Safeguarding autonomous systems: emerging approaches, assumptions and metrics – a systematic literature review

Manuel S. Müller*, Tobias Jung*, Nasser Jazdi*, Michael Weyrich*

**Institute of Industrial Automation and Software Engineering, University of Stuttgart, Stuttgart, DE70550 Germany (Tel: +49-711-685-67306; e-mail: manuel.mueller@ias.uni-stuttgart.de).*

Abstract: Autonomous systems gain more and more interest in research and society. However, they bring new challenges in safeguarding these systems. This contribution orders those new challenges and provides an overview of already existing concepts and approaches to solve those challenges of safeguarding autonomous systems. Moreover, existing metrics for safeguarding of autonomous systems are systematically reviewed. The presented concepts and approaches are of different domains, namely, ground, nautical and aerial vehicles, industrial robots and smart manufacturing, and medical and healthcare. Finally, the concepts and approaches are discussed concerning the following points: Main ideas, parallels existing in different domains, which ideas can be transferred from one domain to another, which high-level tasks were adressed and which assumptions were made.

Keywords: autonomous systems, safety, safety-critical systems, safety-constraint systems.

1 INTRODUCTION

Autonomous systems gain more and more interest in research and industry. First autonomous systems like robot vacuums are available and driven by billions of investments, great success is celebrated. This boom reflects in market capitalization (ark-funds.com, 2021) and numbers of publications from the academia perspective. However, as (Abbass et al., 2018) pointed out, there are proportionally few autonomous systems involved in industries today. One reason is the challenge of providing system safety despite unpredictability of open environments. Identified as bottleneck in the introduction of autonomous systems, the question of safety of autonomous systems moves to the focus: How can safe behavior of all these autonomous systems be guaranteed? Which emerging approaches do exist in the different domains? Which assumptions do they make? How do the approaches measure the safety, i.e. which metrics do they use? This systematic literature review tries to cluster the state of the art in the most relevant industrial domains according to (McKee et al., 2018). The goals are to identify common ideas of emerging approaches, find cross-domain synergies and make the safety measurable.

In order to reach these goals, we start with related work in Section 2. In Section 3, we specify our methodology and formally formulate the goals in concrete research questions (RQ). The analyzed publications are presented and assumptions are outlined in Section 4. The RQs shall be discussed in Section 5. The paper ends with a conclusion and an outlook (Section 6).

2 BACKGROUND

There are other surveys on properties related to safety like trust (Shahrdar et al., 2018), security modelling (Jahan et al., 2019) or the influence from security on safety (Koschuch et al., 2019). In addition, subareas of safety like active fault diagnostics (Punčochář and Škach, 2018) or fault tolerant control methods (Fritz and Zhang, 2018) are reviewed. Moreover, surveys on autonomous systems that scratch the topic of safety (McKee et al., 2018; Hayat et al., 2016) or surveys on aspects of autonomy like artificial intelligence that review the safety criterion are available (Tong et al., 2019; Juric et al., 2020). An overview on the most related existing literature reviews is provided in Table 1. However, to the best of our knowledge there is no systematic survey on safety concepts for autonomous systems of different domains right now.

Table 1. Table of related surveys

Survey	Focus	relation
(McKee et al., 2018)	Advances and challenges in autonomy	Safety challenges of autonomy
(Osborne et al., 2019)	Technologies for certification of unmanned aerial systems	Safety concepts for unmanned aerial systems
(Liu et al., 2013)	Risk evaluation in failure mode and effect analysis	Spotting safety risks, metrics
(Zhang and Li, 2020)	Test and verification of neural networks at design time	Improve reliability of machine learning
(Osborne et al., 2019)	Unmanned aerial vehicles	Review of safety criterion

Although there is some research on safety and autonomous systems, especially in the field of autonomous systems, taxonometry is not yet unified (Müller et al., 2021). For this reason, in this section the wording is recapitulated.

In this paper, we consider the following **definitions**: an autonomous system is “a delimited technical system, which systematically and without external intervention, achieves its set objectives despite uncertain environmental conditions” (Müller et al., 2021). It has four main characteristics: “(1) systematic process execution, (2) adaptability, (3) self-governance and (4) self-containedness” (Müller et al., 2021). According to (McKee et al., 2018), there are autonomous systems in the domains of vehicles, robots and plants, and medical and healthcare. Safeguarding is the process of providing safety. Safety is a subarea of dependability. Dependability and security make a system robust, i. e. by aoiding critical system states due to external factors like faults (Ratasich et al., 2019). Moreover, dependability contributes to resilience, i.e. the automatic recovery of a critical system state (Ratasich et al., 2019). Consequently, the main tasks of safeguarding are to avoid critical states and – if they occure – to detect them and to automatically recover from them.

Safety according to DIN/ISO 61508 deals with the systematic reduction of a risk to a tolerable residual risk. The risk considers the probability of a negative (side) effect of the system on the environment multiplied by the severity of the effect, the extent of damage. The complementing norm ISO/PAS 21448:2019 extends the component-centric view on the safety term by a functionality-oriented view.

3 METHODOLOGY AND GOALS

In order to get an overview on the topic, we started with a bibliographic analysis of the term “autonomy*” on WebOfScience using VosViewer. From this bibliometric analysis, the predominant role of vehicles in the area of autonomous systems gets obvious. They are divided into the field of vehicles operating on land, represented with synonyms like “road vehicle”, “car”, “road user”, “ground vehicle”, vehicles in the air (uav, drone, …) and vehicles on water (vessel, auv, …). In addition, robots applied to different domains are important. Methodologically, algorithms, scenarios and (their) simulations seem to predominate the current discussion.

In order to identify emerging approaches and analyze their assumptions and safety metrics, a systematic literature review was conduced. Methodologically, we took up the transfer approach by (Kitchenham and Charters, 2007) bringing the empirical study to the technical domain. Figure 1 illustrates the selection model. The selection model uses the databases ieeexplore.ieee.org (IEEE) and www.sciencedirect.com (SD) for searching publications related to safety and autonomy that appeared in the period from 1st January 2016 to 7th April 2020. We limited our search to the past four years because we expected most promising emergent approaches in this period. First, the papers were selected by title and metadata using the search term “autonomy* AND safe*” including different variations like “autonomy AND safety”. The queries were limited to title and metadata, as well as the listed databases to avoid many false positives.

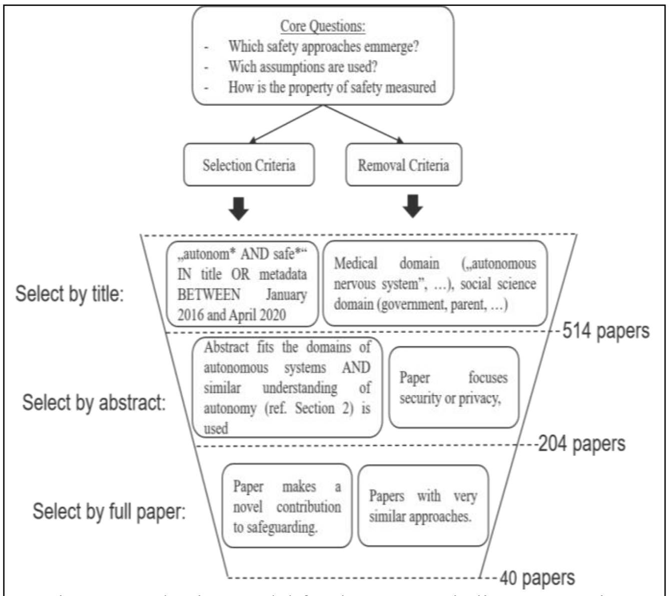

Core Questions:
- Which safety approaches emmerge?
- Wich assumptions are used?
- How is the property of safety measured
Selection Criteria
Removal Criteria
Select by title:
„autonom* AND safe*“ IN title OR metadata BETWEEN January 2016 and April 2020
Medical domain („autonomous nervous system“, …), social science domain (government, parent, …)
514 papers
Select by abstract:
Abstract fits the domains of autonomous systems AND similar understanding of autonomy (ref. Section 2) is used
Paper focuses security or privacy,
204 papers
Select by full paper:
Paper makes a novel contribution to safeguarding.
Papers with very similar approaches.
40 papers

Figure 1: Selection model for the systematic literature review.

