Supporting the Introduction of Machine Learning in Safety-Critical Aerospace Industries (Level M)

David Wood (2198230W)

February 2020

Abstract

Problems which were too complex to solve using traditional programming methodologies are increasingly becoming possible as a result of innovations in machine learning. Machine learning algorithms "figure out" how to solve problems from data and this raises questions as to their suitability for inclusion in safety-critical aerospace systems, where predictability is paramount. Development of unmanned aerial vehicles and pilot assistance systems, made possible by machine learning, create a tension between innovation and safety that current functional safety standards do not address. In this report, current state-of-the-art research into the integration of machine learning and aerospace systems will be presented. As the automotive industry has more experience with the interactions between functional safety and machine learning, particularly in advanced driver assistance systems and autonomous vehicles, existing research on the use of machine learning in safety-critical automotive systems will also be presented to supplement the lack of available research from the aerospace industry. Furthermore, this report will present a hazard analysis technique which extends HAZOP and sneak analysis methods to identify inadequacies in training data used by machine learning components. Through this technique, quality and breadth of training data can be improved, reducing the occurrence of accidents and failures which result from rare real-world scenarios that had not previously been considered. Finally, an evaluation of this technique will be presented where the results of software engineers from industry using exclusively HAZOP will be compared with the results from using the technique presented by this report to identify hazards which result from training inadequacies.

1 Background

Traditional software systems are composed of hand-crafted logical rules which describe the behaviour of those systems. In contrast, software systems which utilise machine learning automatically formulate the rules from data. Unsurprisingly, machine learning techniques are becoming increasingly popular with software engineers looking to implement systems which were too complex to develop traditionally, including those in safety critical contexts.

Machine learning is an umbrella term for a bewildering selection of algorithms, each attempting to solve a different problem with different trade-offs. The ability of machine learning to "figure out" how to solve a problem from the data makes it well-suited to solve problems like classification - the categorisation of data; regression - the prediction of numerical values; and clustering - the grouping of similar data. Algorithms for machine learning generally vary in three main components: representation, evaluation, and optimisation [4].

Any system where failure can threaten life or the environment is a safety critical system. Safety critical systems are commonplace in modern society, and are increasingly dependent on the correctness of software for their primary functions. Aerospace applications - aviation and space systems - both military and commercial, are classic examples of safety critical systems.

Utilisation of machine learning technologies in aerospace systems opens up a wealth of opportunity for technological innovation, but not without risk. Existing functional safety standards for aerospace weren't written with machine learning technologies in mind.

As much of the innovation in the aerospace industry happens in a military context, there isn't a wealth of existing research on the integration of machine learning. However, integration of machine learning in the automotive industry is a subject with vast amounts of current research. As such, research from the automotive industry will be used extensively throughout this report to enrich the techniques developed in this paper for the aerospace industry.

In the aerospace industry, pilot assistance systems and autonomous unmanned aerial vehicles (UAVs) are being developed which use machine learning. These systems share many characteristics with Advanced Driver Assistance Systems (ADAS) and Autonomous Vehicles (AV) being developed in the automotive industry [14][9].

Functional safety is part of the overall safety of a system that depends on an automatic protection system which must respond correctly to inputs and predictably respond to failures.

IEC 61508 [6] is the umbrella standard for functional safety and covers all electric, electronic, and programmable electronic safety-related systems (E/E/PE).

DO-178C is the standard by which the certification authorities for aerospace systems - Federal Aviation Administration, European Union Aviation Safety Agency and Transport Canada - approve software-based aerospace systems.

ISO 26262 [8] derives from IEC 61508 and is responsible for functional safety of road vehicles from the automotive industry. SOTIF or ISO/PASS 21448 [7] is a new standard that applies to functionality which demands situational awareness from the operator in order to be safe.

DO-178C defines Design Assurance Levels or DALs to measure the risk of a component. Similarly, ISO 26262 defines Automotive Safety Integrity Levels or ASILs and IEC 61508 defines Safety Integrity Levels or SILs.

