Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P7137-cd

A systematic review of risk metrics for AI and autonomous systems

Ingrid Bouwer Utne

Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Norway. E-mail: ingrid.b.utne@ntnu.no

Supervisory risk control (SRC) is a concept for risk-aware operational decision-making, enhancing the safety and intelligence of autonomous systems. Autonomous systems may support and exceed human performance, but new types of risks are introduced, for example, related to mission complexity and challenges with situation awareness. There are many types of autonomous systems, both crewed and uncrewed, operating in low to high degrees of autonomy, and systems may also switch in between these. The foundation of SRC is constituted by risk assessment and control system design, as well as artificial intelligence (AI). One or more risk models are integrated with the mission planning and/or guidance layer of the autonomous control system. A challenge is, however, how to measure the risk in a way that represents both safe systems and operations, and that can be utilized by the control system, e.g., for path planning. Furthermore, the human supervisor also needs information about the risks to support situation awareness. Hence, risk metrics that sufficiently integrate spatial and temporal information, evaluate "instantaneous" and "long-term" risk, as well as the consider the effect of uncertainty are needed. Therefore, the objective of this paper is to provide an overview of existing metrics for measuring risk and evaluate their usefulness for autonomous systems operating in unstructured environments. The paper also suggests potential directions for further research and development in the area.

Keywords: Risk metrics, Autonomous systems, Robotics, Artificial intelligence, Supervisory risk control, Decision-support.

1. Introduction

Autonomy is increasingly being implemented as a functionality in land-based and maritime transportation, in aerospace, robotics, and in the marine industry. There are both crewed and uncrewed autonomous systems operating in a high to low degree of autonomy (DoA), impacting mission risks. Systems may also vary between different DoA, i.e., dynamic autonomy, which requires shared control between the system and the human. For systems in low DoA, capabilities such as situation awareness (SA), mission planning and replanning may be limited, whereas the opposite is the case for systems in high DoA, although human supervisors will still be "in the loop", e.g., in control centers. Research on human-autonomy collaboration has increased recently (e.g., Veitch & Alsos, 2022). Systems with dynamic autonomy must have risk-based decision support that give the human supervisors sufficient time and information to be able to take over control in emergencies (Hogenboom et al., 2020).

Successful missions with highly autonomous systems in unstructured environments, such as the ocean, require improved safety, intelligence, and operational capabilities, supported by supervisory risk control (SRC) (Utne et al., 2020). SRC is a novel framework that enables an autonomous system to make risk-informed decisions. Advanced online risk assessment and monitoring capabilities are implemented into control algorithms. The term "risk" is often mentioned in current works on robotics and autonomy, but hardly use a systematic and holistic hazard identification, risk assessment and online risk models to provide risk-informed information to the control algorithms in operation, such as the SRC.

A challenge with integrating systematic risk assessment into robotic decision – making, such as for SRC, is related to available risk metrics. A risk metric can be defined as a "quantity – and a measurement procedure and a method for determining the quantity – that provide information about the level of risk related to the study object in a specified future context"

(Rausand & Haugen, 2020). Hence, the choice of risk metric is crucial as it impacts the information communicated from risk analysis (Edwin et al., 2016). Generally, risk is measured as the average or expected loss over a time period, such as a year. The reason is that risk analysis often is applied in the system design phase. How to measure risk in operation, in a shorter time frame, however, and determine when the system performance drops below the acceptable threshold or boundaries, remains a challenge.

Johansen & Rausand (2014) present an overview several risk metrics, but these are not feasible for operational decision-making (Yang & Haugen, 2015). In SRC, risk metrics need to be transformed into quantitative criteria and constraints for control systems to make decisions. Existing "risk metrics" in robotics typically use the expected cost and worst-case metrics (Majumdar & Pavone, 2020). Measuring the risk level of systems and operations, however, is challenging to capture by a single number as there is a need to consider a broader risk spectrum and potentially also ethical decision dilemmas.

