Reliable Unmanned Autonomous Systems: Conceptual Framework for Warning Identification during Remote Operations

1st Rialda Spahić
Dept. of Engineering Cybernetics
Norwegian University of Science and Technology
Trondheim, Norway
rialda.spahic@ntnu.no

2nd Vidar Hepsø
Dept. of Geoscience and Petroleum
Norwegian University of Science and Technology
Trondheim, Norway
vidar.hepso@ntnu.no

3rd Mary Ann Lundteigen
Dept. of Engineering Cybernetics
Norwegian University of Science and Technology
Trondheim, Norway
mary.a.lundteigen@ntnu.no

***Abstract*—In the offshore industry, unmanned autonomous systems are expected to have a permanent role in future operations. During offshore operations, the unmanned autonomous system needs definite instructions on evaluating the gathered data to make decisions and react in real-time when the situation requires it. We rely on video surveillance and sensor measurements to recognize early warning signals of a failing asset during the autonomous operation. Missing out on the warning signals can lead to a catastrophic impact on the environment and a significant financial loss. This research is helping to solve the issue of trustworthiness of the algorithms that enable autonomy by capturing the rising risks when machine learning unintentionally fails. Previous studies demonstrate that understanding machine learning algorithms, finding patterns in anomalies, and calibrating trust can promote the system's reliability. Existing approaches focus on improving the machine learning algorithms and understanding the shortcomings in the data collection. However, recollecting the data is often an expensive and extensive task. By transferring knowledge from multiple disciplines, diverse approaches will be observed to capture the risk and calibrate the trust in autonomous systems. This research proposes a conceptual framework that captures the known risks and creates a safety net around the autonomy-enabling algorithms to improve the reliability of the autonomous operations.**

***Index Terms*—autonomous systems, machine learning, reliability engineering, risk assessment, calibrated trust**

I. Introduction

The advancements in technology are changing the way the industry handles risk. What used to be a tedious or dangerous job for a human can be replaced by an unmanned autonomous system (UAS). This replacement can enhance safety, work efficiency, and knowledge of the operating environment. As a type of artificial intelligence [1], machine learning (ML) is at the forefront of research in the context of reliable UAS. Autonomy is "an unmanned system's own ability of integrated sensing, perceiving, analyzing, communicating, planning, decision-making, and acting/executing to achieve its goals" [2]. Recent research on autonomous systems identifies common challenges within risk and trust of ML algorithms that enable autonomy [3], [4]. Formal definitions of technical tests and evaluation of UAS [5] highlight challenges of lacking the quantitative definitions of emergent behavior, human trust, reliability and resilience [3]. However, to ensure the UAS's ability to act and make decisions to achieve the mission's goal, it is critical to explore the concept of calibrated trust [6]. Calibrated trust is the process of adjusting the trust level of human operators with the actual reliability of a system [6], - or trusting the machine will do as intended within a specific environment [7].

ML allows computing systems to learn how to do tasks from significant amounts of data, rather than being programmed (human instructed) [1]. Therefore, there is a rising need to understand how ML capabilities can be integrated into existing systems engineering, and design processes [3]. The performance of ML algorithms can measure the majority of UAS's capabilities, inevitably measuring the system's reliability. However, the software and system reliability engineering for UAS incorporating ML is not a trivial task. ML integration is experiencing significant limitations, including black box algorithms or algorithm explainability [8], scalability, and limited structural approach to problem-solving. The data, often impacted by biases, is another limitation related to ML. Bad quality data can lead the ML algorithms to result in poor predictions or decisions, and eventually, unintended harm [9].

During offshore operations, the UAS relies on integrated sensors and video input for surveillance, intervention, and inspection of the assets and the environment. The role of the UAS is to recognize warning signals from the environment or the inspected asset, trigger warning signals, and report them to the offshore control center or operator control rooms in real-time. Unintended ML outcomes can significantly impact the environment, the asset, and the UAS itself. The environmental disruptions can stay unnoticed and develop to critical states, such as disruptive water states or chemical leaks. Similarly, unobserved corrosion, chemical leaks, material degradation, cracks, misplaced objects, and biological growth on assets are just a few examples of the potential issues. This problem

This research is a part of BRU21 – NTNU Research and Innovation Program on Digital and Automation Solutions for the Oil and Gas Industry (www.ntnu.edu/bru21) and supported by Equinor.

can lead to a catastrophic impact on the environment and significant financial loss for the industry. Knowing how to respond and prepare the data for anticipated insights is a challenge in dynamic operations. The industry needs more knowledge on reliable, and time-efficient UAS operations [10].

The contribution of this paper is a Warning Identification Framework (WIF) for UAS incorporating ML. The WIF incorporates managing resilience, ensuring the system's ability to plan, prepare and react to the potential occurrence of unwanted and disruptive events. While designing this framework we consolidate knowledge on reliability and resilience engineering, risk assessment, and human-machine teaming approach to UAS. In this paper, we:

1) Provide a multidisciplinary approach to the safety concerns of current systems incorporating ML through the lenses of risk assessment's future.
2) Propose a global framework based on a shared understanding of gaps in ML of a particular application instead of solutions based on specific ML algorithm enhancements or global change of data gathering processes.