DAL E is the least stringent and roughly compares to ASIL A (there's no corresponding SIL). DAL B is the second-most stringent safety level and corresponds to ASIL D and SIL 3. DAL A is the most stringent and corresponds to SIL 4 (there's no corresponding ASIL). ASILs are similar to Safety Integrity Levels (or SILs) from IEC 61508 and measure the degree of rigor required to reduce the risk associated with a component.

Recent innovations in computer science, such as those which make machine learning techniques practical, aren't addressed by the ISO 26262 or DO-178C standards. As a result, there is a need for new techniques which address the tension between innovation and safety that has been created.

In particular, there are two primary areas where DO-178C and ISO 26262 fail to adequately anticipate the unique challenges brought by the introduction of machine learning techniques - hazard analysis and development process/software techniques.

Hazard Analysis and Risk Assessment (HARA) methods are recommended by ISO 26262 to pinpoint hazardous events in a safety critical system and ensure that mechanisms are established to mitigate the hazards.

There is a broad spectrum of potential hazards introduced by machine learning techniques into automotive and aerospace systems.

By introducing advanced driver assistance systems into automotives, the human operator of the vehicle can become complacent and over-estimate the capabilities of the vehicle [11]. Furthermore, as operators become increasingly reliant on automotive vehicles, their skill level can diminish and become insufficient to adequately correct for malfunctions [3].

In addition, the "correctness" of machine learning algorithms depends heavily on the appropriateness of the dataset that was used in training, an issue which also exists in aerospace systems [2]. ISO 26262 assumes that

behaviour of components can be fully specified, which creates challenges for the integration of machine learning-based components. It is unclear how to guarantee that the training dataset used is comprehensive enough so that all environments and scenarios that might be encountered by the system are accounted for.

Existing hazard analysis techniques are not sufficient to address the wide range of hazards that result from the machine learning techniques being integrated into automotives.

2 State of the Art

Bhattacharyya et al. [2] discuss the changes required to civil aviation certification processes which currently require correct system behaviour be completely specified. In their report, the authors describe the motivating applications of adaptive control and artificial intelligence algorithms, including autonomous operation of UAVs. In addition, the authors provide a detailed overview of the current certification process and requirements and of common adaptive control and artificial intelligence algorithms.

Furthermore, Bhattacharyya et al. detail the challenges faced for certification of these systems, including:

- Definition of comprehensive set of requirements (that are both complete and can be decomposed) is challenging due to the dynamic runtime nature of these systems.
- Difficulty in verification of a comprehensive set of requirements (were such a set of requirements to be created).
- Requirement that source code be managed internally to prevent unauthorised changes is in stark contrast to current development practices for machine learning systems.
- Lack of determinism and conventional design artifacts from adaptive systems.

Bhattacharyya et al. conclude with potential mitigation strategies - education, modified certification standards, new verification approaches, architectural mitigations, and licensing changes.

Brookhuis [3] provides an overview of the behavioural impacts that automated vehicle guidance systems in aerospace and automotive can have (such as advanced driver assistance systems). Brookhuis discusses the increased

likelihood of failure due to complexity as automation is introduced; the chance of increased reaction time and need for operator awareness; and the risk of complacency as the operator becomes (over)reliant on the automated system.

Amodei [1] et al. detail five practical research problems which relate to safety of artificial intelligence - avoiding negative side-effects, avoiding reward hacking, scalable oversight, safe exploration and robustness to distributional shift. The authors conclude that with the realistic possibility of machine-learning based systems being integrated into aerospace projects, industrial processes and health-related systems, that a unified approach is needed to prevent these systems from causing unintended harm.

Koopman and Wagner [10] describe the interdisciplinary challenges involved autonomous vehicles in aerospace and automotive. The authors discuss the issues with existing standards and certifications as they relate to systems which act autonomously. In addition, the difficulty in reasoning about the correctness of a machine learning system is raised, as well as the challenges in testing; security and finding cost-effective hardware.