The goal of the selection model was to provide a representative set of new contributions to safeguarding autonomous systems from the literature. From the 204 included publications those were excluded that did not comprehensively contribute to the RQs or cover already mentioned approaches. Moreover, we excluded those focusing on security influencing safety since (Leccadito et al., 2018; Koschuch et al., 2019) already addressed this. Based on these exclusion criteria, 40 papers are discussed in this survey. The papers are complemented by further background.

The goal of this systematic literature review is to cluster the state of the art in the domain of safeguarding autonomous systems reviewing different domains. Common ideas of emerging approaches and cross-domain synergies shall be uncovered. Finally, this systematic literature review aims to make the safety measurable.

The first goal of finding cross-domain similarities of emerging approaches is connected with the question of which clusters of tasks or problems respectively need to be solved in order to reach safety for autonomous systems. A similar Research Question (RQ) is posed by (Zhou et al., 2019) in the domain of dependable cyber-physical systems. From a pragmatic point of view, our RQ are:

(RQ1): Which overall tasks need to be addressed in order to reach safe, autonomous systems?

Related to the first goal and crucial for the second goal of reaching synergies between the different domains is the understanding of which mindset is behind the respective emerging approaches and which assumptions are made. We generalized this question from (Burton et al., 2017) focused on safety of machine learning in automated driving. Providing a safety case for certification requires an arguing strategy. This is especially true if you consider a complex system where you rely on assumptions. For this reason, we derived the RQ:

(RQ2): Which assumptions are made building the safety argument?

Finally, we took up the idea of (Murphy and Schreckenghost, 2013) to survey the metrics in the domain of safeguarding. Since some problems in one domain sometimes are already solved in another domain, you rely on similar metrics and a common way of measuring. For this reason, we derived:
(RQ3): Which similar metrics occur cross-domain and how is safety measured?

4 ANALYSIS OF DIFFERENT APPORACHES AND THEIR ASSUMPTIONS

In this section the included articles are presented according to their domain, namely architectures and general approaches, ground vehicles, aerial vehicles, nautical vehicles, industrial robots and production, and medical systems. If applicable, the underlying assumptions and safety metrics are outlined.

4.1 Architectures and General Approaches

Motivated by the paradigm of safety by design several architectures are proposed in different domains and with varying focus. (Kunifuji, 2017) reflects the application of the Heterogeneous Real-time Integrating System (HRTIS) (T. Kunifuji and H. Ito, 2012) architecture to safety-related decentralized and autonomous railway control systems. Therefore, the architecture is extended by safety aspects which are the focus of the paper. However, the paper remains very abstract in the concepts to reach the property of safety assuming things like fault detection just work through internal and external self-test units. As no specific modules for handling uncertainty of the environment are proposed, this approach seems to assume static conditions. The discrete state architecture for swarms of autonomous systems by (Vistbakka et al., 2019) provably guarantees safe reaction at runtime based on Event-B (Abrial, 2010) and the assumption of underlying functions, providing binary states like "connection is lost" or "unknown object detected". The architecture focuses on communication. (Hägele and Söffker, 2016) propose a monitoring architecture for autonomous aerial systems of five modules, namely (1) system state surveillance, (2) communication state surveillance, (3) degree of freedom surveillance, (4) system boundaries surveillance, and (5) safety zones surveillance. The monitoring architecture is complemented by safety appraisal module assessing the current situation and estimating the risks. To handle this task, Hägele et al. propose the Strictly Formalized Situation-Operator-Modeling technique. This is a formalized version of (Söffker, 2008) describing world-object interaction. The architecture is complemented by a safety handling module overriding the high level control. Due to runtime decisions and limitation to fall-back actions, Hägele and Söffker claim situation surveillance and control without action space explosion. The letter of (Hägele and Söffker, 2017) complements the approach by a situation-aware estimator of tolerable risk. The proposed architecture focuses on aerial drones, thus assuming domain-specific safety tasks like safety zone surveillance. Assumptions made are: Reliable hazard detection and prediction techniques available, state model and hazard model available, atomic state transitions and action space modeled during design time. More assumptions from (Hägele and Söffker, 2017) are: Simplified vehicle model, predefined emergency actions, risk assessment at design time is reliable, i.e. risk remaining the same, and failsafe control. The article of (Spislaender and Saglietti, 2018) provides an approach for verification of safety properties that can be expressed by Computational Tree Logic properties in unrestricted Extendet Finite State Machines (EFSM). The approach extends conventional model checking by decidability for unrestricted EFSMs. The main idea is to transform a verification problem into a test coverage problem. To cope with the decidability problem, heuristics are exploited: “Due to inherent limits of decidability, a heuristic search for universal paths fulfilling property-specific coverage criteria was carried out via simulated annealing” (Spislaender and Saglietti, 2018). Simulations provide evidence pro/con universal properties with automated test case generation.

Besides the architectures there are several general approaches providing methodology to cope with safe autonomous systems. The approach for safety of learning components by (Tuncali et al., 2018) utilizes linear N-dimensional decision boundary separating safe from unsafe states. This decision boundary is called barrier certificate, where a “barrier certificate is a differentiable function B from the set of states of the dynamical system to the set of reals” (Tuncali et al., 2018). The approach transforms a safety verification problem to the identification of that barrier certificate based on simulations. A posteriori, the synthesized barrier certificate is verified. The barrier certificate is a level set of a generator function, assumed as positive function that decreases along the system trajectory. The approach assumes that these generator functions or barrier certificates do exist and are known. The demonstration of the concept is done on a low dimensional problem which is implicit assumption.
(Tadewos et al., 2019) proposes an approach to automatically generate a behaviour tree of an autonomous vehicle that satisfies both goals: search and delivery, and safety. In a nutshell, the concept is an automatic transformation from dynamic differential logic to behaviour trees. The approach assumes leaf nodes of behaviour trees to generate discrete states (*running, success, fail*) derived from modularized sub-functions. This information is propagated through the whole graph. Any node or sub-node needs to return in order to get a result like "success" or "fail" and robots are assumed to perform “a single action at a time” (Tadewos et al., 2019).
The contribution of (Ezekiel and Lomuscio, 2017) combines fault injection and model checking to model-based fault injection. The approach is designed for generic temporal-epistemic specifications. The core idea is to assess multi agent systems (MAS) against their specification in the presence of (injected) faults. On their way, the authors formalize diagnosability in temporal-epistemic specification and provided a toolkit that analyses fault-tolerance and diagnosability. The approach assumes random, stuck at, or inverted faults of Boolean, Integer or enumeration type. The faulty behaviour is assumed to be triggered whenever the injection action is performed. If the specification still holds although specific fault injected, the system is considered fault-tolerant to this specific fault.
(McAree et al., 2016) propose a model-based design process from interfering discrete logic based coverage check over simulation to real-world testing. The assumed use case, an inspection drone, is limited to a small number of safety

constraints so 100% coverage was reached. Therefore, the compliance to the safety constraints, i.e. keeping enough distance from obstacles, does not run into computational complexity problems.

(Yan et al., 2019) provide an accidental causal scenario search algorithm exploiting the concept of Systems-Theoretic Process Analysis (STPA) (Leveson, 2012) for fully-automatic operation systems. As a result of the STPA, potentially unsafe control actions are identified. The provided approach extends this process by automatically identifying causal scenarios that lead to unsafe control actions in five steps: (1) determine unsafe control, (2) find module's failure modes, (3) eliminate unrelated ones, (4) eliminating causal scenarios not leading to hazard and (5) eliminate operational scenarios without hazards occurring. The approach assumes that a low, fixed number of scenarios define the autonomous system since state space explosion is a concern. Since preconditions are done during design time, the models may not leave their scope e.g. by environmental uncertainties or wear.

(Ramos et al., 2019) discuss human-technic interaction for collision avoiding focusing on human tasks, the possible failures arising from them, and the quantification of the risk. Therefore, they propose the hierarchical task analysis based on IDAC cognitive model which "is an operator behaviour model developed based on many relevant findings from cognitive psychology, behavioural sciences, neuroscience, human factors, field observations, and various first- and second-generation [Human Reliability Analysis] methodologies" (Ramos et al., 2019).

