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-P7886-cd

Safe AI vs Safe use of AI

Rune Winther
Institute for Energy Technology, Norway. E-mail: rune.winther@ife.no

Rune Fredriksen
Institute for Energy Technology, Norway. E-mail: rune.fredriksen@ife.no

Recent advances in the development of AI has, predictably, led to a massive increase in research on how to make AI safe. While this is a valid and important quest, this paper emphasizes the difference between requiring an AI to be safe, and that AI is *used* safely. Using AI to control a system with potentially serious safety risks, even with the goal of making the system safer through the use of AI, doesn't necessarily imply that the AI itself must be safe. Focusing too much on making AI safe could result in a two-fold problem: 1) We cannot fully utilize the potential of AI; (because) 2) We struggle with demonstrating adequate safety for the AI. It is our opinion that problem 2 often can be avoided, and problem 1 alleviated *because* problem 2 is not relevant. This, however, requires a somewhat different way of thinking about AI in safety-critical systems than seems often to be the case. In this paper we discuss these problems, illustrate them with examples, and show that there is already much knowledge on how to achieve documented safety of AI-enabled systems without having to provide safety assurance for the AI itself.

Keywords: AI, safety, assurance

1. Introduction

A recent survey of approaches to using Artificial Intelligence in Safety-Critical Systems within industry and transportation Perez-Cerrolaza et al. (2024), covering close to 300 sources, documents two things very clearly: Firstly, the last 5-8 years have seen a significant increase in research related to AI and safety, which is not surprising given the success of AI in this period. Secondly, there is much focus on how to assure the AI itself is safe, so that it can be used to manage safety.

There is a dilemma here: Requiring the AI itself to be safe is going to put constraints on the AI, thus reducing its capabilities and therefore the value of using AI in the first place. While there are good reasons to research the possibility of using AI as part of safety functions, we are concerned that this will have limited value in reality and that this could lead to a suboptimal utilization of AI in safety critical contexts.

The goal of this paper is to debate and clarify the differences between the concepts of "safe AI" and "safe use of AI", and look into some of the alternative strategies for creating benefits from AI in systems where safety is a concern. We will also argue that using AI in safety-critical systems is often more a concern for reliability than it is for safety.

This paper primarily discusses AI within the context of Cyber Physical Systems (CPS), and our focus is on systems where failures could potentially cause serious harm to people or the environment. This means that we address the use of AI for autonomous control of physical systems/processes, i.e. industrial systems, transportation, etc. Our focus is on technical aspects, and we will therefore not consider human-AI interaction as part of this paper. While we do not make any specific assumptions on the type of AI, we are concerned with enabling the most capable AI possible. This means that Deep Neural Networks (DNNs) is of special interest for our research.

This paper is structured as follows: Chapter 2 gives a brief review of approaches to "AI and safety". In Chapter 3 we will argue that focusing on making AI itself safe could cause suboptimal use of AI in safety-critical systems, while Chapter 4 shows that benefits from AI can be realized without the AI becoming safety-critical. Chapter 5 builds on Chapter 4, showing that appropriate

choice of system architecture may even allow adaptive (thus continuously improving) AI as a central element in safety-critical systems. Fact is that the theory and methods for doing this to a large extent already exists. Chapter 6 presents a general discussion on the topics of this paper, before Chapter 7 summarizes and concludes the paper.

2. AI and safety - A brief review

The topic of using AI in safety-critical systems is not a new one. In fact, there was a lot of activity already 20 years ago. Giving a complete overview of the history and status-quo is not possible within the context of this paper, so we refer to Perez-Cerrolaza et al. (2024) and Bloomfield and Rushby (2024) for more comprehensive presentations. Between them, they provide several hundred references to work of relevance to safety and AI. It is also worth noting that AI and safety is addressed in standardization activities, e.g. ISO/IEC5469 (2024). In this paper, we limit the presentation to aspects of relevance for the discussion in this paper.

There are many approaches on how to utilize AI in safety-critical systems, and many ways to manage safety in AI-enabled systems. As described in Bloomfield and Rushby (2024), from approaches focusing on system dependability by minimizing the need for trust in the AI, to approaches seeking to establish trust in the AI itself, there is in reality a spectrum of approaches. Perez-Cerrolaza et al. (2024) defines several categories for both the use of AI, and for approaches to manage and document safety.