Therefore, the objective of this paper is to provide an overview of existing metrics for measuring risk and evaluating their usefulness for autonomous systems and operations. The paper also suggests potential directions for further research and development in the area. To the author´s knowledge this is the first paper attempting to "bridge" risk metrics and risk characterizations from the risk science domain with robotics, with a particular focus on path planning of relevance for autonomy. Artificial intelligence (AI) in this context is related to SRC, i.e., intersecting with high level mission planning for an autonomous system.

The paper is structured as follows: Section 2 presents the SRC concept, Section 3 gives an overview of typical risk metrics related to risk assessments and path planning (robotics). Section 4 presents an evaluation of these metrics, Section 5 discusses the implications and future research needs, whereas Section 6 states the conclusions.

2. Autonomous systems and supervisory risk control

Autonomous systems contain sensors which provide information for sensing, perception, modelling and decision making. Such data is used for SRC (Utne et al., 2020).

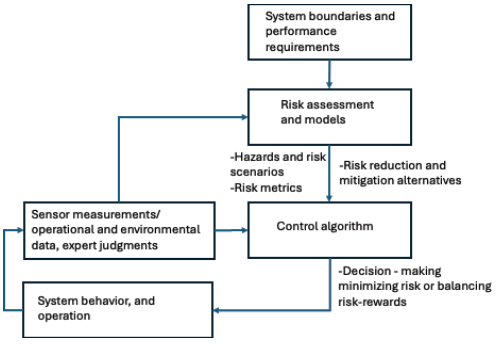

Figure 1. Overview of the SRC concept. See Utne et al. (2025) for more details.

A main difference between the SRC framework, shown in Figure 1, and existing approaches, e.g., for autonomous vehicles, is that the latter often are focused on collision risk and motion models (Katrakazas et al., 2019). The review by Raveendran et al,. (2022) mostly addresses the processing industry, and none include online sensor data and modelling for autonomous systems explicitly. Vagale et al. (2021) reviews collision risk algorithms for path planning of autonomous surface vessels (ASV) and states that collision risk assessment typically is based on only one or two "risk factors". To achieve safe and intelligent systems, however, the risk spectrum must be covered through more extensive risk analysis and modeling, which is included in SRC (Utne et al., 2020).

Thieme et al. (2021) suggest that a control system may use risk information: (i) directly from a risk model (e.g., a Bayesian network); (ii) as decision or optimization criteria; (iii) as a constraint or for modifying a constraint in the control algorithm; or (iv) in risk maps for path planning. (e.g., graph search, etc). Currently, some of these alternatives have been implemented and tested in control system (Utne et al., 2025). These studies revealed that that the traditional risk metrics are challenging to implement, since most were developed for risk measurements in the system design phase, and not for supporting operational decisions.

3. Measuring risk

3.1. *Expressing risk*

There are different definitions of risk, but commonly a risk triplet is used (Kaplan & Garrick, 1981):

$$R = \langle s, p, c \rangle. \quad (1)$$

where s is scenario, p and c are the probability and consequence of that scenario respectively. Generally, the main focus is on risk to people, i.e., fatalities and injuries, but risk to the environment and material assets are also considered. Different risk metrics have different properties, and this may lead to different risk estimations and needs for risk reduction (Haugen & Kristiansen, 2023).

3.2. *Risk metrics for autonomous systems*

Average values over time may be representative for long-term operations but may not be relevant for autonomous systems making decisions within seconds to minutes. The expected cost and worst-case metrics correspond to a risk neutral perspective (former) and a risk conservative perspective (latter). Such perspectives are related to the willingness by humans for risk acceptance depending on potential benefits, controllability, and consequences (Haugen & Kristiansen, 2023).

In AI, utility functions are applied and these may include risk-reward aspects (Russel & Norvig, 2014). A utility function transforms costs into real values, or "utilities", e.g., the maximizing expected utility. Benrabah et al. (2024) review traversability risk assessments for autonomous ground vehicles and classify navigation algorithms into sensor-based and map-based based on a characterization of risk. The sensor-based approaches follow the obstacle boundary, and then the risk is most often included as the minimum distance to the obstacle. The map-based approaches use an environmental map as input, developed from vehicle and terrain data. Generally, the map characterizes risk as the probability of occupancy, probability of traversability, slope, curvature and roughness, object density, elevation, speed, gaussian distribution, deformation of wheels, and cost. Occupancy grids are specifically focused on collision risk, and as an example, the risk of crossing a path can be formulated as:

$$R(\mathcal{P}_{[0,i]}) = 1 - \prod_{j=0}^{k}(1 - P_j) \quad (2)$$

where P_j is the probability of occupation of the jth cell. Benrabah et al. (2024) present several variants of similar probabilistic definitions.