(Allal et al., 2017) consider human error in maintenance on port, namely the case of sea chest strainers. They exploite Human Reliability Analysis Event Trees and Technique for Human Error Rate Prediction (Swain and Guttmann, 1983). Based on their analysis, Allal et *al*. propose error barriers and recovery mechanisms. As in the other paper on human modelling, the risk assessment was performed manually in advance. This required static conditions.

(Valdez Banda et al., 2019) propose a design phase process of five steps for systematically and holistically analyse (potential) hazards and manage them according to the autonomous systems operative context. The approach is based on STPA (Leveson, 2012) and STAMP (Leveson, 2004) and targets the safeguarding of an autonomous ferry. As other design time approaches, this concept relies on fixed number of potential accidents and hazards with a manageable amount of countermeasures. There is no flexibility to unknown scenarios and risk is estimated by current statistics, which may change over time.

Recapitulation: The examined contributions show that no single architecture has yet emerged that cross-domain solves all relevant safety requirements. Within a domain, one can try to develop a template for the design of safety-relevant systems. This could be in the form of design guidelines, patterns etc. ensuring that the essential safety aspects for the specific domain are considered. It remains an open research question whether it is possible to develop a general safety architecture. The examination of the contributions shows that the topics *monitoring* and *learning* across all domains are of particular importance.

4.2 Ground Vehicles

Ground vehicles represent the domain with most publishing activities on safety and autonomy in the last four years. The most studied problem was safe action planning, specifically collision avoidance. The paper of (Vaskov et al., 2019) contributes to this topic applying Reachability-based Trajectory Design (RTD) to the automotive domain. RTD provides provable safe trajectories in real-time. Compared to nonlinear model predictive control and randomly explored random trees, Vaskov et al. claim better performance. The presented version builds on static obstacles, i.e. all obstacles have the same behavior.

(Pecka et al., 2018) propose an approach that incorporates probability bounds of unwanted trials into Contextual Relative Entropy Policy Search method, a reinforcement learning derivate. However, the authors propose to replace Gradient Policy by a cautious physics simulator certifying safety of the policy. Core assumption of this approach is that such a cautious simulator exists, i.e. the simulator never classifies unsafe policies safe.

(J. Chen et al., 2018) propose an imitative learning approach, aiming to control uncertainties. They extend existing imitation learning methods to include a better performance and safety control based on a safe set. Speed and distance dependent ellipses define the safety areas that are not affected by safety controls based on a safe set.

Another approach focuses on stability of autonomous driving in uncertain environments (Nagasaka and Harada, 2016). However, safety and smoothness play a major role, too. The approach makes use of LIDAR and radar for object tracking which behavior planner, path planner and speed planner rely on. The references (look-up table) for subsequent online planning are several paths calculated based on a detailed digital map. This map contains plenty information, e.g. the lane position and associated traffic rules. For simplicity, the authors reduce the online path-planning problem to two dimensions. Next, Bezier curves interpolate the reference path to the current situation. However, this approach relies on the up-to-dateness of the very detailed map and reliable object tracking, and location and speed have to characterize the obstacles appropriate, which did not always work well in their experiments.

Other approaches for safe action planning intent to prove safe action planning formally. The authors (Konda et al., 2019) contribute to prove safety of autonomous systems in chaotic environments like traffic circles. They focus on collision avoidance and lane keeping in traffic circles using control barrier functions. They assume traffic circles as disks but propose to composite them with others to handle complex safety constraints. To construct provably correct control barrier functions, the authors propose nominal evasive maneuvers, which are standard actions to take in order to get out of a risky situation. In order to get a mathematical proof, roundabout is simplified to 2d and position as well as speed of road users is assumed known and precise.

Another approach (S. Magdici and M. Althoff, 2016), which is also based on safe sets, provides an emergency trajectory in case an unexpected event occurs. It uses variable models for a very specific subset, namely navigation, to ensure the safety of the vehicle during operation. Further concerns are limited resources or limited capabilities of the specific algorithms. In order to complement different algorithms, the approach (Ramakrishna et al., 2019) exploits weighted simplex strategy based supervised safety control. Weighted simplex strategy uses high performance but unverified controller but activates high assurance controller whenever high performance controller tends to violate safety constraints. However, each controller has its strength, which is why the controller output is weighted. The authors compared simple weighted simplex strategy to context-sensitive weighted simplex strategy. The first strategy just computes weighted controller output by comparing supervisor and learning component where the latter incorporates context information learned by reinforcement learning. Moreover, the authors provide a Bayesian network based risk monitor providing estimated probability of the system staying in safe region. Evidence for confidence increases over time. Finally, the approach reliably identifies safe turn region and the ranges for commands keeping the vehicle on track. The assumptions behind this approach are mainly that the two incorporated algorithms complement each other in a way that covers all scenarios the system may get into. Moreover, the estimation, which control algorithm should takeover must be reliable although a situation occurs for the first time.

Moreover, there are contributions to safety assessment. In their approach, (Xu et al., 2019) develop a quantitative approach for safety assessment. The Evaluation of safety relies on operational verification based on Stochastic Hybrid Automata. The approach regards the decision-making as a periodic control system with monitored constraints. Since autonomous systems' safety mainly depend on the recognition of the environment by means of tons of space-time data, they reduce the complexity using representation of the environmental data in an abstract formalized feature model. For their evaluation, they take Single-Lane Roundabout Scenario to show how to verify quantitative properties of the safety assessment with UPPAAL SMC. The authors assume a knowledge library that provides a probability distribution based on historical data. Moreover, all vehicles must obey the traffic rules. For safety judgement, the authors subdivide the roundabout in segments where only one vehicle may enter a segment.

In their model based safety analysis (MBSA) approach, (Tlig et al., 2018) propose using modular numerical simulation platform for handling safety considerations during design and validation phase. Therefore, they transfer MBSA from avionic domain to Traffic Jam Chauffeur, an autonomous driving function deriving potentially critical scenarios. The modelling is tree-like and very abstract. Because of the very abstract point of view, smart sensors (Radar and Camera) presenting presence and distance of identified objects are assumed. Moreover, Confidence of the sources is subdivided in four discrete levels where the authors assume error-free conservative fusion, where most serious information wins. Moreover, they neglect errors in control part.

In their tutorial, (Cheng et al., 2019) present a toolbox that combines three previously proposed mechanisms namely (1) quantitative k-projection coverage, a new dependability metric, (2) a formal reasoning engine assuring proper generalization, and (3) runtime neuron activation pattern monitoring, which searches neuronal activation patterns in historical activations for similarities. All those methods are designed to reduce risk of design fault of artificial neural networks as they are widely used in the development of autonomous driving systems. The aim is to prove the reduction of uncertainties in design time in a well-structured manner. The approach assumes that all hardware and software faults are handled by classic safety methods and therefore do not occur.

Another monitoring approach (Philippe et al., 2016) propose a linear Model Predictive Control (MPC) surveilled by a safety module that balances both trajectory smoothness (comfort) and safety. Safety monitor computes risk according to the MPC model accuracy and the predicted error. Starting from classic MPC problem, the authors derive linearization due to complexity problems. However, this come with the assumption that linear behavior is adequate since actuator saturation do not occur and tracking precision is sufficiently high.

In order to meet the need for safety in unknown environments (A. Bajcsy et al., 2019) propose a real-time safety analysis based on Hamilton Jacobi accessibility to calculate the Backward Reachable Set in real-time. Here the main idea is to create a specific model, i.e. an environment map, according to a fixed scheme at runtime and to identify a secure environment in this map using the Backward Reachable Set.

(Machin et al., 2018) also contribute to this topic. Their Safety MOnitoring Framework (SMOF) generates automatically safety rules based on safety margins. The approach builds on formal verification techniques and hazard analysis to synchronize regulations. However, the needs trusted information sources.