A broad way of classifying approaches to using AI in safety-critical systems is as either black-box or white/grey-box. The difference being that in black-box approaches you make no assumptions about the inner structure or workings of the AI, while in the white/grey box case you assume access to at least some information, and/or are able to manipulate the internal workings of the AI. Within each of these there is a multitude of approaches, and the following is just a brief and simplified glimpse.

Techniques relevant for white/grey-box approaches include formal verification Vassev (2016) and various monitoring techniques. Formally verifying safety properties of AI is very attractive, because it provides a high level of trust. However, formally verifying DNNs with large numbers of parameters is not a practical reality, and we will not go further into this topic. Monitoring was among the early proposed approaches Cukic et al. (2006), and can be done in many ways. One possibility is to use monitoring to control how the AI evolves during training, another to monitor the output from the AI in order to check its validity before passing it on. Formal approaches and monitoring may also overlap, e.g. in cases where formally specified rules are used in the monitoring of the AI. Again, we refer to Perez-Cerrolaza et al. (2024) for a thorough presentation of the various approaches.

Black-box approaches are in reality limited to monitoring/evaluation of the AI's output, and the use of "safety-bag" techniques. Safety-bag means that the system has an architecture where failure of the AI can be prevented from causing hazardous system states. Monitoring of the AI's output must in this case be done without knowledge of the AI itself, i.e. it must be based on system models. An important distinction exists between monitoring the AI's outputs, and monitoring the system's state. In many cases, the state of a system will change slowly enough to allow monitoring of the AI's effect on the system before deciding whether an intervention is necessary. The latter approach allows the AI the most "freedom", i.e. interventions are done at the last possible moment.

Monitoring and safeguarding systems based on models is in reality an important topic of its own, and has become increasingly important as the use of advanced autonomy (with or without AI) has increased. Ames et al. (2019) provide a recent discussion of control theory in the context of safety, while García and Fernández (2006) and Osborne et al. (2021) discusses the possibility of safe online learning of control systems. ASTM (2021) provides a specific way to go about this, through defining Run Time Assured (RTA) architectures. This provides an architectural framework for developing a system which provides run-time

assurance as an alternative to design-time assurance to fulfill safety requirements for an unassured or complex function (e.g. AI). An early approach similar to this, but specifically aimed at AI, was discussed in Winther (2006). It is also worth mentioning that Bloomfield and Rushby (2024) systematically lists a number of relevant architectural concepts.

Although non-trivial, the various techniques for real-time monitoring of black-box control algorithms represent approaches with potentially great value for the practical use of AI in safety-critical systems, exactly because they treat AI as a black-box.

Before we end this chapter we also need to address the issue of how to convincingly argue that an AI-enabled safety-critical system is safe. While there are standards and guidelines emerging, addressing AI in the context of safety, e.g. ISO/IEC5469 (2024); DNV (2023), we will focus on certain specific principles we consider most flexible and capable as a general approach to documenting and communicating assurance of AI-enabled safety-critical systems. A methodology that has proven to work well for making safety cases for complex systems is the use of graphical argumentation notations, such as e.g. Goal Structuring Notation (GSN) SCSC (2021).

What type of evidence we need to produce to argue adequate safety depends directly on how the AI is used in the system. If the AI is a key safety-element, we need evidence addressing the AI specifically. In cases where safety is managed by other means, e.g. safety-bag techniques, the evidence need to address architectural issues and the safety of non-AI elements.

Hawkins et al. (2021) has proposed a GSN-based methodology directly addressing the assurance of ML-elements, called Assurance of Machine Learning for use in Autonomous Systems (AMLAS). The aim of AMLAS is to support systematic integration of safety assurance while developing ML components, and the generation of evidence justifying the safety when integrating ML-components into autonomous systems. AMLAS consists of a 6-stage iterative process, where the use of assurance argument patterns are an important part. These are standardized pieces of argumentation described using GSN-notation. Hawkins et al. (2023) presents a practical application of AMLAS.

While AMLAS is limited to the ML/AI components themselves, Bloomfield et al. (2021) presents safety case templates for the broader context of autonomous systems, using many of the same principles as AMLAS, but also includes approaches to manage situations where ML/AI is an *element* of a bigger system.

We will now turn our focus to what we consider to be potentially problematic about aiming for safe AI.

