation systems and vocal human-machine interfaces. The realisation of AWs aims to accommodate several disabilities by means of multi-modal interfaces. One aspect to be carefully addressed is the safety assessment and possible certification of AWs. In fact, in presence of AI, together with specific regulations such as "ISO 7176-14:2022 Wheelchairs"[1], other requirements and guidelines should be considered, such as the ones included in the EU Artificial Intelligence Act[2]. In case AWs also monitor biomedical parameters, further certifications might be required.

2 THE REXASI-PRO PROJECT

The recently started European project named *Reliable and Explainable Swarm Intelligence for People with Reduced Mobility* (REXASI-PRO)[3] set challenging objectives to implement trustworthy swarm intelligence based on the cooperation of AWs and drones. 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 AWs and flying robots to allow a seamless door-to-door experience for people with reduced mobility. Among project objectives, three main challenges are identified:

(i) the usage of AI techniques in different use-cases for AWs;
(ii) the development of autonomous flying drones that are able to replace the need for human supervision within indoor and outdoor environments;
(iii) the implementation of mixed collaborative environments, where devices can communicate with each other to manage emergencies.

AI is used to develop safety assistant, driving assistant, route assistant, and social navigation modules to allow AWs to adapt to unknown environments and situations. To overcome the limitations of ground vehicles, drones can map the surroundings from any perspectives that allows to see obstacles and elaborate their representation in a 3D map. The generated map is then used to build a simulation environment where the AW dynamics is recreated. This allows to ensure trusted social navigation and collisions avoidance, to preserve the safety and well-being of wheelchair-bound persons.

As for any complex systems, AWs are divided in simpler components following a modular approach. Therefore, holistic safety assessment can also be based on a modular approach, starting from single components, such as the sensors, and moving up to subsystems, such as the environmental sensing one, and considering the interfaces between them. In the context of social robotic navigation, trustable environmental sensing is an essential aspect that is crucial to guarantee robustness against uncertainties, internal malfunctions, and external disturbances. Sensor systems include smart sensors elaborating raw data coming from environmental measurements and transforming it into useful information such as event/threat detection.

[1] https://www.iso.org/standard/72408.html
[2] https://artificialintelligenceact.eu
[3] https://rexasi-pro.spindoxlabs.com

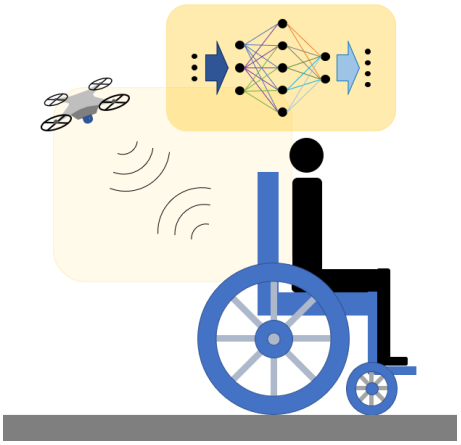

Figure 1: Safety-monitored wheelchair-drone system.

For the specific AW case-study, a smart-sensing subsystem is used 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 mitigated by applying the principle of "no single point of failure" together with strategies that rely on technology diversity. To assess the trustworthiness of the system, a model-based evaluation is 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 PROBABILISTIC SAFETY MODELLING

In order to address the challenges related to quantitative safety assessment, we propose a multi-agent, multi-modal and self-adaptive sensing system 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, a possible approach is based on *voting*, where the output is based on the agreement of most detectors. By analysing and tracking outputs of sensors and their detection performance over time, it is possible to score their reputation, weight their contribution accordingly, and even exclude those that are no more considered *reputable*, i.e., those that could then negatively affect the outcome of the decisions. Through appropriate reconfiguration, the sensing system can consider a subset of detectors to keep the required safety level. In other words, the sensing system is able to self-adapt when internal faults occur or when an exogenous environmental condition affect detection performance.