Schumann et al. [13] discuss the application of artificial neural networks (ANNs) to high-assurance systems. Traditional verification & validation approaches are deemed insufficient for neural-network based applications, particularly those which are trained in an online fashion. Schumann et al. cite non-determinism and the inherent incompatibilities between the situations where neural networks excel and those that safety critical systems aim to avoid.

Stolte et al. [15] present a detailed hazard analysis and risk assessment (according to ISO 26262) for an unmanned protective vehicle. Stolte et al. demonstrate that conventional HARA approaches are of limited suitability, particularly for applications with a wider functional range. In addition, the authors discuss the need for a clarification of terminology between autonomous vehicles and functional safety.

Salay et al. [12] analyse the issues that arise when machine learning systems are used in conjunction with functional safety standards. The authors identify five key areas where ML can impact the ISO 26262 standard: identification of hazards; fault and failure modes; the use of training sets; level of ML usage; and required software techniques.

Henriksson et al. [5] present the results of a study to adapt the ISO 26262 standard to enable machine learning. Henriksson et al. conducted a qualitative analysis on the challenges faced when developing automotive software which depends on deep learning based on the experience of two experts on functional safety. The authors argue that the requirements that

previously existed on the actual application should be moved to the training phase where the network is created; that model sensitivity is more important than traditional branch coverage; and that test case design must be more thorough. This impact of this research is weakened by the low number of participants.

3 Proposed Solution

This report will present a technique for hazard analysis which is designed to identify gaps in the training data used in machine learning components and thus eliminate one of the primary challenges for existing hazard analysis techniques, as introduced in Section 1.

By highlighting the inadequacies in the training data being used in a machine learning component, the technique presented in this report, a hybrid of HAZOP and Sneak Analysis, will improve the robustness of machine learning components in the safety critical system by reducing the number of scenarios which are not accounted for in training.

HAZOP is a hazard and operability study. HAZOP can be used throughout the different phases of a project and is normally performed by four-to-six people. A HAZOP safety study is normally led by a chairman who is experienced in performing safety studies and one participant will record the findings of the group.

Comprehensive evaluations of a process produced by HAZOP are the result of combining a set of guidewords with parameters systematically. In a system, components that have a meaningful design intent, contribute to system complexity and add potential hazards are chosen as nodes. For each node, guidewords, such as `MORE`, `LESS`, `REVERSE`, `EARLY`, and `AFTER`, are combined with parameters, such as "flow", "temperature", "pressure", and "composition", systematically to identify potential hazards.

HAZOP can be extended with more guidewords and parameters by specific industries and companies, making it a particularly flexible hazard analysis technique. For example, the nuclear industry will often add additional terms which relate to radiation.

Sneak Analysis is a technique developed in the aerospace industry to determine why spacecraft rockets might accidentally fire, or not fire when required. The problem that Sneak Analysis was designed to solve has clear parallels to a machine learning component in an autonomous vehicle, which may respond incorrectly, or not respond at all, to an event.

Sneak Analysis defines six separate paths for error - sneak flow, sneak

indication, sneak label, sneak energy, sneak reaction and sneak procedure or sequence.

- **Sneak flow** describes the unintended flow of information or material from one area to another. This can be a result of human error or equipment failure. In machine learning, sneak flow is most similar to when a model uses part of the input data which is unintended when classifying the input. For example, a machine learning model used in a UAV might be recognising a parked car in an image due to the pavement in training photos, an assumption that would not always hold throughout the world.

- **Sneak indication** describes when indicators for processes are wrong or ambiguous. Indicators that are relied on by machine learning algorithms may not be reliable in some rare real-world scenarios. For example, in an autonomous drone or vehicle, the colour of another vehicle could be very similar to the colour of the sky in some lighting conditions and result in a vehicle assuming that it is safe to continue.

- **Sneak label** describes when there is incorrect or ambiguous labelling of indicators or equipment. In supervised machine learning, where example input-output pairs are provided, ambiguous or incorrect labelling could result in wrong output. For example, the colours of traffic signals might be labelled correctly for one geographical area but the labelling could be incorrect in another which could result in dangerous manoeuvres from an autonomous vehicle.