3. Why focusing on safe AI might not be the best strategy

A fact that is important to remember regarding AI and safety is that AI is software, and the level of assurance needed for safety-critical software will also be required for safety-critical AI. The question, and focus of much of the ongoing research, is therefore on how to practically achieve this.

A fundamental requirement for assurance is transparency and predictability. If we cannot determine how a system behaves, we cannot trust it. One of the most basic types of information on a system's behavior comes from testing. When testing, we can treat the system as a black box, and don't have to make any assumptions on its internal workings. However, there are substantial limitations regarding the level of assurance this can provide. As demonstrated more than three decades ago, solely basing assurance on testing is not feasible Butler and Finelli (1993); Littlewood and Strigini (1993). Although there are strategies that to some extent alleviate the problem, e.g. through bootstrapping Bishop et al. (2021), statistical approaches will have limited value with regard to providing assurance. This means that to achieve any level of assurance reasonable for safety-critical AI, we need knowledge of, and/or control with, the inner workings of the AI. This means that we are, in reality, limited to white/grey-box techniques, which implies constraints on the AIs we use.

Obvious options are:

- Using AI-techniques for which assurance is possible.
- Controlling how machine-learning (ML) models evolve during training. Which would require that we identify and formulate rules/requirements the AI must comply with.
- Run-time monitoring of the AI's output, which means that we must establish rules the output must satisfy.

The first option implies a serious limitation on what AI-techniques can be used, and ML-based approaches will be challenging. Monitoring and controlling how an ML-based model evolves during training (the second approach above) provides more options, but ensuring that all rules/requirements have been identified and satisfied, is not a trivial task. This approach will likely have to be conservative, and thus put significant limitations on the algorithm's possibility to learn optimally. The third option is the one that provides the greatest degree of freedom, and it is supported by established methods and theory (e.g. Ames et al. (2019)). However, this approach will limit the AI's behavior in the way that any output not considered acceptable by the monitor will be discarded, even if that output in reality is "smarter" than that of a conventional algorithm. I.e., the smartness of the AI is limited by the monitor.

The effect of these issues is that aiming for safe AI will likely limit the effect we can get from using AI, thus potentially representing a sub-optimal strategy for utilizing AI in safety-critical contexts. Whether this is the case, however, depends on whether we can utilize AI in a safety-critical system while limiting the criticality of the AI itself. This is the topic of the next chapter.

4. Realizing the benefits of AI in safety-critical systems without the AI becoming safety-critical

Safety is never the primary function when we develop new systems. Safety is something we sometimes need to manage, in order to get the functionality we need and want. The primary function of a car or an aircraft is to transport people and things from one place to another, but because the transport may fail in ways that can cause harm, we need to ensure that these systems, in addition to being effective and efficient in doing their primary task, also are safe. Thus, whenever we consider using AI in a system where safety is an issue, it is paramount that we really understand what we want to get from using the AI. As we will see later, using a common approach to managing safety, using AI in safety-critical systems will often be a reliability problem, not a safety problem. The following example, while somewhat simplified compared to the real world, illustrates the points above.

A modern car is typically equipped with cameras and AI image analysis capabilities, enabling the car to establish a real-time model of its surroundings. This model can be used for many things, e.g. adaptive cruise control, conflict avoidance (warning of other cars in driver's blind spots), lane assist and collision avoidance. Of these, only the last two are truly safety functions. Adaptive cruise control and conflict avoidance are not critical by themselves, because safety is managed by the collision avoidance function. Using AI as part of adaptive cruise control and conflict avoidance is sensible, because this will enable more sophisticated handling and smoother driving than is possible with simpler methods. It is also not problematic with regard to safety, because safety is managed outside these functions. While it is undoubtedly tempting to use AI also for lane assist and collision avoidance it would require us to establish a high level of assurance for the AI. The question is whether AI is necessary to realize these safety functions. Looking closer, we see that (in this case) it is not:

- Lane assist can be achieved by e.g. having a radiating cable (antenna) in the road base layer, detectable by sensors on the car. This is similar to the classic case of robots following white or black lines, and only require a few lines of basic programming.
- Collision avoidance can be achieved using LIDAR/RADAR and simple rule-

based logic.