A very specific approach for autonomous bus transit systems by (Han et al., 2019) targets the special technical challenges of buses namely extra dimension, complex ego-system, i.e. varying number of passengers, complicated structure of kinematics and dynamic constraints etc., and highest degree of safety guarantee in urban environment. Therefore, the approach provides an optimization concept for number and position of different sensors dependent on bus size and sensor fusion performance. Moreover, the authors propose an extended motion-planning algorithm based on closed-loop Rapidly-exploring Random Trees (CL-RRT) updating the environmental model at each iteration. The algorithm evaluates the randomly sampled trajectories with probabilistic safety constraints. Since the approach mainly considers the sensor quality the assumption made is that the sensors monitor the environment at any time in an appropriate way and thus the sensors provide all information needed. Moreover, the approach relies on static object detection model.

Recapitulation: In this section, we surveyed contributions to autonomous ground vehicles. Most dominantly, safe action planning approaches are proposed. However, runtime monitoring and advanced safety analysis like specific model-based design approaches gain interest. Keeping the models up-to-date is still a challenge.

4.3 Nautical Vehicles

Map setup and synchronization, situation awareness, and safe path planning and collision avoidance are dominant topics for autonomous nautical systems.

Motivated by providing situation awareness to autonomous ships, (Murray and Perera, 2018) propose a data-driven vessel trajectory prediction algorithm for the time horizon of 5-30 minutes. They assume the Automatic Identification System to reliably detect all vessels. The artificial intelligence exploits this data to predict the vessels trajectories combining Single Point Neighbor Search Method (SPNSM) with Multiple Trajectory Extraction Method (MTEM). SPNSM is a fast trajectory prediction algorithm consisting of trajectory isolation, data clustering around a given position, and iterative prediction. The latter interpolates the new position based on the old one and the similar historical trajectories detected. MTEM complements SPNSM in the more accurate speed estimation improving the data set the SPNSM relies on. Moreover, a statistical analysis is proposed to improve the results to outperform state-of-the-art. The assumption is knowing the trajectories of other ships in the environment, thus providing situation awareness. Moreover, it is assumed that enough representative historical trajectories are available to learn from. The approach does not provide a solution for abnormal behaviour of other ships or the context (e.g. weather).

(Yoo et al., 2018) propose a stochastic path planning algorithm under model uncertainties for simulated underwater gliders. Equipped with minimal sensing capabilities their state models degrade due to unforeseen disturbances till the glider returns to surface and receives GPS signal. Under the assumption of valid Markov property they formulate two recursive objective functions and exploit stochastic (Fast Marching Tree)*-algorithm (Janson et al., 2013) for finding a trade-off between travel cost and safety. So, another assumption is that these objective functions can be derived.

(Hernández et al., 2018) propose another approach contributing to path planning for autonomous underwater vehicles. Their framework proposes the modules mapping, planning and mission handling. In this paper, focus is on including motion constraints, plan doable trajectories and perform risk estimation. Planning is done iteratively exploiting last known solution to safe computational power. The core extension is doing path planning on (Rapid Random Trees)*-algorithm (Karaman and Frazzoli, 2011) modified by the Dubins vehicle model (Savla et al., 2005) for more computational efficiency. The main contribution to safety is the introduced risk function. It combines (1) path length and clearance (2) predefined heuristically risk zones and (3) direction vector risks considering moving direction. However, the increase in efficiency is bought by constraints, i.e. assumptions. First, the models in use are simplifications as the 3d vehicle just considers 2d mapping. Moreover, Dubbin's vehicle model is a simplification bound to constraints. The definition of heuristics for risk zones demands for well-known properties of these areas. Applied risk checking however, demands for certainty in choosing unreachable unexplored regions. Taking last best solution implies implicit assumption of moderate changing environmental conditions.

(Shen et al., 2017) present an approach to create a grid-world map of underwater terrain. The map is sliced in 2-D planes and modelled in multi-level coverage trees. The trees are generated online whereas the autonomous vehicle calculates the seabed reconstruction offline. As shown by simulation, the accurate map contributes to safety since it enables path planning far away from obstacles. Furthermore, the obstacle density in neighboring cells are analyzed. The approach assumes bounded search space for a single model, namely the map. In this bounded space the coverage tree reveals unsearched areas. Moreover, the task is mapping implying no trade-off between environmental knowledge and economically reach goals has to be taken.

Recapitulation: Solutions for concrete nautical problems already exist. Especially for underwater vehicles, building maps and allocating the own position on it is a major concern, since no satellite positioning is available. As most of the approaches only address the field of path planning and collision avoidance under these specific constraints, they are hardly transferable to general problems.

4.4 Aerial Vehicles

The approaches for safeguarding aerial vehicles cover a wide range from data fusion to interpretable decision processes.

(Hasan et al., 2019) propose an autonomous parachute safety system for hexacopters operating outside the visual range. The approach uses the eXogeneous Kalman filter (Johansen and Fossen, 2017) to reliably detect altitude drops from various data sources to activate the fall-back system in case of emergency. The approach assumes white noise and bias as well as Lipschitz nonlinear uncertainty. Suitable tuning parameters and linearizability of the nonlinear observers are assumed.

The paper of (Yel and Bezzo, 2018) describes a Gaussian Process Theory based reachability analysis approach for handling signal losses of autonomous aerial drones. In order to accelerate processing time at runtime, the authors rely on a pre-computed library of primitive trajectories of the drone under various disturbances and the maximum deviations. At runtime, the Gaussian Process regression estimates the maximum deviation of the vehicle for its current trajectory. As time without signal present passes the potential, deviation grows and model uncertainty raises. In the moment, signal gets lost, the drone operates on its model up to the time that either the signal comes back or the reachability analysis intervenes. The reachability analysis intervenes in the moment, where the predicted worst-case deviation violation of a safety constraint (e.g. collision with obstacle). In this case, the monitoring triggers a predefined signal recovery manoeuvre. The prediction includes the uncertainty added by the recovery manoeuvre. However, since the recovery manoeuvre is hardcoded, the approach relies on a fixed set of actions, which brings the system back to a safe state. Moreover, since primitives are pre-computed, the approach relies on new disturbances resembles the recorded ones.

The approach of (Snisarevska et al., 2018) deals with the highly stochastic process of start and landing scheduling of airplanes. A scheduling assistance system is proposed which estimates the stochastic distribution of start and landing times and thus balancing safety margin and throughput. Therefore, the approach assumes predefined minimal safety margin given and that the distribution of interest can be accurately estimated.

(Di Franco and Bezzo, 2020) propose applying decision trees at runtime to interpretably monitor quadcopter behaviour and therefore avoid collisions. The baseline concept is inspired from the LIME (Ribeiro et al., 2016) and LORE (Guidotti et al., 2018) concept. Trajectories are classified safe or unsafe according to generated training set trajectories which are transformed according to the current scene. The obstacle model is reduced to 2d circles. If a planned trajectory classified unsafe the re-planning of the trajectory is triggered using another decision tree. So, the trajectory is found which is most similar to the current one and maneuverers the drone in out of risky situation. Since the decision trees are rule-based, Di Franco and Bezzo argue to provide interpretability contrast to the state-of-the-art AI black box models. However, "[s]imilar to other machine learning techniques, we observe that larger and diverse training sets produce more accurate predictions and explanations" (Di Franco and Bezzo, 2020). Therefore, sufficient data sets available are assumed. Moreover, since rule-based systems can hardly cope with overlapping data sets, completely separable datasets with bounded disturbances are assumed, i.e. the context is fixed and whatever worked once is assumed to work again. Moreover, if the problem gets too high-dimensional, the interpretability of the rule set suffers.

In (Vierhauser et al., 2019), Vierhauser et al. propose a design time methodology to create interlocking safety cases, i.e. assign interlocking checkpoints to other safety cases. They target UAV infrastructure. The infrastructure's Safety Case represents general safety goals. Pluggable Safety Cases complementing them by modelling the vehicle-specific attributes indicating compliance with the overall constraints and thus providing safe operation.

Recapitulation: Safety is a huge topic with aerial vehicles with already existing standards. However, those standards only regulate hardware and software, not the needed intelligence for autonomy. Therefore, new approaches are needed to safeguard autonomous aerial vehicles. Currently many approaches are developed for unmanned aerial vehicles, where new safeguarding concepts as well as safety standards emerge. However, standardization for safe autonomy and development of new concepts dealing with safeguarding intelligent algorithms are still in their infancy.