Lefebvre et al. (2016) integrates collision risk in path planning by defining the consequence as a constant cost C_{C_o}, meaning that only the first collision is the focus. The risk associated with path p_{s_g} corresponds to the product of the cost C_{C_o} with the probability that; either (i) the collision occurs during transition from the first state $p_{s_g}(1)$ to the second state $p_{s_g}(2)$ of the path, or (ii) that a collision occurs during the transition from state $p_{s_g}(i)$ to $p_{s_g}(i+1)$ where $2 \leq i < \left|p_{s_g}\right| - 1$ and has not occurred during previous transitions. A conservative approach would be to calculate the cumulative risk along the path.

According to Vagale et al. (2021), the most common risk metrics for ASVs are time to closest point of approach (TCPA) and distance to CPA (DCPA), but also distance of the last-minute avoidance, distance to target vessel, relative bearing, safe passage circle etc. are mentioned.

3.3. *Existing SRC studies and risk metrics*

Bremnes et al. (2020) calculates risk by defining a risk index:

$$ri = \frac{E[Risk|\sigma, ev] - E[Risk]_{bc}}{E[Risk]_{wc} - E[Risk]_{bc}} \quad (3)$$

where $E[Risk|\sigma, ev]$ is the expected risk of loss of an autonomous underwater vehicle (AUV) given evidence ev of the observable variables (in a Bayesian network) and the decision d. $E[Risk]_{wc}$ and $E[Risk]_{bc}$ are the worst and best case expected loss, respectively. The SRC then balances mission utility and risk, using the maximum expected utility (MEU) principle subject to a risk bound. Maidana et al. (2023) calculates the risk of an ASV using the cumulative risk along the path but also includes risk acceptance thresholds that needs to be satisfied. Johansen et al. (2023) uses cost optimization for SRC where the probabilities are calculated in an online risk model with the expected cost of the consequences for an autonomous cargo ship:

$$R(d) = \Pr(severe)\, C_{severe} + \Pr(significant)\, C_{significant} + \Pr(minor)\, C_{minor} + \Pr(none)\, C_{none} \quad (4)$$

Blindheim et al. (2022) uses safety inequalities based on results from a hazard analysis, which are transformed into risk cost terms in a model predictive control algorithm contributing to accumulated system risk levels at any point in time. Rothmund et al. (2023) suggest a heuristic policy to evaluate three decision action

strategies for an industrial drone inspecting confined areas. Each strategy has a consequence, i.e., a cost, if the goal of the task is not achieved. Bremnes et al. (2025) use a risk model for developing risk maps for path planning for an AUV. Here the risk metric is the total mean loss

$$R_{TML}(x,t) \triangleq \sum_{i=1}^{N} \mathbb{E}[\mathcal{L}_i(x,t)] \quad (5)$$

which expresses the total expected loss (probability times consequence) at state x at time t per unit length, summed over all identified hazardous events. This metric is risk-neutral, as it is equally sensitive to low-probability high-consequence events as to high-probability low-consequence events. If more risk-adversement is desirable, an alternative may be the conditional value at risk (CVAR) but this has not been implemented.

The existing SRC studies have mainly focused on using cost as a risk metric, without investigating its feasibility as a risk metric, nor on their advantages and disadvantages.