Ensuring safety by having a ”safety layer”, often accomplished using basic conventional technology, is of course a well-known approach. We see it in nuclear power plants, in the form of reactor trip systems, and railway signaling systems which will go to a fail-safe state (signals set to ”stop”) if any of a number of failure conditions are detected. It is therefore imperative to be conscious on how, and for what, the AI is used, and that using conventional ways of achieving safety would help us avoid making the AI unnecessarily critical.

However, assuming that conventional safety functions are available, using AI as part of higher-level functions in a safety context is not without problems. Unexpected behavior of the AI will likely have the effect of triggering the safety functions, thus causing *reliability issue*. However, managing a reliability issue is clearly preferrable to safety issues, because the assurance needed usually will be much less rigorous.

In the next chapter we will turn our attention to the practical issues on how to exploit AI in safety-critical systems, and specifically address the potential of using reinforcement learning.

5. Practical approaches to enable continuously improving AI in safety-critical systems

The main reason we want to use AI in the control of systems is that AI potentially perform better than conventional models. If we can also allow the AI to continue to learn while in operation, the performance of the AI-model would steadily adapt and improve. The problem is, of course, that assuring safety of continuously improving AI is even more challenging than for pre-trained AI. However, from the previous chapters we know that there are approaches where safety is assured through redundancy, e.g. Run Time Assured (RTA) architectures ASTM (2021) and various strategies for defense in depth Bloomfield and Rushby (2024). Furthermore, because these approaches considers the AI as a black box, it doesn't really matter whether the AI continues to improve or not. As we have seen, there is even research specifically focused on safe online learning of control systems Osborne et al. (2021).

In this chapter, we will discuss practical approaches to enabling reinforcement learning in the control loop of safety-critical systems, addressing both system design issues and how to provide assurance arguments on safety. The theory is already there, so our focus is on *how* to do it. In the following subsections we will address two specific approaches to using AI in safety-critical systems, both are variations of ”AI-upgrading”. When introducing AI to an existing system, we have the benefit of having access to an already assured system.

It should be noted that the following discussion focuses on principles, and therefore is simplified as compared to real-world cases. In the real world, there will be a number of practical and formal issues that must be addressed.

5.1. *AI-upgrade: Scenario 1*

In this scenario we assume that the the system has a verified safety function that is (adequately) independent of the control system, as illustrated in Figure 1. Solutions similar to this are found in e.g. nuclear power plants.

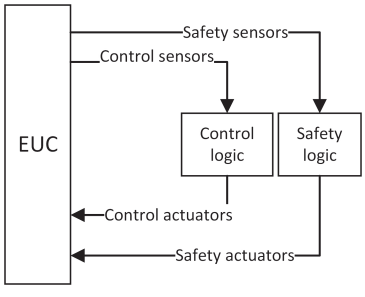

Fig. 1. Control system with an independent safety function. (EUC = Equipment Under Control, e.g. a chemical plant, self-driving car, etc.)

This means that we have:

- A pre-existing control system we know performs fairly well, and in particular does not cause the activation of safety functions more often than is acceptable.
- A safety function that has been assured as adequately safe.

We can exploit this as illustrated in Figure 2. This architecture is a combination of Architecture 2 in Bloomfield and Rushby (2024) and a safety-bag technique where the state of the system is monitored. Note that the monitor is a decision device, and that for every cycle of reading sensors and commanding actuators, the monitor decides whether it is the AI or the pre-existing control logic that shall be allowed to control the process in the next cycle of the control loop. The monitor is placed upstream of the AI because it monitors the *effect* of the AI on the system, not the output from the AI.

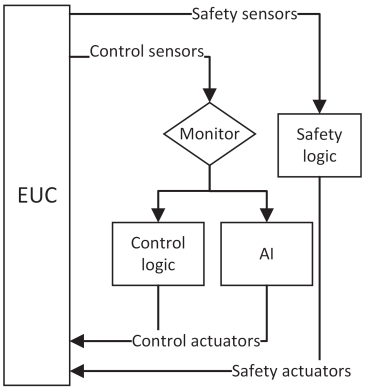

Fig. 2. Introducing AI, exploiting existing resources.

An important point here is that in this specific case the safety-bag technique is used to ensure *reliability*, not safety, because safety is managed by the dedicated safety logic. However, this doesn't mean that safety is necessarily unaffected by the change. In a very simplified form, illustrated using Goal Structuring Notation (GSN) SCSC (2021), arguing for safety must be adapted as seen by comparing Figures 3 and 4.