The ability of self-adaptation is achieved by combining a "Managed Subsystem", which is the system under consideration (i.e., the sensing one), with a "Managing Subsystem" based on the *Monitor, Analyse, Plan and Execute* over a shared *Knowledge* (MAPE-K) feedback loop (see Figure 2). In AWs, the managing part implements the autonomic safety logic over the monitored sensing subsystem within the overall wheelchair-drone system. The managed subsystem is monitored along with the environment, and related data is stored in the knowledge base, which includes all relevant models representing the system from the safety perspective. Data is

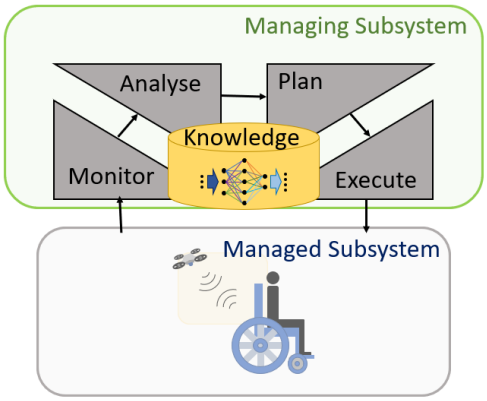

Figure 2: Self-adaptation through MAPE-K feedback loop for safety monitoring of wheelchair-drone system.

continuously collected and analysed to check if pre-defined safety conditions are met and no anomalies are detected. If needed, reconfiguration actions are planned at run-time to keep the system operation safe enough, including exclusion of faulty sensors, issue of alerts, application of speed limitations, switch to manual/remote control, or even fail-safe system stop.

Considering the multi-agent structure of the system, sensors characterised by different technologies are used, affected by different types of internal or external faults. The assumption about the *diversity* of sensing technology/mechanism is essential to exclude correlations between them and common-mode faults. Sensor diversity can be also achieved through different algorithms, parameters choice, and sensor displacement [8].

This 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 [16], "*Diversity* should be taken advantage of in order to prevent vulnerabilities to become single points of failure". More recently, the importance of diversity in machine learning systems has been highlighted to comply with functional safety requirements [5].

TAS operating in real-world environments must cope with several uncertainties such as unpredicted changes, disturbances, and the so-called "unknown unknowns". Properties such as self-adaptation allow to deal with those uncertainties, however they also limit the possibility of employing deterministic verification approaches. One possibility to cope with uncertainties is to adopt probabilistic approaches possibly based on graphical models. *Bayesian networks* (BNs) [15] can be used due to their suitability to represent complex causal relationships between system components, and to visually describe inter-dependencies in an easily interpretable way. The assumption about diversity of sensing technology is essential to exclude correlations between sensors and common-mode faults. BN extensions such as *Dynamic Bayesian Networks* (DBNs) [19] and *Time-Varying Dynamic Bayesian Networks* (TV-DBNs) [20] are also useful to manage time-varying and dynamic aspects of the system [10].

In the specific case-study of vehicle detection, the presence or absence of an automobile on a specific section of the road is detected through sensors using different technologies, e.g. magnetometer sensor, video image processor or radar sensors. Unforeseen elements could interfere with the analysed scene and, depending on the their nature, induce errors on one or multiple sensors.

In the described context, let X denote the actual, non-observable, value to be measured, i.e. the random variable representing the event "vehicle presence or absence". As we assume the sensors to be only partially trustable, the observation of X as returned by a sensor S_i is described by a distinct, observable, variable $O_X^{S_i}$ with the same possible values of X. Let us denote as E_i the exogenous factors possibly inducing a deterioration of sensor trustworthiness. In the case of a video image processor, it could be caused by weather condition that worsen visibility, as for instance fog or rain. We model such a correlation by setting X and E_i as *parents* of the observable variable $O_X^{S_i}$ in the BN. The quantification of BN parameters requires the quantification of conditional probabilities in $P(O_X^{S_i}|X, E_i)$. This corresponds to a confusion matrix to be assessed for each configuration of the exogenous factors in E_i. Assuming no other variables are involved in the measurement process, we have X and E_i representing root nodes of the BN graph (see an example in Figure 3). Note that to complete the BN quantification, also the unconditional probabilities of X and E_i should be assessed. Overall, this defines a joint model of the form:

$$P(X, O_X^{S_1}, \ldots, O_X^{S_m}|E_i, \ldots, E_m) = P(X) \prod_{i=1}^{m} P(O_X^{S_i}|X, E_i), \quad (1)$$

where the actual values of the exogenous factors are assumed to be observed. If this is not the case, a weighted average over the different exogenous configurations should be considered. Such a joint probabilistic model allows to infer information about the variable X in the form of the posterior probability $P(X|O_X^{S_1}, O_X^{S_2}, \ldots, O_X^{S_m})$.