4. Criteria for evaluation of the risk metrics

Table 1 presents an overview of existing risk metrics used in risk assessments and risk monitoring for human fatalities, injuries, damage to the environment, and material assets. The table also includes typical risk metrics used in robotics. The choice of risk metric impacts the estimation of risk and should be considered thoroughly with respect to effects on the decision-making. It is desirable to avoid using risk metrics that underestimate the risk, but being overly conservative may also hinder efficient missions for autonomous systems. To evaluate the risk metrics with respect to feasibility in operation, eight criteria (1-8) have been identified, based on Johansen & Rausand (2014), and the simulation – based testing and experiments with SRC (Utne et al., 2025; Bremnes et al., 2025):

1 Allows for including spatial (e.g., extent) and temporal (e.g., duration) aspects, which are important to be able to reflect dynamic properties of a system in operation.
2 Must allow for continuous or frequent updates based on new data/information.
3 Must be able to distinguish between, or aggregate, different hazardous events.
4 Must be valid, i.e., that the risk metric ‘‘measures’’ risk (cf. Eq. 1), i.e., the metric has a clear verbal and mathematical definition.
5 Must provide comparability to other decision-aspects, to allow for trade-offs and optimization.
6 Must be transparent related to value judgments in terms of aversion factors or monetary evaluation of different consequence dimensions.
7 Must have a clear location in the bow-tie (proactive vs. reactive).

5. Discussion

5.1. *Feasibility of existing risk metrics*

According to Johansen & Rausand (2012), the relevant above-mentioned metrics for the operational phase are IRPA, FAR, IR, PER, PEF. Still, IRPA, PLL and FAR focuses on averages over a year, which is insufficient for operational (real-time) decision-making. Still, they have a clear definition and purpose.

Some metrics include risk aversion factors, such as total risk and CVAR. SRI includes location and affected people, which is desirable in an operational context. Still, none of the metrics include both a temporal and spatial information, except expected loss, in the manner used by Bremnes et al. (2025). PEF may be seen as a way of including injuries in addition to fatalities in the risk calculations (or other consequence aspects). Recovery time may be associated with resilience of systems, which is a desirable characteristic of autonomous systems.

Some of the risk measures or metrics are not defined quantitatively, which makes them challenging to use for SRC and robotic decision-making. The risk matrix has some limitations in this respect but might potentially be used as a “lookup” table for simple rule based systems. Expected losses/costs is the most commonly used way of translating risk into a quantitative measure. The operational risk metrics defined by Yang & Haugen (2015;16) are interesting, but they have not been fully quantified in practice. Risk indicators may be a desirable way to quantify different hazards or risk influencing factors (RIFs), that may be updated regularly in operation. A challenge, however, may be to aggregate the information from indicators into robotic decision-making (SRC). For human operators, this could be done through risk visualization tools. In general, the current risk metrics used in robotics gives a limited view on risk, since only one or a few RIFs are included.

Table 1. Risk metrics, representations, references, and evaluation criteria (Haugen & Kristiansen, 2023; Johansen & Rausand, 2012; Bremnes et al., 2025; Rockafellar & Uryasev, 2000; Leveson and Thomas, 2018; Øien et al., 2001; Øien et al., 2011; Yang and Haugen, 2015; 2016; Mehlhorn et al., 2023; Rødseth et al., 2022; Edwin et al., 2016).