- **Sneak energy** describes the unintended presence or absence of energy. Traditionally, this is the result of unreacted materials in a process. In a machine learning system, this could describe scenarios where the component is unaware of the state of the rest of the system. For example, during an emergency in an UAV, a machine learning component may decide to perform a manoeuvre which isn't possible due to weather conditions or remaining fuel/battery levels.

- **Sneak reaction** describes unintended reactions and is often the result of unanticipated process conditions. Unanticipated conditions in machine learning components could be the result of irregular events. For example, during Christmas, lighting conditions can change as festive lights are added to town centres; or during Halloween, costumes would be more common and could interfere with pattern recognition.

- **Sneak procedure or sequence** describes when events are in an unintended or conflicting order. Sneak procedure or sequence could occur in machine learning components when one-off situations result in conditions changing. For example, in an automotive system, a road traffic accident might mean that there are temporary traffic indicators which show conflicting instructions to the permanent indicators that are also visible.

In traditional sneak analysis, identification of sneak flows is performed on a piping and instrumentation diagram or tree diagram. Remaining sneak paths are identified through checklists (often referred to as sneak clue lists).

The hybrid approach of HAZOP and Sneak Analysis presented in this report will use an extension of HAZOP, with specialised parameters, to identify sneak flows in each of the six kinds of sneak path described previously.

Like HAZOP, this process will typically be performed by four-to-six people. One participant will lead the process as the "chairman" and should have experience with performing safety studies and another participant will keep records. Higher quality results would result from some participants having experience with traditional hazard analysis process and safety critical systems; and other participants having experience with machine learning techniques and their limitations.

For each of the six kinds of sneak path previously identified - sneak flow, sneak indication, sneak label, sneak energy, sneak reaction, sneak procedure or sequence - the chairman will read the description of the sneak path and provide some examples. Once all participants understand the kind of sneak path currently being investigated, a HAZOP session will be conducted using the parameters from traditional HAZOP analyses in addition to those listed in Table 1 and 2 for the current sneak path kind.

Table 1: Extension to HAZOP Parameters (1)

Sneak Flow	**Sneak Indication**	**Sneak Label**
common irrelevant features common angles insufficient variety	lighting conditions visibility (dust, etc.) adversarial input reflections	locale specific

Table 2: Extension to HAZOP Parameters (2)

Sneak Energy	Sneak Reaction	Sneak Procedure or Sequence
fuel/battery levels	public holidays	route diversions
payload stability	astronomical events	emergency services
structural integrity	roadworks	signal interference
operator attention	climate change	
	power outages	

4 Evaluation

To evaluate the technique discussed in this paper, a hypothetical safety-critical aerospace project which utilises machine learning, described in Section 4.1, is evaluated by a group of four recent computing science graduates after brief familiarisation with the relevant processes, the results of which are presented in Section 4.2.

4.1 System Description & Project Context

The project aims at developing an unmanned aerial vehicle or drone for detecting illegal parking in public town centres. The vehicle is operated without supervision and only within designated large car parking areas.

In normal operation, the drone surveys the target area and sends the license plate number of any cars which enter, exit and park within the area to a central system. Due to the limited flight duration of the drone, it will return to a nearby docking station when available power is low and another drone will launch. Furthermore, the drone can only operate within daylight hours, and will return to docking station when visibility is approaching a threshold.

Automated drone flight and license-plate reading are implemented using machine learning techniques and a set of training data.

The drone is prohibited from flying within five meters of ground level, except when docking, and from flying faster than three meters/second. In addition, the drone can be commanded to return to the nearby docking station remotely via radio.

4.2 Hazard Analysis & Risk Assessment

Throughout the hazard assessment performed, participants were instructed to focus on the parts of the system that were specific to automated operation, in order to reduce complexity.

Initially, participants were asked to determine safety goals for the project from the initial project description, the results are shown in Table 3.