What is required, from a safety perspective, is therefore only to show that the introduction of a monitor and the AI will not affect the safety function. Given that this was possible in the original case, there is no reason that this should pose a serious problem.

The main benefits of the architecture in Figure 2 are that safety will be as before, and that the use of the monitor and the pre-existing control logic

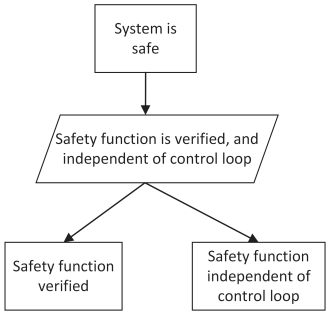

Fig. 3. Safety argument for the architecture in Figure 1.

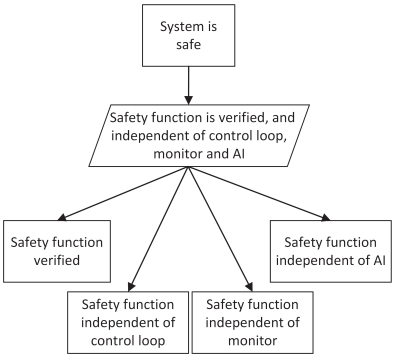

Fig. 4. Safety argument for the architecture in Figure 2.

ensures that the AI doesn't significantly reduce reliability. It is also obvious that this set-up will allow online training of the AI, i.e. reinforcement learning. The "price" we must pay to achieve this, is the development of the monitor and the burden of documenting additional independencies.

5.2. *AI-upgrade: Scenario 2*

We will now look at a slightly different situation, where there is no independent safety function. I.e., safety is managed by the control logic itself. This situation is illustrated in Figure 5. If we want to upgrade this system to AI, the obvious choice is some sort of safety-bag technique, e.g. as illustrated in Figure 6.

For the system in 6, the monitor and the pre-existing control logic must together ensure both safety and reliability. More specifically, as described in Winther (2006), we need to substantiate the following claims:

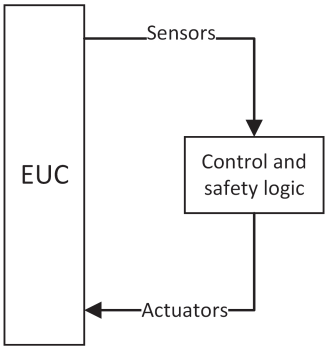

Fig. 5. System where safety is integral to the control logic.

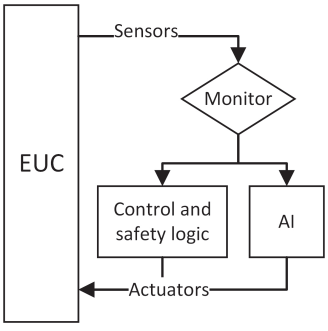

Fig. 6. AI-upgrade of the system in Figure 5.

(1) That the monitor is able to detect when the system is outside defined bounds, and transfer control to the pre-existing control logic before the system enters an unsafe state.
(2) That the pre-existing control logic is able to avoid an unsafe state when given control.
(3) That neither the AI or the monitor can cause the pre-existing control logic to fail in ensuring safety.

Support for claims 1 and 2 can be based on existing theory and methods, e.g. using control barrier functions Ames et al. (2019). Claim 3 is of the same type regularly handled when proving integrity of safety functions, and thus a well-known (if not trivial) issue.

5.3. *Comparing the scenarios*

While the two scenarios utilize the same basic principle for integrating AI into an existing system, there is one major difference between them. In scenario 1, the introduction of AI is primarily a reliability problem. Safety is managed through an independent dedicated safety function, and the pre-existing control logic is only used to reduce the chance of the AI triggering the safety function. In the second scenario, the monitor takes on a much more important role with regard to safety. We also need to show that the pre-existing control logic will remain safe when used as a back-up, and only called upon when a potentially unsafe state is approached.

While scenario 1 is preferrable, we consider scenario 2 to be a viable option in many situations. The key question is how fast the system can change state, and whether it will be possible to do the switching from AI to the pre-existing control logic in time to prevent an unsafe state. Thus, the viability of the approach in the second scenario depends on the nature of the process that is being controlled.