Risk metric	**Representation**	**Description**	**Evaluation criteria**							
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Risk matrix	Matrix defined by a pre-defined set of categories for probability and consequence	A risk matrix is a widely used semi-quantitative way of measuring risk.			+	+		+	+	
Individual risk (IRPA)	$IRPA = \sum_{i=1}^{n} \lambda_i \cdot \Pr(E_i) \cdot \Pr(F\|E_i)$, λ_i is probability of a hazardous event i, E_i is exposure to i, F is fatality.	The probability that an average person is killed during a period of one year due to exposure to the hazardous event i.			+	+	+			+
Localized individual risk (LIRA)	$LIRA\ (x,y) = \sum_{i=1}^{n} \lambda_i \cdot \Pr(F\|HE_i)$, λ_i is the probability of a hazardous event HE_i, F is fatality at location (x,y).	The probability that an average person present at a location, is killed during a year due to a hazardous event HE_i.	(+)		+	+	+			+
Potential loss of life (PLL)	$PLL = n \cdot IRPA$	The expected number of fatalities in a specific population n per year.				+	+			+
Fatal accident rate (FAR)	$FAR = \frac{PLL}{AH} \cdot 10^8$, where AH is the accumulated no. of hours the specific population is exposed to risk.	The expected no. fatalities in a population per 100 mill. hours exposure (~1000 people work 2000 hours per year for 50 years).	(+)			+	+			+
Weighted risk integral (RICOMAH)	$RI_{COMAH} = \sum_{i=1}^{n_{max}} f(n) n^k$, where f(n) is frequency of accidents with exactly n fatalities per year and k is a risk aversion factor.	The expected number of fatalities corrected for risk aversion with respect to a high number of fatalities.				+	+		+	+
Scaled risk integral (SRI)	$SRI = \frac{P \cdot IRPA \cdot T}{A}$, where P is population factor ($\frac{n+n^2}{2}$), T is share of occupancy, A is area, and n is number of persons in the area.	The group risk per surface area (A) per year.	(+)			+	(+)			+
Total risk (TR)	$TR = PLL + a\sigma$, where a is risk aversion factor and σ is standard deviation	The expected number fatalities corrected for risk aversion for extreme events.				+	+		+	+
Potential equivalent fatality (PEF)	$PEF = PLL + 0.1M + 0.01N$	Human injuries and fatalities per year can also be combined into the potential equivalent fatality (PEF) measure, where a major (M) and a minor injury (N) are expressed as $1/10$ and $1/100$ fatality.				+	+		(+)	+
Potential environmental risk (PER)	$f = \lambda \cdot \Pr(E\|S) \cdot \Pr(C\|E)$, where λ is the frequency of spill S, E is exposure to spill for area A and C is defined consequence.	The frequency of a defined consequence category for a certain organism, population, habitat or ecosystem within an area.	(+)	(+)		+	+			+
Recovery time (RT)	$f = \lambda \cdot \Pr(Dd > R$, where λ is frequency of spill, Dd is damage duration and RT is required recovery time.	The probability per year of having an accident that exceeds the time needed by the ecosystem to recover from damage	(+)			+	+			+
Expected economic loss (EL)	$E(D) = \sum_{i=1}^{n} x_{AS_i} \cdot \Pr(AS_i)$, where x_{AS_i} is economic loss related to accident scenario AS_i.	Bremnes et al (2025) extended this to $R_{TML}(x,t) \triangleq \sum_{i=1}^{N} \mathbb{E}[L_i(x,t)]$, which includes state x at time t per unit length (Sect. 3.3).	+	+		+	+	+		+
Monetary collective risk (MCR)	$R_m = \sum_{i=1}^{n} p_i C_i \varphi(C_i) \omega_i$, where ω_i is willingness to pay for averting consequence C_i, C_i is consequence dimension i, p_i is probability of consequence C_i and φ_i is risk aversion factor for C_i.	The expected total monetary loss per year, aggregated and weighted across different damage dimensions (such as fatalities, injuries, disruption of service).				+	+	+	+	+

Table 1 (cont.). Risk metrics, representations, references, and evaluation criteria (Haugen & Kristiansen, 2023; Johansen & Rausand, 2012; Bremnes et al., 2025; Rockafellar & Uryasev, 2000; Leveson and Thomas, 2018; Øien et al., 2001; Øien et al., 2011; Yang and Haugen, 2015; 2016; Mehlhorn et al., 2023; Rødseth et al., 2022; Edwin et al., 2016).