Sensor diversity implies the conditional independence of the different sensor measurements given the latent variable as well as the unconditional independence between the exogenous factors affecting the reliability of the different sensors. For instance, weather conditions are mostly irrelevant to radar sensors, which are subject to multi-path propagation, separability, and sensitivity of radar cross section to the aspect angle [13]. Those assumptions allow to reduce the computation of the posterior probabilities to an inference in a *naive* topology and hence derive a closed-form expression. In practical applications, extensive testing for correlation between components should be considered to evaluate the impact on safety targets. If the above independence relations are not satisfied, BN inference algorithms should be used to obtain the posterior distributions. This appears as the necessary computational counterpart of a higher expressiveness in the modelling phase. Similar considerations hold for the extension of the above static setup to a dynamic one, where the *Markovian* assumption is typically considered to reduce the modelling to two consecutive time steps only.

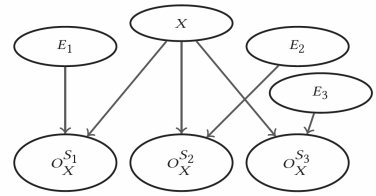

Figure 3: A BN modelling three sensors measuring X.

The BN approach can be linked to the *voting* approach, as described in references [8] and [10]. 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 [10], 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, in REXASI-PRO we leverage on the state-of-the-art in multi-modal sensing, and we employ inherently explainable probabilistic methods based on BN models to dynamically evaluate sensing trustworthiness at run-time. To that aim, we keep alive design-time models, and explore paradigms such as digital twins and autonomic computing, e.g., the MAPE-K feedback loop. The final objective is 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 allows to set up appropriate proof-of-concepts to develop and benchmark novel techniques for probabilistic SIL evaluation within TAS.

4 CONCLUSIONS

Assessing trustworthiness in AWs is essential to ensure their safety in real-world applications. Most powerful AI techniques suffer from opacity and explainability issues when employed in safety-critical applications [9]. The use of probabilistic approaches can support safety assessment by providing a quantitative evaluation of trustworthiness that is applicable to selected subsystems, such as the sensorial one [10]. With such an approach, the pro-active safety achievable with higher intelligence, adaptation and uncertainty management capabilities can be combined with the potential of run-time monitoring and probabilistic model checking enabled by appropriate modeling formalisms, such as Bayesian Networks and their extensions [14]. Together with other trustworthy AI techniques, such as XAI and safety envelopes [4], we believe that this approach can have a great potential in addressing real-world certification challenges of critical autonomous systems [7]. We are currently developing and testing the approach in industrially relevant use-cases within the recently started REXASI-PRO project, where several cross-discipline aspects related to robust, ethical, and legal AI are being investigated.

ACKNOWLEDGMENTS

This work was partly supported by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, grant agreement no. 101070028, funded by the European Union.

Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or of the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] HLEG AI. 2019. High-level expert group on artificial intelligence. , 6 pages. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[3] Algirdas Avizienis, Jean-Claude Laprie, Brian Randell, and Carl Landwehr. 2004. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing* 1, 1 (2004), 11–33. https://doi.org/10.1109/TDSC.2004.2

[4] Julian Bernhard, Patrick Hart, Amit Sahu, Christoph Schöller, and Michell Guzman Cancimance. 2022. Risk-Based Safety Envelopes for Autonomous Vehicles Under Perception Uncertainty. In *2022 IEEE Intelligent Vehicles Symposium (IV)*. 104–111. https://doi.org/10.1109/IV51971.2022.9827199

[5] Axel Brando, Isabel Serra, Enrico Mezzetti, Francisco J. Cazorla, Jon Perez-Cerrolaza, and Jaume Abella. 2023. On Neural Networks Redundancy and Diversity for Their Use in Safety-Critical Systems. *Computer* 56, 5 (may 2023), 41–50. https://doi.org/10.1109/MC.2023.3236523

[6] Long Chen, Yuchen Li, Chao Huang, Bai Li, Yang Xing, Daxin Tian, Li Li, Zhongxu Hu, Xiaoxiang Na, Zixuan Li, Siyu Teng, Chen Lv, Jinjun Wang, Dongpu Cao, Nanning Zheng, and Fei-Yue Wang. 2023. Milestones in Autonomous Driving and Intelligent Vehicles: Survey of Surveys. *IEEE Transactions on Intelligent Vehicles* 8, 2 (feb 2023), 1046–1056. https://doi.org/10.1109/tiv.2022.3223131