Section II gives an overview of the background and related work. Section III details the reliability and resilience engineering, risk assessment, and human-machine teaming theories and concerns. Section IV formalizes the concepts from Section III and illustrates the WIF's role in mitigating risks. Finally, Section V concludes the paper and addresses future work.

II. Background and Related Work

A. Trust Calibration

Recent research shows potential in alleviating risks, enhancing reliability, and influencing trust in autonomous systems. Reliability is an ability to perform as required, without failure, for a given time interval, under given conditions (IEC 192-01-24) [11]. The reliability of an autonomous system directly impacts trust. However, over-trust and under-trust often occur in highly dynamic environments and can pose serious safety and efficiency concerns [6]. Over-trust in the system implies that the human operator overestimated the reliability of a system. Under-trust in the system implies that the human operator estimates that the system should not be trusted with a given task. Okamura et al. [6] describe the trust calibration in autonomous systems in a dynamic environment as an essential process for successful collaboration between humans and systems. Trust calibration incorporates system reliability and continuous system transparency. Okamura et al. [6] argue that trust is a latent construct and therefore challenging to measure. The authors [6] observe human behavior to determine the trust calibration status. They experimented with a drone simulator and observed seventy participants who performed inspection tasks manually or relied on the inspection by an autonomous drone. In the experiment, the participants observed the changing weather conditions in the drone simulator. The participants were required to actively make decisions whether they trust or rely on the autonomous drone to perform inspection tasks within the environmental conditions presented on the simulator. The experiment's goal was to capture the under-trust and over-trust of the participants in the autonomous drone operations. The experiment demonstrated successful detection of miscalibration of trust and adjustment of participants' behaviors, showing trust gaps in collaboration between humans and autonomous systems. The results showed that understanding how the system functions and makes decisions is crucial when trusting the autonomous systems.

B. Explainable Machine Learning

ML is taking over many high-stakes decision-making throughout society [8]. The author [8] defines *black box ML algorithms* as either functions that are too complicated for any human to comprehend or as proprietary functions. Past research highlights that developing explainable algorithms will mitigate some of the problems caused by the black box algorithms [8]. Rudin [8] argues that trying to explain black box algorithms rather than developing explainable ones can support a bad practice and therefore cause harm to society.

The author [8] singles out some of the most prominent challenges of black box and explainable algorithms:

1) *Complexity:* There is a belief that black box algorithms result in top predictive performance when compared to the explainable algorithms that are easier to understand. The author claims that when the data is structured and contains meaningful features, complex classifiers (such as neural networks, random forest, boosted decision trees) and more straightforward classifiers (such as logistic regression and decision lists) perform similarly. Complexity does not imply accuracy, which is also valid for computer vision or image processing algorithms that are often particularly complex.
2) *Faithfulness of explainable algorithms:* Explainable ML algorithms provide interpretations that are not faithful to what the black box algorithm computes. The explainable algorithm does not mimic the black box algorithm but instead tries to interpret it as accurately as possible and provide an explainable alternative to the black box algorithm. The difficulty in creating this interpretation can lead to misalignment with the black box algorithm and endanger the trust in the black box algorithm. Rudin [8] proposes calling these interpreted algorithms 'summary statistics', 'summary predictions' or 'trends of the algorithm' to avoid confusion with the belief that the interpretation should mimic the black box algorithm.
3) *Challenge to incorporate risk estimation within black box algorithms:* The database is a definite collection of data or information that the algorithms learn from and train on to make predictions. Black box algorithms are often incompatible with the situation where information outside the database needs to be combined with a risk assessment. Rudin [8] argues that the black box algorithms are challenging to calibrate with additional information on estimated risk manually. Another downside of these

algorithms is that it is not transparent as to what the risk estimation is.

4) *Explainability leading to human error:* Additional explanations to the black box algorithm can lead to complicated decision-making and leave space for other human error.
5) *Hidden patterns in data:* There is a myth that only black box algorithms can uncover hidden patterns in data. This myth can lead to less trust in the performance of explainable algorithms. The author [8] claims that if the pattern were significant enough, it would be possible to obtain it with an explainable algorithm.
6) *Explainability is difficult to design and develop:* Creating explainable algorithms for specific domains often involves constraints on data dimensions, meaning that explainability requires low-dimensional space. It is challenging to troubleshoot the algorithm or agree on the explainable algorithm's reasoning process for a specific domain. The main challenge lies in the difficulty of developing and designing explainable algorithms.

Explainable ML algorithms lead to increased transparency that is crucial in measuring the fairness of the advanced system's decision-making processes. The *fairness* notion tells if the output of a predicting system is fair or discriminating [12]. Fairness is a rising problem due to the predictive system's tendency towards efficiency and sacrificing anomalies as tolerable collateral damage [12].

C. Errors and Biases in Machine Learning

There is a growing worry about the errors of ML in sensitive domains [13]. Pleiss et al. [13] describe cases of ML errors due to biases in data that have directly impacted human lives. The authors examine the cases of ML classification algorithms and frameworks that constrain these algorithms such that no false-positive or false-negative predictions affect any classified group or that there exists fairness in the classified groups. Their study demonstrated unsettling results that any algorithm with one error constraint (i.e., equal false-negatives across groups) is almost equal to randomizing the percentage of predictions for an existing classifier.