Risk metric	Representation	Description	Evaluation criteria							
			1	2	3	4	5	6	7	8
Conditional expected damage (CED)	$l_1 = E[X\|X \le x_1]$ $l_2 = E[X\|x_1 < X \le x_2]$ $l_3 = E[X\|X > x_2]$, where X is damage variable, x_1 is low level severity and x_2 is high severity.	The conditional expected value given that the consequence severity is above a specified level.		+		+	+		+	+
Conditional value at risk (CVAR)	$R_{CVaR}(x,t) \triangleq \sum_{i=1}^{N} CVaR_{\alpha}[\mathcal{L}_i(x,t)]$ where $\mathcal{L}_i(x,t)$ is the loss at state x at time t per unit length, and VaR_{α} is the cost corresponding to boundary of the $(1-\alpha)$-quantile of the probability distribution.	A risk metric which may be used for analyzing and optimizing the risk of stock portfolios in finance, sensitive to more severe consequences.	(+)	+		+	+	+	(+)	
Safety constraint	"If a hazard occurs, then what needs to be done to prevent or minimize a loss."	May be defined by intervals or single metrics. An example is closest point of approach (CPA) and CED.	(+)	+	+	(+)	+	+		(+)
Risk indicator	"A measurable/operational definition of a risk influencing factor." (Safety indicators are not necessarily quantitative and linked to risk models.)	The risk level in operation may be measured by using risk indicators, which are updated regularly.		+	+	+	+	+		
Activity performance risk	A selection of critical safety parameters influencing risk.	An expression of risk level associated with performing a specific activity.		+	+	(+)	(+)	(+)		(+)
Activity consequence risk	Frequency of occurrence of a specific catastrophic failure scenario	An expression of the effect that completing an activity will have on the risk level after the activity has been completed.			+	(+)	+	+		+
Time-dependent action risk	Indicators derived from operating parameters against operating limits.	An expression of short-term risk variation while performing one or several activities.	(+)	+	+	(+)	(+)	+		(+)
Period risk	Activity performance + interactions.	An expression of risk for a plant or facility over a (normally short) time period.			+	(+)		(+)		
Operational design domain (ODD)/Operational envelope (OE)	The ODD defines the limits within which the driving automation system is designed to operate, and as such, will only operate when the parameters described within the ODD are satisfied. The OE may be defined as "The specific conditions and scenarios under which a given autonomous ship system is designed to function."	The ODD may be specified by a taxonomy that includes scenery (e.g., zones), environmental conditions (e.g., weather) and dynamic elements (e.g., traffic). The OE must cover voyage and operation phases.	(+)	+		(+)	+	(+)		
Risk visualization	An example is the «risk barometer", which translates risk expressed by means of metrics above into a relative percentage value indicated by a "needle". Another example are the risk maps used for path planning.	Both the barometer and risk maps involve defined risk levels expressed through colors reflecting "acceptability" or different aspects of risk. Trends may also be visualized.		+	+	(+)	(+)	+		(+)

Risk maps (see Bremnes et al., 2025) have a potential to include a lot more risk information than the traditional "risk maps" in robotics.

Please note that some existing risk metrics have not been included in the paper due to a high uncertainty in relevance. FN/FE diagrams have been excluded, as these represent societal risk and may give inconsistent evaluations (Johansen & Rausand, 2014). Individual risk of dangerous dose is an environmental risk metric which is focused on toxic chemicals (Johansen & Rausand, 2012). Frequency of hazardous events, such as Loss of main safety function in oil and gas and Core damage frequency in nuclear power production, do not include the consequences specifically (Johansen & Rausand, 2014). QALY/DALY are metrics focused on diseases and life expectancy (Haugen & Kristiansen, 2023). Odds ratios are discussed for airspace (Bati et al., 2021).

Hence, there is currently no risk metric that captures all the criteria (1-8) in a desirable manner, i.e., that is particularly advantageous for measuring online risk in operation, neither for autonomous systems´ decision-making nor for human operators.

5.2. *Future research needs*

In the evaluation of the metrics in the paper, all criteria are weighted equally. It may be worthwhile investigating, whether some are more important than others. Furthermore, since no metric fulfils all criteria, it is necessary to investigate different combinations of metrics, i.e., related to operational and instantaneous decision-making (Yang & Haugen, 2015) and "long-term" risk, e.g., further exploring the approach in (Lefebvre et al., 2016). Furthermore, it is may be feasible to investigate how risk metrics for different DoA in terms of an ODD or operational envelope for an autonomous system combined with risk maps may be utilized and updated regularly during operation representing a system´s current capabilities, uncertainty, and system integrity. A challenge with AI systems, when using machine learning is lack of transparency and interpretability, as well as uncertainty. Hence, development and choice of risk metric becomes increasingly important for such systems. For a human operator individual and aggregated risk maps can be presented through the human-machine-interface (HMI) to represent different RIFs, potential hazardous events, and the total risk level. Here, the spatial and temporal aspects of the risk models are important. How to aggregate different risks (and rewards) into the total risk level for a safe and efficient HMI is an unsolved issue, but risk indicators related to the risk models, reflecting changes in the risk model and the scenarios, could be used as a foundation. These could, for example, also be connected to the expected response time from the human supervisor in case intervention is needed.