6. Discussion

While using AI in systems where safety is important is not without concerns, the capabilities of AI are obviously interesting as the complexity of systems increase. We are concerned, however, that there will be serious limitations on what level of assurance can be achieved for AI, and in particular for the most capable types of AI (typically DNNs). Because certain AI-models are very capable, and in some cases the only realistic option (e.g. in autonomous driving), one can easily end up using AI to perform safety functions. If we are not able to establish the necessary level of assurance for the AI, we are stuck.

However, as we have seen, there are ways to avoid making AI critical, even if it plays a central role in systems where safety is imperative. In fact, much of the theory for doing this already exists. We are concerned there is lack of consciousness about this, and that this will unnecessarily limit the potential benefits of using AI.

We believe a good strategy for using AI in safety-critical systems will be to first explore system design/architectural options to reduce the need for trust in the AI as much as possible. Being conscious that conventional safety functions

might be available, and that there are already in place much knowledge and theory that can be used to reduce the criticality of AI, is essential.

7. Summary and conclusions

In this paper we have seen that there is much knowledge relevant to the exploitation of AI in systems where safety is important. A major point of this paper is the importance of understanding the difference between "safe AI" and the "safe use of AI". The latter includes the former, but will allow the use of more capable AIs than is possible in a "safe AI" approach. As we have shown, there are ways of using AI in safety-critical systems without the AI itself becoming safety-critical. The relevance of such approaches is based on the fact that more often than not, there are other reasons than safety for wanting to use AI.

References

Ames, A. D., S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada (2019). Control barrier functions: Theory and applications. *18th European Control Conference (ECC)*.

ASTM (2021). Standard practice for methods to safely bound behavior of aircraft systems containing complex functions using run-time assurance. *ASTM International*.

Bishop, P., A. Povyakalo, and L. Strigini (2021). Bootstrapping confidence in future safety based on past safe operation. *CoRR abs/2110.10718*.

Bloomfield, R., G. Fletcher, H. Khlaaf, L. Hinde, and P. Ryan (2021). Safety case templates for autonomous systems. *CoRR abs/2102.02625*.

Bloomfield, R. and J. Rushby (2024). Assurance of AI systems from a dependability perspective. *SRI Project 101425, CSL Technical Report SRI-CSL-2024-02R2*.

Butler, R. W. and G. B. Finelli (1993). The infeasibility of experimental quantification of life-critical software reliability. *IEEE Transactions on Software Engineering 19(1)*, 3–12.

Cukic, B., E. Fuller, M. Mladenovski, and S. Yerramalla (2006). Run-time assessment of neural networks control systems. *Methods and Procedures for the Verification and Validation of Artificial Neural Networks*, 257–269.

DNV (2023). Recommended practice for assurance of AI-enabled systems. *DNV-RP-0671*.

García, J. and F. Fernández (2006). Run-time assessment of neural networks control systems. *Methods and Procedures for the Verification and Validation of Artificial Neural Networks*, 257–269.

Hawkins, R., C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli (2021). Guidance on the assurance of machine learning in autonomous systems(amlas). *CoRR abs/2102.01564*.

Hawkins, R., C. Picardi, L. Donnell, and M. Ireland (2023). Creating a safety assurance case for an ml satellite-based wildfire detection and alert system. *Journal of Intelligent & Robotic Systems 108:47*.

ISO/IEC5469 (2024). Tr 5469:2024 artificial intelligence - functional safety and AI systems.

Littlewood, B. and L. Strigini (1993). Validation of ultrahigh dependability for software-based systems. *Communications of the ACM 36*, 69–80.

Osborne, M., H.-S. Shin, and A. Tsourdos (2021). A review of safe online learning for nonlinear control systems. *International Conference on Unmanned Aircraft Systems (ICUAS)*.

Perez-Cerrolaza, J., J. Abella, M. Borg, C. Donzella, J. Cerquides, F. J. Cazorla, C. Englund, M. Tauber, G. Nikolakopoulos, and J. L. Flores (2024). Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Comput. Surv. 56*(7).

SCSC (2021). Goal structuring notation community standard version 3. *SCSC Assurance Case Working Group*.

Vassev, E. (2016). Safe artificial intelligence and formal methods. *In Proceedings of the Leveraging Applicat. of Formal Methods, Verification and Validation: Foundational Techniques*, 704–713.

Winther, R. (2006). Fault tolerance to facilitate the use of artificial intelligence in critical systems. *Proceedings of ESREL 2006*.