Knowing when to react is critical during remote operations. A timely reaction can prevent accidents saving the environmental impact and significant amounts of money. Galaz et al. [4] provide recent research of machine intelligence risks that include algorithmic bias and harms, unequal access and benefits, cascading failures and disruptions, mis- and disinformation, and trade-offs between efficiency and resilience. The authors imply that many foreseeable risks can be acted upon proactively. However, they do not propose actions or algorithms to intervene with ML outcomes' foreseeable risks. The authors [4] highlight that these risks are related to algorithmic biases and their allocative harms. The authors group these biases into training data bias, transfer context bias, and interpretation bias:

1) Training data bias is the erroneous data from which machines learn.
2) Transfer context bias occurs when using ML algorithms and dataset created in/for one environment in another.
3) Interpretation bias is a conflict between ML interpreted results and expected or needed results for further functioning of a system.

Suresh et al. [9] discuss important choices generated over extensive data and build a framework for understanding unintended consequences of ML. The authors identify 'biases' as the most common reason from which unwanted ML consequences arise. The bias represents an unintended or even malicious property of the data [9]. The authors [9] curate through recent work of known ML issues and identify six sources of harm that represent a framework for understanding the unintended ML consequences:

1) Historical bias occurs when the machine learns on historical or available data samples that do not reflect an accurate picture of the world.
2) Representation bias occurs when there is an imbalanced representation of all the data samples in the data set.
3) Measurement bias occurs when what we choose to measure does not relate well to the data samples the machine learns on or when the machine learning task is oversimplified.
4) Aggregation bias occurs when using a one-size-fits-all algorithm for cases with different conditional distributions.
5) Evaluation bias occurs when the evaluation or benchmark data for the ML algorithm does not represent the target measurement.
6) Deployment bias occurs when there is a mismatch between the problem an algorithm is intended to solve and how the algorithm is used.

The authors [9] advise tweaking ML algorithms to mitigate aggregation and evaluation biases in data. They indicate that the framework can communicate knowledge on ML outcomes and possibly facilitate productive solutions on dealing with the harmful consequences.

D. Applications

A significant number of applications are developed for autonomous systems incorporating ML for mitigating risks during operations. As a major task in offshore operations, the UAS are increasingly popular to gather information for risk assessment of the assets or the environment. Condition-monitoring data is often used as additional information for evaluating risk [14]. In offshore operations, monitoring of assets can give real-time information on degradation of the asset material, and the condition-monitoring data provides information on individual degradation process [14]. Some of the applications regarding data assessment on degradation processes such as oxidation, corrosion, fatigue, crack growth are [14]–[17]. Improved design and tweaking of ML algorithms and reconsideration of data gathering and pre-processing methods are the most notable research topics for enhancing the reliability of autonomous systems, and understanding error measurements [5], [18]–[22].

Anomaly detection is an essential process for recognizing unexpected events in the data during operations. Liu et al. [23] explore background biases for anomaly detection in surveillance videos. Their study shows that the algorithms are biased to capture a considerable amount of background information as the basis of predictions. The authors [23] argue that background bias is a problem that exists in the majority of the action recognition algorithms, particularly in deep neural networks. They propose a trainable, area-guided framework for the anomaly detection algorithms to recognize anomalous regions and learn the essence of the anomaly instead of simply remembering the background [23]. Related concerns around anomaly detection algorithms are prominent in research, such as trade-offs and analysis of the algorithms [24], bottleneck identification [25], and large-scale anomaly detection in surveillance videos [26].

III. Multidisciplinary View on Reliable Autonomous Systems

A. Risk Assessment

Risk assessment is a discipline that incorporates structured analysis and identification of possible hazards/threats, their causes and consequences, risk description, quantification, and representation of uncertainties [14]. The terms *risk* and *warning* are often used together or interchangeably. According to [1], *risk* is the possibility of something bad happening at some time in the future, a situation that could be dangerous or have a bad result. Moreover, a *warning* is a statement or an event telling somebody that something bad or unpleasant may happen in the future so that they can try to avoid it [1]. Additionally, a warning is a sign that indicates approaching or threatening risk and may require immediate intervention. Therefore, it is crucial to understand a specific environment or assets' potential risks of failure to understand warning signs and act upon them. The risk assessment should provide a coherent increase of the awareness on risk and attention to safety. The fourth industrial revolution, particularly the internet of things, big data, and artificial intelligence that enables autonomy, changes how we design and develop systems and monitor our environment. This complex network of cooperative systems provides opportunities to improve the systems that monitor, intervene, and inspect the environment or the industrial assets to become more efficient, faster, more flexible, and resilient. However, these systems also generate new weaknesses, hazards and create new risk, somewhat due to new and unknown functional dependencies in and among the systems [14]. Scibilia et al. [27] describe the industry perspective on the definition of autonomy and divide autonomy into six levels, from no automation to a fully autonomous system that does not require human interaction. The authors [27] highlight that the fully autonomous system is multidimensional and incorporates autonomy/automation, data deliberation, and risk assessment. Data deliberation signifies the system's capability to continuously gather data from the environment, analyze it, and compare historical data to make predictions. Risk assessment signifies the system's capability to continuously assess the risk and adjust the criticality of the warnings accordingly, deciding the best risk mitigation policy [27].