6. Conclusions

This paper attempts to bridge risk metrics and risk characterizations in the risk science domain, robotics, and path planning. The paper presents typical risk metrics used in conventional risk assessments, as well as in path planning for autonomous systems. The overall challenge with the "traditional" risk metrics is that measuring changes in operation (seconds-days) is difficult since several measures averages over, e.g., a year. The main challenge with the metrics used in robotics is that these do not use a systematic risk assessment as a basis for deriving RIFs that should be quantified and integrated for risk-aware control algorithms to cover a broader risk spectrum. An important issue is also that the choice of risk metric may impact safety, as all risk metrics have different advantages and limitations.

The challenge of how to measure risk in operation of autonomous systems has so far only been addressed to a limited extent. Hence, further research must particulary address the combined used of risk metrics ("instantaneous" and "long-term", uncertainties, the development of new metrics, and considering the risk information needed for both the autonomous system and the human operator to ensure sufficient situation awareness and safe decision-making.

Acknowledgement

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held

responsible for them. This work is supported by ERC grant (BREACH, ID: 101142277, DOI 10.3030/101142277.)

References

Bati, F., F. Lee, R. Bollschweiler, L. Withington, N. Khoorrami, I. Filipppov and R. Adhikari (2021). Risk metrics to measure safety performance of the national airspace system: Implementation using machine learning. *IEEE/AIAA 40th Digital Avionics Systems Conference (DASC),* USA.

Benrabah, M., CO. Mousse, E. Randriamiarintsoa, R. Chapuis and R. Aufrère (2024). A Review on Traversability Risk Assessments for Autonomous Ground Vehicles: Methods and Metrics. *Sensors*, 24, 1909.

Blindheim, S., TA. Johansen and IB. Utne (2023). Risk-based supervisory control for autonomous ship navigation, *J. Mar. Sc. & Tech.*, 28, 624–648.

Bremnes, JE., CA. Thieme, AJ. Sørensen, IB. Utne and P. Norgren (2020). A Bayesian approach to supervisory risk control of AUVs applied to under-ice operations, *Mar. Tech Soc. J*., 54(4), 16-39.

Bremnes, JE., IB. Utne, TR. Krogstad and AJ. Sørensen (2025). Holistic risk modeling and path planning for marine robotics, *IEEE Journal of Oceanic Engineering*, 50 (1).

Edwin, NJ., N. Paltrinieri and T. Østerlie (2016). Ch. 13. Risk Metrics and Dynamic Risk Visualization, In: *Dynamic Risk Analysis in the Chemical and Petroleum Industry*, Elsevier.

Haugen, S. and Kristiansen, S. (2023). *Maritime transport. Safety management and risk analysis*. 2nd ed. Routhledge, London.

Hogenboom S., B. Rokseth, JE. Vinnem and IB. Utne (2020). Human reliability and the impact of control function allocation in the design of dynamic positioning systems. *Rel. Eng. Sys. Saf*, 194, 106340.

Johansen, I. L. and M. Rausand (2012). Risk metrics: Interpretation and choice, *in Proc. IEEE Int. Conf. Ind. Eng. Eng. Mana*g., Hong Kong, 1914– 1918.

Johansen, IL. and Rausand, M. (2014). Foundations and choice of risk metrics. *Safety Science,* 62, 386–399.

Johansen, T, S. Blindheim, TR. Torben, IB. Utne, TA. Johansen and AJ. Sørensen (2023). Development and testing of a risk-based control system for autonomous ships, *Rel. Eng. Sys. Saf*., 234, 109195.