4.5 Industrial Robots and Production

In their paper, (Bank et al., 2018) present a Linear Temporal Logic based approach for assuring safe operation at runtime. The provided software first generates a formal problem specification and a state machine of the assembly process. Afterwards, the NuSVM BMC solver processes these artefacts generating safe and feasible meta-level plan. Finally, motion planner uses this meta-level plan for execution on shop-floor. Since exploration spaces grows exponentially with the input task's complexity, the software first subdivides the task in subtasks to limit processing time to linear growth. However, this divide-and-conquer approach relies on loose coupling between the subtasks, since emergent effects in their interplay have to be modelled manually in advance.

(Legashev et al., 2019) argue for an online certification unit for semi-automated re-certification of autonomous robotic systems in case of legal requirement changes or new insights. The envisioned cloud platform enables user to monitor telemetry data and trigger new certification tests. The certification module takes changed requirements and selects test cases for execution on the autonomous robot. The test module then executes the corresponding test cases by simulating or providing the input information, reads out the telemetry data and judges the result of the test. However, this approach assumes reliable connection to the cloud, the limitation that the test cases (even considering ill-functioning devices) cannot aggravate system state and that this test cases match any individual system. Moreover, since there is no hint of how new test cases get on the system, a mechanism providing this service is assumed.

(Omori et al., 2018) present an approach for safe way planning of bipedal robots proposes the fusion of visual (RGBD) and tactile information determining safe ground to step onto. The safety concept uses object detection for identifying know objects. Safed in a database, the safety cost for stepping on the object is loaded and the algorithms calculate optimal path. However, if the database does not contain a specific object, a safety test mode carefully steps onto this object, measures reacting force, the safety metric, and therefore calculates safety cost. The approach assumes all objects to be large enough and static in their properties, and detection to be reliable.

Recapitulation: The field of industrial robots and smart manufacturing differs significantly from the three previous fields. On the one hand, it is much more static. Consequently, external influences are limited. The field is reduced to the shop floor, thus encapsulated from the outside. On the other hand, more diverse components interact with each other. Therefore, different approaches are needed for industrial robots and smart manufacturing than for vehicles. This complexity of the various components can also be seen in the differing approaches presented in this section.

4.6 Medical and Healthcare

In the medical and healthcare domain, very few publications handle both autonomy and safety. (Ye et al., 2020) provide a robot monitoring framework for semi-autonomous brain-biopsy. Moreover, the contribution provides a concrete example of how to develop a safe system from scratch to product. Safety is conducted with hardware and software providing collision avoidance. The approach is validated through simulation and experiment. However, no classification for safety is provided and the concept is very specific for its use case. (Ma et al., 2019), the authors propose a concept to flexibly control endoscope. Optimum control at minimum movement inside the body safes the patient from unintended wounds. Experimental comparison to Rigid-Endoscope shows promising results.

(Haidegger, 2019) provides a comprehensive survey on autonomous medical robots and the time course of their development. They therefore provide a reason of why only few contributions are found under the mentioned keywords. Providing a classification into different degrees of autonomy and defining procedures for measuring the degree of autonomy (DoA, LoA), they conclude that only sub-functions or sub-tasks of the robots are equipped with some sort of autonomy (LOA 1-3). They propose the development of new standards for pushing development in this area. Besides that, (Haidegger, 2019) predict near breakthroughs of autonomy in medical domain.

Recapitulation: Medical applications are very safety critical. For this reason, many hurdles must be overcome to get a certification for an autonomous medical system. Therefore, only a few very specific autonomous systems for very specific use cases exist today.

5 DISCUSSION

Taking the domains spotted by (McKee et al., 2018), approaches for safeguarding can be found for all of the domains they suggest. However, there are major differences in the number of publications found for the specific domains. While ground vehicles unite *14* contributions, medical robots and safeguarding smart manufacturing only count *3* and *2* contributions. However, there are similarities between the domains from an abstract point of view, on which the research questions posed in Section 3 focus. The research questions are discussed each in a separate subsection.

5.1 Which overall tasks need to be addressed in order to reach safe, autonomous systems? (RQ1)

From a very abstract point of view, all reviewed papers match into five fundamental clusters of tasks, namely (1) monitoring, (2) safe action planning, (3) model synchronisation and data fusion, (4) comprehensive decision making, and (5) advanced risk assessment. Monitoring focuses on the recognition and the judgement of unexpected or unwanted environmental conditions or system states. Safe action planning considers decision making that cares about safety constraints. Model synchronisation and data fusion copes with providing adequate information based on latest process data. This ensures decisions made under right assumptions. The comprehensive decision making considers interpretability of the decisions made. And advanced risk assessment provides the information of on what the system has to care for. Based on the literature research, the authors are of the opinion that these five basic clusters must exist in one form or another in all autonomous systems to ensure safety. Since the autonomous systems' characteristic is to adapt to new situations, the system must notice that it has to adapt. Without (1) monitoring, you cannot detect changes in the environment. Having recognized a change in the environment, the system must independently act on this change requiring for (2) safe action planning. However, this action must be based on models since the actions are chosen at runtime. Obviously, the decisions based on the models are only guaranteed to be safe if the information they rely on, i.e. the models are correct. Because the autonomous system adapts, models also have to be adapted which is represented in (3). Because there are that many changes in the system over time, forecasting the system's behaviour in a black-box manner gets really hard. Therefore, the system itself must provide human operator with interpretable information about the decision making process, referred to in (4). This information sets the operator in position to judge and predict system behaviour and therefore make use of the system without endanger himself and its environment. Finally, since the risks get hidden under the veil of complexity, (5) advanced risk assessment is obligate.

However, parts of the five fundamentals may be supplemented by assumptions like the orthogonality of the sensor data where no data fusion is needed, or safe environment where neither safety concerns on action planning nor comprehensive decision making is essential. If he leaves out any part, the developer must be really sure that the underlying assumptions are explicitly stated and hold for the envisioned use case.

From the reviewed papers, 9 papers consider (1) monitoring, 20 papers consider (2) safe action planning, 12 papers consider (3) model synchronisation and data fusion, 2 papers consider (4) comprehensive decision making, and 12 papers consider (5) advanced risk assessment. It is no surprise that most publications care about risk assessment and safe task planning since risk assessment is an evolution of best practice focusing on special issues in autonomous systems and safe action planning is absolutely fundamental. However, the huge discrepancy to the other clusters reveals a gap. Specifically, the comprehensive decision making is crucial to acceptance and understanding of those new systems. Moreover, the underrepresentation of monitoring is surprising since many approaches generally cope with this topic, e.g. fault-detection (Punčochář and Škach, 2018). It would be very interesting which approaches apply to autonomous systems. Finally, concerning model synchronization, mainly map actualization is considered. However, closing knowledge gaps in the models, fusing different information sources and judging the reliability of them remains a big issue. (Hasan et al., 2019) and (Müller et al., 2019) contribute to this field, but much more research is needed. Furthermore, the approaches mostly focus on one single aspect, seldom on two or more. A holistic cross-domain approach combining all five aspects is still missing.

As the metrics differ, the safety architectures, design patterns etc. differ as well across the domains. Nevertheless, they follow similar goals and implement the same fundamental tasks. We identified three basic clusters of tasks which are (1) monitoring, (2) safe action planning, and (3) model synchronisation and data fusion. They appear crucial for any safe autonomous systems, since addressed in every domain. However, although not addressed in every domain, (4) comprehensive decision making is very important, too. Moreover, (5) advanced risk assessment can be identified as common cluster. Table 2 lists the percentage of approaches contributing to the respective task over the cited publications. Note that the approaches might contribute to more than one task. Moreover, there are general approaches, which treat the topic without a concrete application to one single domain.