Naturally, the digital future is shaping the future of risk assessment. According to [14], six underlying factors impact the advancement of risk assessment:

1) Knowledge, information, and data available for analyzing and computing the risk are continuously growing.
2) Modeling capabilities and computational power are continuously advancing, making more accessible simulations and large-scale data analysis.
3) The increasing complexity of the advancing systems made of heterogeneous elements (hardware, software, human) leads to system behaviors challenging to predict or explain.
4) The risk assessment extends to cover managing risk in a systematic way that includes the occurrence of the risk, prevention, mitigation, emergency crisis management, and restoration [14].
5) Recognition that risk varies over time and accordingly, the effectiveness of the mitigation measures changes.
6) Cyber-physical systems require solid frameworks for safety and security assessment.

Zio [14] highlights that description of the risk and future risk assessment is conditioned on available knowledge. However, it is equally important to address the incomplete knowledge or the unknowns within the risk assessment. According to the available knowledge, Flage et al. [28] classify the events in risk assessment to:

1) Unknown-unknown events that are new and unknown to everyone.
2) Unknown-known events that are new to risk analysts but have been recognized by someone else.
3) Known-unknown events with weak background knowledge and justified indications that a new, unknown type of event or scenario could occur in the future.
4) Known-known events that are known to the analysts performing the risk assessment and for which there is existing evidence.

In autonomous systems that incorporate ML, unknown events require novelty detection and anomaly detection approaches. *Novelty detection* is the task of classifying test data that differ in some respect from the data that are available during training [29]. Anomaly detection detects the anomalies unrelated to the training data [30]. Both anomalies and novelties occur rarely and are dealing with unexpected events in the data. We can argue that the most dangerous events are the unknown ones because otherwise, we can take action to prevent them. Accordingly, Flage et al. [28] argue that known-unknown events are representative of *known risks* that become apparent in *new conditions*. However, the unknown-unknown, unknown-known, and known-known events can be associated with negligible probabilities of occurrence.

B. Reliability Engineering for UAS

The system reliability engineering and reliability assessment are practical ways to manage risk and support decision-making for safe, reliable, and efficient operation of complex engineering systems [31]. According to [32], *reliability engineering* is an engineering discipline for applying scientific know-how to a component, product, plant, or process in order to ensure that it performs its intended function, without failure, for the required time duration in a specified environment. Reliability engineering involves an iterative process of reliability assessment and improvement, and the relationship between the two processes [33]. Autonomous systems can change their behavior in response to unanticipated events during operation [34]. However, assessing the reliability of an autonomous system varies depending on the autonomy levels of the system. Previous research on autonomy levels includes the work of Huang et al. [35] who developed Autonomy Levels for Unmanned Systems that specifies metrics to assess autonomous systems capabilities. As the enablers of autonomy, the reliability engineering approach to ML algorithms is similar to traditional software reliability assessment. Abstractly, ML performs perception tasks and informed decision-making; thus, most systems that incorporate ML will naturally include standard software components [3]. Reliability growth modeling that characterizes how the reliability of a system increases during testing [3] is one of the standard approaches to software reliability assessment. In ML, the reliability growth measures the accuracy as a fraction of correct predictions divided by a total number of predictions [3]. Consequently, reliability and accuracy in ML are commonly synonymous terms [3].

According to [36], there are four technical components of reliable software:

1) Fault prevention - avoiding faults during design and development of systems through enforcement of good design methods.
2) Fault removal - the process of enforcing formal inspection and testing systems until eliminating all visible faults while not creating any new faults.
3) Fault tolerance - the survival attribute of a system.
4) Fault/failure forecasting - the process of establishing reliability models, failure data, fault/failure relationships, analysis, and interpretation of system behavior.

A reliable system has a capability to function until the system desists under *expected* circumstances. Moreover, a reliable system is a representation of the resilience engineering results.

C. Resilience Engineering for UAS

Resilience engineering brings together the system safety concepts, reliability of a system, analysis and handling uncertainties, risks, and survivability of a system. According to [1], resilience is the ability to recover quickly after something unpleasant, such as shock or an injury, the ability to return to its original shape. Hollnagel [37], who was at the forefront of resilience engineering, has developed three premises of resilience engineering that showcase limitations and issues in resilience engineering:

1) The conditions of performance are underspecified.
2) Unfavorable events can be attributed to a combination of normal performance uncertainties.
3) Safety management cannot be based on error probabilities and calculations.

These premises demonstrate the limitations within current safety engineering and pose guidelines for the continuous evolution of resilience engineering.

Fig. 1. Basic functions of a resilient system, adopted from [38]

Vachtsevanos et al. [38] illustrate the expected basic functioning of a resilient system through anticipation of undesirable events, the monitoring of performance, and the response to warnings or threats (see Figure 1). This kind of system implies proactive measures and readiness to adapt to the variability of circumstances making it less susceptible to a hazardous environment. As a result, a resilient UAS is flexible and capable of returning to a normal functioning state after experiencing disturbances.