Kaplan S. and BJ. Garrick (1981). On the Quantitative Definition of Risk. *Risk Analysis*, 1(1), 11-27.

Katrakazas C., M. Quddus, W-H Chen (2019). A new integrated collision risk assessment methodology for autonomous vehicles. *Accident Analysis and Prevention*, 127, 61-79.

Lefebvre, N., I. Schjølberg and IB. Utne (2016). Integration of risk in hierarchical path planning of underwater vehicles, *IFAC-PapersOnLine, 49(23)*, 226-231.

Leveson, N. and JT. Thomas (2018). *STPA Handbook.* MIT.

Maidana, RG., SD. Kristensen, IB. Utne and AJ. Sørensen (2023). Risk-based path planning for preventing collisions and groundings of maritime autonomous surface ships, *Oc. Eng*., 290, 116417.

Majumdar, A. and M. Pavone (2020). How should a robot assess risk? Towards an axiomatic theory of risk in robotics, in *Proc. 18th Int. Symp. Robot. Res*., Puerto Varas, 75–84.

Mehlhorn, MA., A. Richter, YAW. Shardt (2023). Ruling the Operational Boundaries: A Survey on Operational Design Domains of Autonomous Driving Systems, *IFAC OnLine* 56-2 , 2202–2213.

Rausand, M. and S. Haugen (2020). Risk assessment. Theory, methods and applications. Wiley, US.

Raveendran A., VR. Renjith and G. Madhu (2022). A comprehensive review on dynamic risk analysis methodologies. *Journal of Loss Prevention in the Process Industries*, 76, 104734.

Rockafellar, R. and S. Uryasev (2000). Optimization of conditional value-at risk, *J. Risk*, 2, 21–42.

Rothmund, SV., CA. Thieme, IB. Utne and TA. Johansen (2023). Bayesian Approach to Risk-Based Autonomy, with Applications to Contact-Based Drone Inspections, *J. Int. & Rob. Sys*, 109, 31.

Russel, S and P. Norvig (2014). *Artificial intelligence. A modern approach*, Upper Saddle River, NJ, Prentice Hall Series in Artificial Intelligence 3rd ed.

Rødseth, ØJ., LAL. Wennersberg, H. Nordahl (2022). Towards approval of autonomous ship systems by their operational Envelope. *J. Mar. Sc. Tech.*, 27, 67–76

Thieme, CA., B. Rokseth, I.B. Utne (2021). Risk-informed control systems for improved operational performance and decision-making, *Proc ImechE Part O: J. Risk and Reliability*, Special issue: Autonomous Systems Safety, Reliability, and Security, 1-23.

Utne, IB, B. Rokseth and A. J. Sørensen, J. E. Vinnem (2020). Towards supervisory risk control of autonomous ships, *Rel. E. S. Saf.,* 196, 12020, 06757.

Utne, lB., TA. Johansen and Sørensen, AJ. (2025). Towards Risk-based Rationality in Autonomous Systems and Operations. In *Proc. RAMS Conference*, Florida US.

Vagale, A., R. Bye, R. Oucheikh, O. L. Osen and TI. Fossen (2021). Path planning and collision avoidance for autonomous surface vehicles II: a comparative study of algorithms, *J. Ma. Sci. Tech*., 26, 1307-1323.

Veitch E. and OA. Alsos (2022). A systematic review of human-AI interaction in autonomous ship systems. *Safety Science*, 152, 105778.

Yang, X. and S. Haugen (2015). Classification of risk to support decision-making in hazardous processes, *Safety Science,* 80, p. 115–126, 2015.

Yang, X. and S. Haugen (2016). Risk information for operational decision-making in the offshore oil and gas industry. Safety Science, 86, 98-109.

Øien, K., 2001. Risk indicators as a tool for risk control. *Rel. Eng. Sys. Saf*., 74, 129–145.

Øien, K., Utne, IB., Herrera, I. 2011. Building safety indicators: Part 1 – Theoretical foundation. *Safety Science*, 49 (148-161.