Table 3: Safety Goals

ID	Safety Goal
SG01	Maximum velocity must not be exceeded
SG02	Minimum distance from ground must not be exceeded
SG03	Detection and reaction to relevant obstacles must be ensured.
SG04	Operating area constraints must be followed.
SG05	Low-power states must result in landing in designated areas.
SG06	Response to remote operator intervention must be ensured.
SG07	Poor visibility conditions must result in landing in designated areas.
SG08	Recoverable malfunctions must result in landing in designated areas.
SG09	Unsuitable conditions must result in landing in designated areas.

During the bulk of the session, which lasted 45 minutes, participants produced a list of hazards using only HAZOP analysis and then produced additional hazards using the techniques detailed in this paper.

Finally, participants were asked to share their thoughts on the effectiveness of the technique. Most participants commented that the technique helped them consider scenarios that they would not have otherwise. However, some participants commented that they found a subset of the sneak paths described by the technique confusing or unintuitive.

The complete list of hazards identified are available in Table 4 (without this paper's contribution) and Table 5 (with this paper's contribution, deduplicated).

5 Conclusion

In this report, the state-of-the-art research in integration of machine learning and functional safety in the automotive and aerospace industries was

presented, with a explicit focus on the applicability of any problems and solutions to the aerospace industry.

Furthermore, this report presented a hazard analysis technique extending HAZOP and sneak analysis methods to identify inadequacies in training data used by machine learning components. As shown in the evaluation, quality and breadth of training data can be improved which could result in reduced accidents and failures that occur due to rare real-world scenarios that had not previously been considered.

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete Problems in AI Safety. *arXiv:1606.06565 [cs]* (July 2016). arXiv: 1606.06565.

[2] Bhattacharyya, S., Cofer, D., Musliner, D., Mueller, J., and Engstrom, E. Certification considerations for adaptive systems. In *2015 International Conference on Unmanned Aircraft Systems (ICUAS)* (June 2015), pp. 270–279. ISSN: null.

[3] Brookhuis, K. A., Waard, D. d., and Janssen, W. H. Behavioural impacts of Advanced Driver Assistance Systems–an overview. *European Journal of Transport and Infrastructure Research 1*, 3 (June 2001). Number: 3.

[4] Domingos, P. A few useful things to know about machine learning. *Communications of the ACM 55*, 10 (Oct. 2012), 78–87.

[5] Henriksson, J., Borg, M., and Englund, C. Automotive safety and machine learning: initial results from a study on how to adapt the ISO 26262 safety standard. In *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems* (Gothenburg, Sweden, May 2018), SEFAIS '18, Association for Computing Machinery, pp. 47–49.

[6] IEC/TC 65/SC 65A. IEC 61508-1:2010. https://www.iec.ch/functionalsafety/.

[7] ISO/TC 22/SC 32. ISO 21448:2019. https://www.iso.org/standard/70939.html.

[8] ISO/TC 22/SC 32. ISO 26262-1:2018. https://www.iso.org/standard/68383.html.

[9] Koopman, P., and Wagner, M. Challenges in Autonomous Vehicle Testing and Validation.

[10] Koopman, P., and Wagner, M. Autonomous Vehicle Safety: An Interdisciplinary Challenge. *IEEE Intelligent Transportation Systems Magazine 9*, 1 (2017), 90–96. Conference Name: IEEE Intelligent Transportation Systems Magazine.

[11] Parasuraman, R., and Riley, V. Humans and Automation: Use, Misuse, Disuse, Abuse:. *Human Factors* (Nov. 2016). Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[12] Salay, R., Queiroz, R., and Czarnecki, K. An Analysis of ISO 26262: Using Machine Learning Safely in Automotive Software. *arXiv:1709.02435 [cs]* (Sept. 2017). arXiv: 1709.02435.

[13] Schumann, J., Gupta, P., and Liu, Y. Application of Neural Networks in High Assurance Systems: A Survey. In *Applications of Neural Networks in High Assurance Systems*, J. Schumann and Y. Liu, Eds., Studies in Computational Intelligence. Springer, Berlin, Heidelberg, 2010, pp. 1–19.

[14] Spanfelner, B., Richter, D., Ebel, S., Wilhelm, U., and Patz, C. Challenges in applying the ISO 26262 for driver assistance systems.