D. Human-Machine Teaming Perspectives

Human-Machine Teaming (HMT) is a relationship between humans, the machine, and their interdependencies. The goal of HMT is to build trustworthy, transparent, predictable, adaptable, and reliable systems that incorporate artificial intelligence, to create effective human-machine teams [7]. HMT requirements [7] for an adaptable autonomous system include:

1) Multiple options or paths for recovery from a single problem (among which allowing humans to specify problem at different levels of abstraction);
2) On-demand adjustment of autonomy;
3) System degradation and failure resistance (the system shall be tolerant and fail gracefully maintaining its safety [7]).

For highly effective HMT, the most relevant requirements during the development and design stage of the autonomous systems are to ensure safe and effective systems during operations in complex, contested, unanticipated, and dynamic environments [7]. Calibrated trust (i.e., trusting the autonomous

system will do what it is supposed to do within a particular environment) and shared understanding (i.e., shared perception between human-to-machine and machine-to-human) are fundamental HMT concerns. A long-term strategy is to achieve an intuitive, shared, and bidirectional information flow between humans and machines [7].

IV. Warning Identification Framework

The UAS incorporating ML can understand the operating environment and decide their reactions to the changes in the environment. During asset surveillance, the UAS can detect anomalies in sensor measurements that can suggest possible risks or early warning signals. A risk indicates the possibility of asset disturbance, and a warning signifies the early sign of a disturbance that can require immediate reaction. The anomalous events during remote operations, such as a measured crack on the pipeline during surveillance for the offshore oil and gas industry, can be extreme and unlikely. The rarity of such measurements leads to very little evidence in data. The rare measurement can be dismissed or even unnoticed by the anomaly detection ML algorithm (as discussed by [23]). The autonomous systems' ability to detect warnings or risks is not merely about building a tool; it is about creating a long-term strategy. The UAS needs to have the possibility to react to these warnings when the situation requires it.

This section proposes an early concept of a Warning Identification Framework (WIF) to guide the planning of UAS incorporating ML in addressing the known risks and recognizing the warning signals accordingly (see Figure 2). The process of development and integration of ML into a system is referred to as the *ML lifecycle* [39]. The ML lifecycle consists of four stages: Data Management, Model Learning, Model Verification (as the activities during which machine-learned models are produced), and Model Deployment (the deployment of ML component along with the other software components in the system) [39]. The Data Management stage is responsible for the acquisition of the data that can be used to predict future data or to perform other kinds of decision making under uncertainty [40]. The planning of the Data Management stage is often underestimated. However, with the trust, reliability, and explainability issues that ML encounters, it is critical to have a clear understanding of the purpose of ML incorporated into a more extensive system. The WIF intends to address trust calibration, errors and biases, and explainability for the UAS that depend on ML algorithms (as shown in Figure 2). This framework bases on concepts and mitigation methods from Risk Assessment, Reliability Engineering, Resilience Engineering, and Human-Machine Teaming (as shown in Figure 2). Two of the factors of future risk assessment, according to [14], are the recognition of knowledge and data growth and the need for solid frameworks for the safety assessment of cyber-physical systems.

Therefore, the WIF consists of three segments:

1. Identifying the Risk: The first step in WIF is identifying the risk of experiencing disturbances in the form of rare anomalies (i.e., concerning pipeline surveillance) from available knowledge, historical insights, and domain expert inputs. In this step, Risk Assessment provides insights into risk definition based on available knowledge [14] and focusing on known events that become apparent in new conditions [28]. Known risks or vulnerabilities provide knowledge on the sequence of events that can lead to the asset or environmental disruption, frequency of occurrence of these events, and consequences of the disruption. These factors are a part of formal characterizations and representations of risk described in [41]. An extended definition, by [42], describes the knowledge of risk through defining the set of disturbance scenarios, set of consequences and, quantified uncertainties. Furthermore, a reliable system is capable of normal functioning under expected or ordinary circumstances. These circumstances are a part of the risk scenario definition. This step requires developing models based on existing knowledge to *identify* risks.

2. Hierarchy of Warning Signals : Hirerachization or ranking of the warning signals is a description of the sequence of the events that may evolve into a disturbance that requires immediate intervention. This hierarchy provides the early-to-late-warning evolution of a disturbance by defining the criticality of a warning signal. Adjusting the criticality of warning signals is a part of Risk Assessment. This adjustment allows for fault forecasting, as a characteristic of a reliable system that incorporates the analysis of warning signal relationships. Finally, as a resilient system, analysis and adjustment of the criticality of warning signals allow the system to incorporate a shared understanding of anticipation and monitoring of the disruptions. This step requires domain experts to develop models based on existing knowledge for *describing* risks.

3. Orchestration of Actions : Knowing how to respond to the emerging disturbance is one of the critical elements of reliable UAS [27], [38]. The orchestration of UAS actions is an essential task in remote operations. This step incorporates the reliable system capabilities to prevent, remove or tolerate the disruptions and a resilient system capacity to respond to the emerging situation. This step satisfies the requirement and expectation of HMT for an autonomous system to adjust the autonomy on demand. The ability for the UAS to systematically and intelligently recognize and act upon warning signals gives the system the capability of being proactive and reactive. A proactive UAS expects and captures weak signals before anomalies occur. A reactive UAS communicates and responds to the emerging situation.