Table 2. Table of cross-relations between domain and task

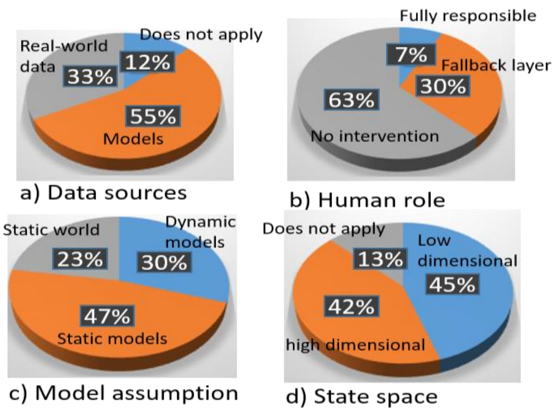
Domain	**Addressed in X % of the approaches**				
	(1)	(2)	(3)	(4)	(5)
General	22.2	30.6	0.0	0.0	47.2
Ground vehicle	14.3	52.4	16.7	0.0	16.7
Nautical vehicle	0.0	62.5	25.0	0.0	12.5
Aerial vehicle	6.7	16.7	70.0	6.7	0.0
Smart manufacturing	0.0	16.7	16.7	33.3	33.3
Medical and Healthcare	50.0	50.0	0.0	0.0	0.0

In conclusion, the focus on the identified tasks differs because the requirements on the systems vary. It also shows in which domain practitioners might start to search if the respective task becomes important in his or her domain.

5.2 *Which assumptions are made building the safety argument? (RQ2)*

The assumptions made in the reviewed approaches refer to the environment model, the human responsibility, dimensionality of the state space and the available data. Results are visualized in Figure 2. Environmental model is assumed static (23%), dynamic but with static models (47%) or dynamic with changing models (30%). The high rate of static assumed environment is a comprehensive simplification since model management is often neglected as found in Section 4.1. However, it is surprising since coping with unpredictable environments is the core difference between automated systems and autonomous systems. Only 9 approaches fundamentally update their models and all these approaches only consider the map. Combining the concept of the Digital Twin with artificial intelligence (Jazdi et al., 2020) may help to fill the gap. But much research remains to be done. Concerning operator's role, it is not surprising that most approaches (63%) do not involve humans, since this is property of full autonomy. However, 30% of the approaches consider human-in-the-loop, where 25% of these approaches explicitly assess human error risk. It is worth noting that state space negatively correlates with full autonomy. Most approaches (45%) face a proportionally low-dimensional state space. High-dimensional problems generally relay on human fall-back layer. The data needed to power the approaches is manly simulated or delivered by a model (55%). Other approaches (12%), namely design time risk assessment does not relay on any operation time information. 33% of the approaches relay on historical real-world training data.

Related to the assumptions is the safety argument. As visualised in Figure 3, in this area four strategies are spotted in the articles, namely mathematical prove (17.5%), test and evidence collection (70%), systematic decision-making (32.5%), and safety by design (7.5%). In this context, safety by design refers to design measures like special architecture or model-based analysis with the goal to improve safety. Low popularity of strict proves is reasonable since guaranteeing preconditions is quite hard in unpredictable environments. Safety by design is a weak argument for the same reason. Using test and evidence leading matches the expectation since it is the state of the art approach. However, this approach is not practical for open world systems. Therefore, a new strategy argues about the way decisions are made. This approach seems reasonable against the background of modern jurisdiction, which judges mainly on the basis of the course of an accident and enjoys growing popularity.



a) Data sources b) Human role c) Model assumption d) State space

Figure 2: Assumptions on data sources, human role, models and state space

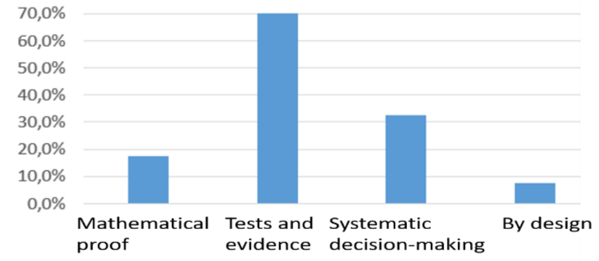


Figure 3: safety argument for the safety cases

5.3 *Which similar metrics occur cross-domain and how is safety measured? (RQ3)*

The main metric for safety across all domains is the risk of causing harm to the environment, i.e. the product of the extent of damage and probability of occurrence. This main metric is defined in basic norms like ISO 61508. However, these two terms (probability of occurrence and loss) are themselves abstract. The question therefore arises as to how these two components of risk are measured. It has to be considered that autonomous systems differ a lot across the domains, e.g. nuclear power plants vs. consumer drones. Therefore, the commonalities of the safety concepts cease the more the domains differ. While under vehicles the creation of a reliable map and collision-free trajectory planning are connecting elements, the focus for industrial robots, namely human-machine cooperation, differs significantly. This discrepancy also shows off in the choice of metrics. The extent of damage is estimated on a case-by-case basis using risk assessment. The methodology is similar in principle. However, the probability of occurrence can usually only be estimated indirectly and is often approximated by specific variables based on the risk assessment (50%). In the field of vehicles, at least the distance to obstacles can be identified as a common metric (55%). Otherwise, the differences are significant even within a domain. The metrics are derived from the safety assessment and are dedicated for the specific use cases. In this regard, there is no uniform metric measuring probability of occurrence and losses. Figure 4 shows the main clusters of the used metrics. In the field of vehicles and robots, the distance to the obstacle is often used as a metric to measure the occurrence probability of collisions, with collisions being assigned a uniformly high worst-case cost. In addition, detection rate is a relatively common means of estimating probability in object detection. The detection rate refers to the ratio of properly detected objects to missed or misclassified objects. This can be for instance the rate of detected critical scenarios, the classification of dangerous trajectories etc. In this case, a cross-entropy table estimates the extent of damage. However, about the half of the reviewed publications derived their metrics from the risk analysis. As a result, it is worth to discuss the proposed methodologies of how to identify the appropriate safety metric. In this context, the STPA is proposed by some authors, e.g. (Yan et al., 2019; Valdez Banda et al., 2019). However, the analysis of how to derive appropriate domain specific safety metrics is left for future work.

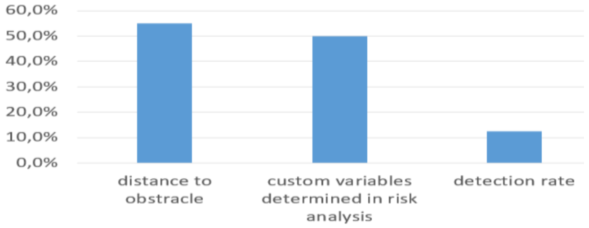


Figure 4: metrics for measuring safety

6 CONCLUSIONS

As the demand for autonomous systems increases and the requirements for their use in uncertain environments grow, the development of new, adapted safety concepts is of crucial importance. According to the systematic search criteria discussed in Section 2, *40* approaches from five domains remain. Bibliometric analysis indicates that these papers represent the literature quite well.

Towards the **question of cross-domains clusters and similarities**: An important trend that is emerging in the safeguarding of autonomous systems is to complement the safety assessment in the design phase with a safety assessment during operation in order to overcome the problem of state space explosion. Therefore, model-based approaches are gaining importance. Nevertheless, it remains a great challenge to make decisions based on a situation- and context-related safety analysis during operation in a comprehensible way.

Concerning the **question of which mindset** or more specifically which assumptions are behind the safety cases of the emerging approaches, the following is notable: As complexity increases, one trend to argue for the safety of an autonomous system is to make the way the system makes decisions transparent and plausible. In contrast to the black-box approach, which shows that the system reacts as desired in every plausible situation, this grey-box approach requires significantly fewer test cases. However, the creation of human-interpretable decision-making processes capable of dealing with changing environments are an open problem. Something like a guardian is needed to protect against unsafe actions during the execution phase according to a systematic and reasonable strategy. Moreover, most approaches still assume a static environment, which is difficult to guarantee for autonomous systems.