[15] Stolte, T., Bagschik, G., Reschka, A., and Maurer, M. Hazard analysis and risk assessment for an automated unmanned protective vehicle. In *2017 IEEE Intelligent Vehicles Symposium (IV)* (June 2017), pp. 1848–1855. ISSN: null.

Table 4: Evaluation Results (before)

ID	Function	Malfunction	Hazardous Scenario or Consequence	S	Rationale	E	Rationale	C	Rationale	A	SG
1	Flight	Rotor breakage due to collision	Drone loses control and crashes	S3	Falling drones could pose serious risk to people or property	E3	People would have a small window to avoid a falling drone	C3	Collision could be unpredictable and fallout is swift	ASIL B	SG03
2	Flight	Control loss due to wind speeds	Drone loses control and crashes	S3	Falling drones could pose serious risk to people or property	E3	People would have a small window to avoid a falling drone	C2	Crashes could come quickly after disruption but not instantly, weather can be anticipated	ASIL B	SG09
3	Flight	GPS signal loss	Drone strays outwith intended zone	S2	Increases likelihood of other hazards	E1	GPS signal is generally reliable and bounds could be hardcoded	C1	Operator can be deployed to fetch drone before further hazards	ASIL A	SG04
4	Remote Communication	Signal interference	Drone continues to fly despite potential upcoming dangerous conditions which could result in unsafe manoeuvre	S3	Unsafe manoeuvre could put people or property at risk	E1	Signal interference is rare	C1	Operators could be deployed to remove interference or reduce distance to drone	ASIL B	SG06
5	Landing	Debris blocking landing location	Drone is unable to land and could crash if power supply runs out as a result	S3	Falling drones could pose serious risk to people or property	E3	People would have a small window to avoid a falling drone	C1	Operators could be deployed to remove debris	ASIL A	SG05
6	Landing	Landing location out-of-range	If drone cannot get to landing zone to charge then it could fall	S2	Falling drones could post serious risk to people or property	E1	Deployment sites would have sufficient landing zones coverage	C1	Operators could be deployed to manually guide drone	ASIL A	SG05

7	Flight	Control loss due to damage from weather	If drone is damaged due to weather then it could react unpredictably or fail in some scenarios	S1	Could result in crash that risks people or property	E2	Falling drones could pose serious risk to people or property	C2	Regular inspection can reduce likelihood of failure	ASIL A	SG08

Table 5: Evaluation Results (after, without including previous results)

ID	Function	Malfunction	Hazardous Scenario or Consequence	S	Rationale	E	Rationale	C	Rationale	A	SG
8	Flight	Reflections from objects interfere with object recognition	Training data does not include blinded images which could impact flight decisions	S2	Could result in collisions	E2	Erratic flight can be hard to avoid but might be obvious before incident	C3	Source of reflections are unpredictable	ASIL D	SG03
9	Flight	Manual operators aren't paying attention	Intervention to prevent other hazards is impaired	S2	Increases likelihood of other hazards	E1	Proper training and policies can prevent hazard	C1	Entirely under control of project	ASIL A	SG06
10	Flight	Dust in the air interferes with cameras	Training data does not include blinded images which could impact flight decisions	S2	Could result in collisions	E2	Erratic flight can be hard to avoid but might be obvious before incident	C2	Weather conditions can be predicted	ASIL B	SG03
11	Flight	Changes in terrain (e.g. snowfall) interferes with object recognition	Training data does not contain data from all seasons	S1	Could render drone inoperable and result in crashes	E1	People would have a small window to avoid a falling drone	C2	Can be avoided in design phase	ASIL C	SG03

12	Flight	Solar eclipse results in sudden poor lighting conditions	Training data does not include zero lighting conditions	S1	Crashes can cause serious damage to people or property	E1	People would have a small window to avoid a falling drone	C1	Solar events are very predictable	ASIL A	SG03

Key: ID: Identifier for hazardous scenario, S: Severity* (S0–S3), E: Exposure* (E0–E4), C: Controllability* (C0–C3), A: ASIL Rating (QM, ASIL A–D), SG: ID of Safety Goal