Finally, the three steps of WIF satisfy the HMT requirements for an explainable functioning of a system with a shared understanding of intentions and multiple approaches to a single problem.

Warning Identification Process: Inspired by the conceptual model of Process Performance Indicator (PPI) [14], Figure 3 illustrates the *Warning Identification Process (WIP)*. PPI reflects on the degree of system objective satisfaction and describes the disruptive events leading to unwanted disruptions of the operation. The WIP demonstrates the development and identification of the warning signals by the UAS, guided by the WIF Hierarchy of Warning Signals. Displayed are stages

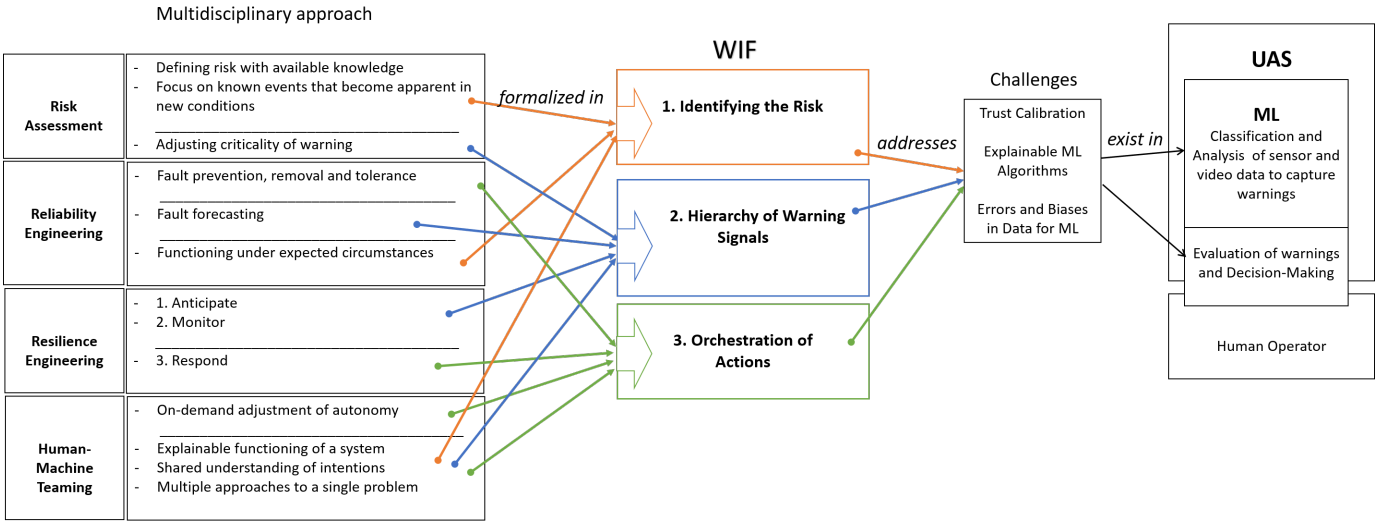

Fig. 2. Multidisciplinary approach formalized in WIF to address challenges in UAS ML

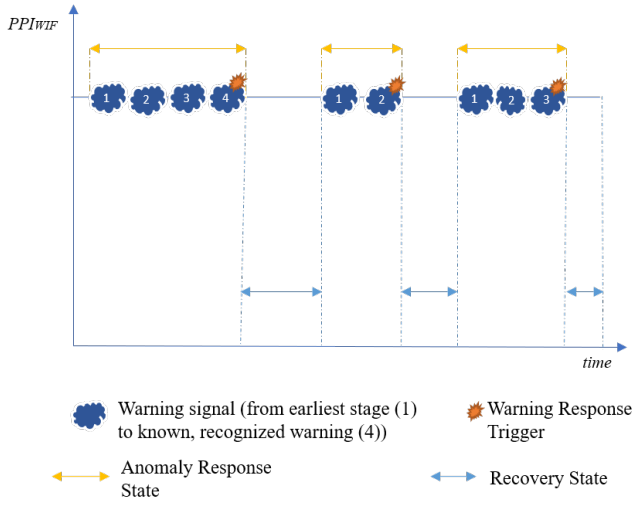

Fig. 3. Warning Identification Process

of warning from *one* to *four*, where *one* represents the earliest stage of the warning sensed by UAS, and *four* represents the latest and recognized warning that requires action. *Anomaly Response State* is the period from when the UAS detects an anomaly until it recognizes it as a warning. During the *Warning Response Trigger*, the UAS takes action and responds to the disturbance following the WIF Orchestration of the actions. Finally, the *Response State* is when the UAS returns to the natural flow of the operation, between the recognized disturbance and a new expected disturbance.

Application : The complexity of algorithms that enable autonomy makes it challenging to control, identify and characterize potential disruptions and react to the consequences. As a future factor in risk assessment, [14] highlights the need for extending the frameworks of risk assessment for complex, interconnected systems that support critical infrastructures. Disruptive events and safety barrier failures typically occur due to degradation processes [14]. Introduction of the condition-monitoring or surveillance data in WIF can give insight into the disruptive process, such as degradation, and prioritize the monitored variables. The WIF can complement the ML processes incorporated in UAS towards condition monitoring, surveillance, and intervention-based risk assessment. The proposed WIF provides a scalable, explainable and structural approach to dynamic risk assessment alongside ML.