The **question of how to measure safety** seems easy to answer. According to the definition of ISO 61508, safety is the absence of unacceptable risks. For this reason, safety is measured in terms of risk. However, risk, defined as the product of probability and loss, is difficult to measure directly. Therefore, most safety metrics are domain-specific and correlate with risk rather than representing risk directly. In the domain of vehicles and robots, the likelihood of collision is dominant where the distance to obstacles serve as metric. Either tables or a single overestimation estimates the loss of the collision. Therefore, the methodology for determining the appropriate risk metrics is the critical issue. In this area, STPA is a new complementary approach to classic FMEA. However, a detailed analysis of how to identify the appropriate metrics for measuring the safety of autonomous systems is not the focus of this paper and needs further research.

Based on the literature review we spot the following **emerging trends**. The first trend is the focus on runtime risk assessment to complement classical risk assessment. New risk analysis methods such as STPA and statistical or machine learning based methods are used here. Moreover, some classical methods are experiencing a renaissance when combined with machine learning. One example are barrier certificates, which provide safety guarantees in combination with neural networks. In addition, there is a trend towards the explainability of artificial intelligence (AI) procedures, on the one hand by using classical procedures such as Model Predictive Control as a supplement, and on the other hand by improving the explainability of AI procedures themselves.

In our survey, we spotted three **open issues** in safeguarding autonomous systems, resulting from the analysis of the three RQs discussed, namely (1) taking the studied approaches, they often highlight the uncertainties of the environment as great challenge, requiring for adaption of the system. The question of how to provide models with adaptability in order to recognize and react on changed conditions requires further work. (2) The literature review shows that emerging approaches tend to focus on safeguarding at runtime. However, the question of how to guarantee safe action at runtime despite uncertain environment remains. (3) The complexity of autonomous systems raises and the incorporation of machine learning makes the systems hard to understand. However, the success of a safety case depends on the appropriate assumptions and therefore on proper understanding. This requires for solutions on the field of how decision processes become human-interpretable.

7 REFERENCES

Abbass HA, Scholz J, Reid DJ. Foundations of trusted autonomy. Cham, Switzerland: Springer Open, 2018.

Abrial J. Modeling in Event-B - System and Software Engineering. undefined 2010.

Allal AA, Mansouri K, Qbadou M, Youssfi M. Task human reliability analysis for a safe operation of autonomous ship. In: 2nd International Conference 2017; 2017. p. 74–81.

ark-funds.com. ARKQ - Autonomous Technology & Robotics ETF by ARK Invest, https://ark-funds.com/arkq; 2021 [accessed July 2, 2021].

Bank HS, D'souza S, Rasam A. Temporal Logic (TL)-Based Autonomy for Smart Manufacturing Systems. Procedia Manufacturing 2018; 26: 1221–9.

Burton S, Gauerhof L, Heinzemann C. Making the Case for Safety of Machine Learning in Highly Automated Driving. In: Tonetta S, Schoitsch E, Bitsch F, editors. Computer safety, reliability, and security. LNCS sublibrary. SL 2, Programming and software engineering. 10489. Cham, Switzerland: Springer; 2017. p. 5–16.

Cheng C, Huang C, Nührenberg G. nn-dependability-kit: Engineering Neural Networks for Safety-Critical Autonomous

Driving Systems. In: 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD); 2019. p. 1–6.

Di Franco C, Bezzo N. Interpretable Run-Time Monitoring and Replanning for Safe Autonomous Systems Operations. IEEE Robotics and Automation Letters 2020; 5(2): 2427–34.

Ezekiel J, Lomuscio A. Combining fault injection and model checking to verify fault tolerance, recoverability, and diagnosability in multi-agent systems. Information and Computation 2017; 254: 167–94.

Fritz R, Zhang P. Overview of fault-tolerant control methods for discrete event systems. IFAC-PapersOnLine 2018; 51(24): 88–95.

Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F. Local Rule-Based Explanations of Black Box Decision Systems, 2018.

Hägele G, Söffker D. Strictly Formalized Situation-Operator-Modeling technique for fall-back layer modeling for autonomous or semi-autonomous systems requiring software-based fail-safe behavior. In: IEEE International Conference 2016; 2016. p. 886–891.

Hägele G, Söffker D. Safety unit-based safe behavior assurance for autonomous and semi-autonomous aerial systems: Requirements, concept, and simulation results. In: IEEE Intelligent Vehicles Symposium 2017; 2017. p. 1546–1551.

Haidegger T. Autonomy for Surgical Robots: Concepts and Paradigms. IEEE Transactions on Medical Robotics and Bionics 2019; 1(2): 65–76.

Han F, Li D, Hao Q. Autonomous Driving Framework for Bus Transit Systems Towards Operation Safety and Robustness*. In: IEEE Intelligent Transportation Systems 2019; 2019. p. 2778–2784.

Hasan A, Tofterup V, Jensen K. Model-Based Fail-Safe Module for Autonomous Multirotor UAVs with Parachute Systems. In: International Conference on Unmanned 2019; 2019. p. 406–412.

Hayat S, Yanmaz E, Muzaffar R. Survey on Unmanned Aerial Vehicle Networks for Civil Applications: A Communications Viewpoint. IEEE Commun. Surv. Tutorials 2016; 18(4): 2624–61.

Hernández JD, Moll M, Vidal E, Carreras M, Kavraki LE. Planning feasible and safe paths online for autonomous underwater vehicles in unknown environments. In: IEEE/RSJ International Conference 2018; 2018. p. 1313–1320.

Jahan F, Sun W, Niyaz Q, Alam M. Security Modeling of Autonomous Systems. ACM Comput. Surv. 2019; 52(5): 1–34.

Janson L, Schmerling E, Clark A, Pavone M. Fast Marching Tree: a Fast Marching Sampling-Based Method for Optimal Motion Planning in Many Dimensions, 2013.

Jazdi N, Ashtari Talkhestani B, Maschler B, Weyrich M. Realization of AI-enhanced industrial automation systems using intelligent Digital Twins. Universität Stuttgart, 2020.

Johansen TA, Fossen TI. The eXogenous Kalman Filter (XKF). International Journal of Control 2017; 90(2): 161–7.

Juric M, Sandic A, Brcic M. AI safety: state of the field through quantitative lens, 2020.

Karaman S, Frazzoli E. Sampling-based Algorithms for Optimal Motion Planning, 2011.

Kitchenham B, Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering, 2007.

Konda R, Squires E, Pierpaoli P, Egerstedt M, Coogan S. Provably-Safe Autonomous Navigation of Traffic Circles. In: IEEE CCTA 2019. [Piscataway, New Jersey]: IEEE; 2019. p. 876–881.

Koschuch M, Sebron W, Szalay Z, Török Á, Tschiürtz H, Wahl I. Safety & Security in the Context of Autonomous Driving. In: IEEE International Conference 2019; 2019. p. 1–7.

Kunifuji T. Safety Technologies in Autonomous Decentralized Railway Control System. In: IEEE 13th International Symposium 2017; 2017. p. 137–142.

Leccadito M, Bakker T, Klenke R, Elks C. A survey on securing UAS cyber physical systems. IEEE Aerosp. Electron. Syst. Mag. 2018; 33(10): 22–32.

Legashev LV, Letuta TV, Polezhaev PN, Shukhman AE, Ushakov YA. Monitoring, Certification and Verification of Autonomous Robots and Intelligent Systems: Technical and Legal Approaches. Procedia Computer Science 2019; 150: 544–51.

Leveson N. A new accident model for engineering safer systems. Safety Science 2004; 42(4): 237–70.

Leveson N. STPA: A New Hazard Analysis Technique. Cambridge, Mass.: The MIT Press, 2012. 1 online resource.

Liu H-C, Liu L, Liu N. Risk evaluation approaches in failure mode and effects analysis: A literature review. Expert Systems with Applications 2013; 40(2): 828–38.

Ma X, Song C, Chiu PW, Li Z. Autonomous Flexible Endoscope for Minimally Invasive Surgery With Enhanced Safety. IEEE Robotics and Automation Letters 2019; 4(3): 2607–13.

McAree O, Aitken JM, Veres SM. A model based design framework for safety verification of a semi-autonomous inspection drone. In: 2016 UKACC 11th International Conference on Control (CONTROL); 2016. p. 1–6.

McKee DW, Clement SJ, Almutairi J, Xu J. Survey of advances and challenges in intelligent autonomy for distributed cyber-physical systems. CAAI Transactions on Intelligence Technology 2018; 3(2): 75–82.