[7] M.L. Cummings and David Britton. 2020. Chapter 6 - Regulating safety-critical autonomous systems: past, present, and future perspectives. In *Living with Robots*, Richard Pak, Ewart J. de Visser, and Ericka Rovira (Eds.). Academic Press, 119–140. https://doi.org/10.1016/B978-0-12-815367-3.00006-2

[8] Francesco Flammini. 2013. Model-Based Analysis of 'k out of m' Correlation Techniques for Diverse Redundant Detectors. *International Journal of Performability Engineering* 9, 5 (2013). https://doi.org/10.23940/ijpe.13.5.p551.mag

[9] Francesco Flammini, Cristina Alcaraz, Emanuele Bellini, Stefano Marrone, Javier Lopez, and Andrea Bondavalli. 2022. Towards Trustworthy Autonomous Systems: Taxonomies and Future Perspectives. *IEEE Transactions on Emerging Topics in Computing* (2022), 1–13. https://doi.org/10.1109/TETC.2022.3227113

[10] Francesco Flammini, Stefano Marrone, Roberto Nardone, Mauro Caporuscio, and Mirko D'Angelo. 2020. Safety integrity through self-adaptation for multi-sensor event detection: Methodology and case-study. *Future Generation Computer Systems* 112 (2020), 965–981. https://doi.org/10.1016/j.future.2020.06.036

[11] Q.P. Ha, L. Yen, and C. Balaguer. 2019. Robotic autonomous systems for earthmoving in military applications. *Automation in Construction* 107 (2019), 102934. https://doi.org/10.1016/j.autcon.2019.102934

[12] Hongmei He, John Gray, Angelo Cangelosi, Qinggang Meng, T. M. McGinnity, and Jorn Mehnen. 2020. The Challenges and Opportunities of Artificial Intelligence for Trustworthy Robots and Autonomous Systems. In *2020 3rd International Conference on Intelligent Robotic and Control Engineering (IRCE)*. 68–74. https://doi.org/10.1109/IRCE50905.2020.9199244

[13] Martin Holder, Philipp Rosenberger, Hermann Winner, Thomas D'hondt, Vamsi Prakash Makkapati, Michael Maier, Helmut Schreiber, Zoltan Magosi, Zora Slavik, Oliver Bringmann, and Wolfgang Rosenstiel. 2018. Measurements revealing Challenges in Radar Sensor Modeling for Virtual Validation of Autonomous Driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 2616–2622. https://doi.org/10.1109/ITSC.2018.8569423

[14] Manfred Jaeger, Kim G. Larsen, and Alessandro Tibo. 2020. From Statistical Model Checking to Run-Time Monitoring Using a Bayesian Network Approach. In *Runtime Verification: 20th International Conference, RV 2020, Los Angeles, CA, USA, October 6–9, 2020, Proceedings 20*. Springer International Publishing, 517–535.

[15] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

[16] Jean-Claude Laprie. 2008. From dependability to resilience. In *38th IEEE/IFIP Int. Conf. On dependable systems and networks*. G8–G9.

[17] Mohammad Reza Mousavi, Ana Cavalcanti, Michael Fisher, Louise Dennis, Rob Hierons, Bilal Kaddouh, Effie Lai-Chong Law, Rob Richardson, Jan Oliver Ringer, Ivan Tyukin, and Jim Woodcock. 2023. Trustworthy Autonomous Systems Through Verifiability. *Computer* 56, 2 (2023), 40–47. https://doi.org/10.1109/MC.2022.3192206

[18] Manuel Müller, Timo Müller, Behrang Ashtari Talkhestani, Philipp Marks, Nasser Jazdi, and Michael Weyrich. 2021. Industrial autonomous systems: a survey on definitions, characteristics and abilities. *at - Automatisierungstechnik* 69, 1 (2021), 3–13. https://doi.org/10.1515/auto-2020-0131

[19] Pedro Shiguihara, Alneu De Andrade Lopes, and David Mauricio. 2021. Dynamic Bayesian Network Modeling, Learning, and Inference: A Survey. *IEEE Access* 9 (2021), 117639–117648. https://doi.org/10.1109/ACCESS.2021.3105520

[20] Zhaowen Wang, Ercan E Kuruoğlu, Xiaokang Yang, Yi Xu, and Thomas S Huang. 2011. Time varying dynamic Bayesian network for nonstationary events modeling and online inference. *IEEE Transactions on Signal Processing* 59, 4 (2011), 1553 – 1568. https://doi.org/10.1109/TSP.2010.2103071