V. Conclusion and Future Work

Implementing machine learning techniques in a standardized practice that incorporates reliability is still a matter of early development. During remote UAS operations, any unintended misbehaviors of the UAS can have severe environmental and financial consequences. With an increase in UAS employment in remote offshore operations, we observe a noticeable need to validate and improve the ML processes that enable autonomy, further supporting critical decisions during the UAS operations. The proposed framework, the Warning Identification Framework, attempts to improve the warning signal detection of UAS during remote operations, address the shared understanding of UAS ML intentions, and prevent unintentional consequences of machine learning.

Future research will further identify suitable tools to apply in each step of the Warning Identification Framework and apply the technical approach to the use-case of pipeline surveillance in the offshore oil and gas industry. The relationship between reliability and resilience engineering, risk management, and human-machine teaming expectations, such as calibrated trust, is an emerging area that plays a critical role in developing reliable autonomous systems incorporating machine learning. The authors will expand the proposed framework from the novelty and anomaly detection perspectives as a part of future research.

REFERENCES

[1] Oxford University Press, "Oxford Learner's Dictionaries — Academic English," 2021. [Online]. Available: https://www.oxfordlearnersdictionaries.com/

[2] H.-M. Huang, "Autonomy levels for unmanned systems (ALFUS) framework," National Institute of Standards and Technology, Gaithersburg, Tech. Rep. September, 2004. [Online]. Available: https://doi.org/10.6028/NIST.sp.1011-I-2.0

[3] A. Gula, C. Ellis, S. Bhattacharya, and L. Fiondella, "Software and system reliability engineering for autonomous systems incorporating machine learning," in *Proc. - Annu. Reliab. Maintainab. Symp.*, vol. 2020-Janua, 2020.

[4] V. Galaz, M. Centeno, P. W. Callahan, A. Causevic, T. Patterson, I. Brass, S. Baum, D. Farber, J. Fischer, D. Garcia, T. McPhearson, K. Levy, D. Jiménez, B. King, and P. Larcey, "Machine intelligence, systemic risks, and sustainability," in *Beijer Discuss. Pap. Ser. No. 274*, no. 274, 2021.

[5] F. Macias, "The Test and Evaluation of Unmanned and Autonomous Systems," *ITEA J.*, vol. 29, pp. 388–395, 2008.

[6] K. Okamura and S. Yamada, "Calibrating Trust in Autonomous Systems in a Dynamic Environment," *Proc. 42nd Annu. Meet. Cogn. Sci. Soc.*, pp. 1–6, 2020.

[7] P. Mcdermott, C. Dominguez, N. Kasdaglis, M. Ryan, I. T. Mitre, and A. Nelson, "Human-Machine Teaming Systems Engineering Guide," The MITRE Corporation, Tech. Rep., 2018. [Online]. Available: https://www.mitre.org/publications/technical-papers/human-machine-teaming-systems-engineering-guide

[8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019. [Online]. Available: http://dx.doi.org/10.1038/s42256-019-0048-x

[9] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," 2019.

[10] S. O. Johnsen, T. Bakken, A. A. Transeth, S. Holmstrøm, M. Merz, E. I. Grøtli, S. R. Jacobsen, and R. Storvold, "Safety and security of drones in the oil and gas industry," Tech. Rep., 2019.

[11] "IEC 60050 - International Electrotechnical Vocabulary - Details for IEV number 192-01-24: "reliability"," 2015. [Online]. Available: https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=192-01-24

[12] K. Makhlouf, S. Zhioua, and C. Palamidessi, "On the applicability of ML fairness notions," *arXiv*, pp. 1–32, 2020.

[13] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5681–5690, 2017.

[14] E. Zio, "The future of risk assessment," *Reliab. Eng. Syst. Saf.*, vol. 177, no. March, pp. 176–190, 2018.

[15] M. Compare, F. Martini, S. Mattafirri, F. Carlevaro, and E. Zio, "Semi-Markov Model for the Oxidation Degradation Mechanism in Gas Turbine Nozzles," *IEEE Trans. Reliab.*, vol. 65, no. 2, pp. 574–581, 2016.

[16] Z. Zeng, R. Kang, and Y. Chen, "Using PoF models to predict system reliability considering failure collaboration," *Chinese J. Aeronaut.*, vol. 29, no. 5, pp. 1294–1301, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.cja.2016.08.014

[17] S. Jiang, W. Zhang, J. He, and Z. Wang, "Comparative study between crack closure model and Willenborg model for fatigue prediction under overload effects," *Chinese J. Aeronaut.*, vol. 29, no. 6, pp. 1618–1625, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.cja.2016.10.002

[18] D. De Dominicis and D. Accardo, "Software and sensor issues for autonomous systems based on machine learning solutions," in *2020 IEEE Int. Work. Metrol. AeroSpace, Metroaerosp. 2020 - Proc.*, 2020, pp. 545–549.