Müller J, Gabb M, Buchholz M. A Subjective-Logic-based Reliability Estimation Mechanism for Cooperative Information with Application to IV's Safety. In: 2019 IEEE Intelligent Vehicles Symposium (IV); 2019. p. 1940–1946.

Müller M, Müller T, Ashtari Talkhestani B, Marks P, Jazdi N, Weyrich M. Industrial autonomous systems: a survey on definitions, characteristics and abilities. at - Automatisierungstechnik 2021; 69(1): 3–13.

Murphy RR, Schreckenghost D. Survey of metrics for human-robot interaction. In: Staff I, editor. 2013 8th ACM/IEEE International Conference on Human-Robot Interaction. [Place of publication not identified]: IEEE; 2013. p. 197–198.

Murray B, Perera LP. A Data-Driven Approach to Vessel Trajectory Prediction for Safe Autonomous Ship Operations. In: 2018 Thirteenth International Conference on Digital Information Management (ICDIM); 2018. p. 240–247.

Nagasaka N, Harada M. Towards safe, smooth, and stable path planning for on-road autonomous driving under uncertainty. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). [Piscataway, New Jersey]: IEEE; 2016. p. 795–801.

Omori Y, Kojio Y, Ishikawa T, Kojima K, Sugai F, Kakiuchi Y, et al. Autonomous Safe Locomotion System for Bipedal Robot Applying Vision and Sole Reaction Force to Footstep Planning. In: IEEE/RSJ International Conference 2018; 2018. p. 4891–4898.

Osborne M, Lantair J, Shafiq Z, Zhao X, Robu V, Flynn D, et al. UAS Operators Safety and Reliability Survey: Emerging Technologies towards the Certification of Autonomous UAS. In: 2019 4th International Conference on System Reliability and Safety (ICSRS 2019). Piscataway, NJ: IEEE; 2019. p. 203–212.

Pecka M, Šalanský V, Zimmermann K, Svoboda T. Autonomous flipper control with safety constraints. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018. p. 2889–2894.

Philippe C, Adouane L, Thuilot B, Tsourdos A, Shin H. Safe and Online MPC for Managing Safety and Comfort of Autonomous Vehicles in Urban Environment. In: IEEE 19th International Conference 2016; 2016. p. 300–306.

Punčochář I, Škach J. A Survey of Active Fault Diagnosis Methods. IFAC-PapersOnLine 2018; 51(24): 1091–8.

Ramakrishna S, Dubey A, Burruss MP, Hartsell C, Mahadevan N, Nannapaneni S, et al. Augmenting Learning Components for Safety in Resource Constrained Autonomous Robots. In: 2019 IEEE 22nd International Symposium on Real-Time Distributed Computing (ISORC); 2019. p. 108–117.

Ramos MA, Utne IB, Mosleh A. Collision avoidance on maritime autonomous surface ships: Operators' tasks and human failure events. Safety Science 2019; 116: 33–44.

Ratasich D, Khalid F, Geissler F, Grosu R, Shafique M, Bartocci E. A Roadmap Toward the Resilient Internet of Things for Cyber-Physical Systems. IEEE Access 2019; 7: 13260–83.

Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?". Association for Computing Machinery (ACM), 2016.

Savla K, Bullo F, Frazzoli E. On Traveling Salesperson Problems for Dubins' vehicle: stochastic and dynamic environments. In: 44th IEEE Conference on Decision and Control; 2005. p. 4530–4535.

Shahrdar S, Menezes L, Nojoumian M. A Survey on Trust in Autonomous Systems. In: Arai K, Kapoor S, Bhatia R, editors. Intelligent Computing. Advances in Intelligent Systems and Computing. 857. Cham: Springer; 2018. p. 368–386.

Shen Z, Song J, Mittal K, Gupta S. Autonomous 3-D mapping and safe-path planning for underwater terrain reconstruction using multi-level coverage trees. In: OCEANS 2017 - Anchorage; 2017. p. 1–6.

Snisarevska O, Sherry L, Shortle J, Donohue G. Balancing throughput and safety: An autonomous approach and landing system (AALS). In: ICNS 2018. [Piscataway, New Jersey]: IEEE; 2018.

Söffker D. Interaction of intelligent and autonomous systems – part I: qualitative structuring of interaction. Mathematical and Computer Modelling of Dynamical Systems 2008; 14(4): 303–18.

Spislaender M, Saglietti F. Evidence-Based Verification of Safety Properties Concerning the Cooperation of Autonomous Agents. In: 44th Euromicro Conference 2018; 2018. p. 81–88.

Swain AD, Guttmann HE. Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report, 1983.

T. Kunifuji, H. Ito. Realization of Flexible Railway System by Heterogeneous Real-Time Autonomous Integrating Architecture. In: 2012 32nd International Conference on Distributed Computing Systems Workshops; 2012. p. 390–399.

Tadewos TG, Shamgah L, Karimoddini A. Automatic Safe Behaviour Tree Synthesis for Autonomous Agents. In: 2019 IEEE 58th Conference on Decision and Control (CDC); 2019. p. 2776–2781.

Tlig M, Machin M, Kerneis R, Arbaretier E, Zhao L, Meurville F, et al. Autonomous Driving System: Model Based Safety Analysis. In: 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). [Place of publication not identified]: IEEE; 2018.

Tong W, Hussain A, Bo WX, Maharjan S. Artificial Intelligence for Vehicle-to-Everything: A Survey. IEEE Access 2019; 7: 10823–43.

Tuncali CE, Kapinski J, Ito H, Deshmukh JV. Reasoning about Safety of Learning-Enabled Components in Autonomous Cyber-physical Systems, 2018.

Valdez Banda OA, Kannos S, Goerlandt F, van Gelder PHAJM, Bergström M, Kujala P. A systemic hazard analysis and management process for the concept design phase of an autonomous vessel. Reliability Engineering & System Safety 2019; 191.

Vaskov S, Sharma U, Kousik S, Johnson-Roberson M, Vasudevan R. Guaranteed Safe Reachability-based Trajectory Design for a High-Fidelity Model of an Autonomous Passenger Vehicle. In: 2019 American Control Conference (ACC); 2019. p. 705–710.

Vierhauser M, Bayley S, Wyngaard J, Xiong W, Cheng J, Huseman J, et al. Interlocking Safety Cases for Unmanned Autonomous Systems in Shared Airspaces. IEEE Transactions on Software Engineering 2019: 1.

Vistbakka I, Troubitsyna E, Majd A. Multi-Layered Safety Architecture of Autonomous Systems: Formalising Coordination Perspective. In: IEEE 19th International Symposium 2019; 2019. p. 58–65.

Xu v, Li Q, Guo T, Ao Y, Du D. A Quantitative Safety Verification Approach for the Decision-making Process of Autonomous Driving. In: International Symposium on Theoretical 2019; 2019. p. 128–135.

Yan F, Zhang S, Tang T. Autonomous Train Operational Safety assurance by Accidental Scenarios Searching. In: IEEE Intelligent Transportation Systems 2019; 2019. p. 3488–3495.

Ye M, Li W, Chan DTM, Chiu PWY, Li Z. A Semi-Autonomous Stereotactic Brain Biopsy Robot With Enhanced Safety. IEEE Robotics and Automation Letters 2020; 5(2): 1405–12.

Yel E, Bezzo N. Fast Run-time Monitoring, Replanning, and Recovery for Safe Autonomous System Operations. In: IEEE/RSJ International Conference 2018; 2018. p. 1661–1667.

Yoo C, Anstee S, Fitch R. Stochastic Path Planning for Autonomous Underwater Gliders with Safety Constraints. In: IEEE/RSJ International Conference 2018; 2018. p. 3725–3732.

Zhang J, Li J. Testing and verification of neural-network-based safety-critical control software: A systematic literature review. Information and Software Technology 2020; 123: 106296.

Zhou P, Zuo D, Hou KM, Zhang Z, Dong J, Li J, et al. A Comprehensive Technological Survey on the Dependable Self-Management CPS: From Self-Adaptive Architecture to Self-Management Strategies. Sensors (Basel) 2019; 19(5): 1033.