[19] C. A. Thieme and I. B. Utne, "Safety performance monitoring of autonomous marine systems," *Reliab. Eng. Syst. Saf.*, vol. 159, no. March 2016, pp. 264–275, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.ress.2016.11.024

[20] B. Cline, R. S. Niculescu, D. Huffman, and B. Deckel, "Predictive maintenance applications for machine learning," *Proc. - Annu. Reliab. Maintainab. Symp.*, 2017.

[21] Y. P. Fang and E. Zio, "An adaptive robust framework for the optimization of the resilience of interdependent infrastructures under natural hazards," *Eur. J. Oper. Res.*, vol. 276, no. 3, pp. 1119–1136, 2019.

[22] J. Zhang, M. Xiao, and L. Gao, "An active learning reliability method combining Kriging constructed with exploration and exploitation of failure region and subset simulation," *Reliab. Eng. Syst. Saf.*, vol. 188, no. January, pp. 90–102, 2019. [Online]. Available: https://doi.org/10.1016/j.ress.2019.03.002

[23] K. Liu and H. Ma, "Exploring background-bias for anomaly detection in surveillance videos," *MM 2019 - Proc. 27th ACM Int. Conf. Multimed.*, pp. 1490–1499, 2019.

[24] S. Elbaum, S. Kanduri, and A. A. Amschler, "Anomalies as precursors of field failures," *Proc. - Int. Symp. Softw. Reliab. Eng. ISSRE*, vol. 2003-Janua, pp. 108–118, 2003.

[25] O. Ibidunmoye, F. Hernández-Rodriguez, and E. Elmroth, "Performance anomaly detection and bottleneck identification," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 1–35, 2015.

[26] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," *arXiv*, pp. 6479–6488, 2018.

[27] F. Scibilia, K. S. Tungland, A. Røyrøy, and M. B. Asla, "Energy industry perspective on the definition of autonomy for mobile robots," in *Commun. Comput. Inf. Sci.*, vol. 1056 CCIS, 2019, pp. 90–101.

[28] R. Flage and T. Aven, "Emerging risk - Conceptual definition and a relation to black swan type of events," *Reliab. Eng. Syst. Saf.*, vol. 144, pp. 61–67, aug 2015.

[29] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, jun 2014.

[30] M. Masana, I. Ruiz, J. Serrat, J. van de Weijer, and A. M. Lopez, "Metric Learning for Novelty and Anomaly Detection," *arXiv*, no. abs/1808.05492., 2018. [Online]. Available: http://arxiv.org/abs/1808.05492

[31] A. K. Verma, S. Ajit, and D. R. Karanki, *Reliability and Safety Engineering*, 2nd ed. Springer-Verlag London, 2016. [Online]. Available: https://www.springer.com/gp/book/9781447162681

[32] D. Kiran, "Reliability Engineering," in *Total Qual. Manag.* Elsevier, jan 2017, pp. 391–404. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128110355000271

[33] "Chapter 7 - Reliability Engineering," in *Lees' Loss Prev. Process Ind. (Fourth Ed.*, fourth edi ed., S. Mannan, Ed. Oxford: Butterworth-Heinemann, 2012, pp. 131–203. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123971890000070

[34] D. Watson and D. Scheidt, "Autonomous systems," *Johns Hopkins APL Tech. Dig. (Applied Phys. Lab.*, vol. 26, pp. 368–376, 2005.

[35] H. M. Huang, K. Pavek, B. Novak, J. Albus, and E. Messina, "A framework for Autonomy Levels for Unmanned Systems (ALFUS)," *AUVSI's Unmanned Syst. North Am. 2005 - Proc.*, no. June, pp. 849–863, 2005.

[36] M. R. Lyu, *Handbook of Software Reliability Engineering*, F. C. Spencer Marjorie, Ed. United States of America: The McGraw-Hill Companies, Inc., 1996, vol. 37, no. 7. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167923603000204

[37] E. Hollnagel, "Resilience Engineering: A New Understanding of Safety," *J. Ergon. Soc. Korea*, vol. 35, no. 3, pp. 185–191, jun 2016. [Online]. Available: http://dx.doi.org/10.5143/JESK.2016.35.3.185http://jesk.or.kreISSN:2093-8462

[38] G. Vachtsevanos, B. Lee, S. Oh, and M. Balchanos, "Resilient Design and Operation of Cyber Physical Systems with Emphasis on Unmanned Autonomous Systems," *J. Intell. Robot. Syst. Theory Appl.*, vol. 91, no. 1, pp. 59–83, jul 2018. [Online]. Available: https://doi.org/10.1007/s10846-018-0881-x

[39] R. Ashmore and C. Paterson, "Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges," Tech. Rep., 2019.

[40] K. P. Murphy, "Machine Learning: A Probabilistic Perspective," The MIT Press, Tech. Rep., 2012.

[41] S. Kaplan and B. J. Garrick, "On The Quantitative Definition of Risk," *Risk Anal.*, vol. 1, no. 1, pp. 11–27, 1981. [Online]. Available: https://doi.org/10.1111/j.1539-6924.1981.tb01350.x

[42] T. Aven and O. Renn, *Risk Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 121–158. [Online]. Available: https://doi.org/10.1007/978-3-642-13926-0_8