Are Quantitative Safety Targets for Railways Useful for Disruptive Technologies?

Andrew Gabler and Pravin Hiremath
Acmena Group

Andrew.Gabler@acmena.com.au; Pravin.Hiremath@acmena.com.au

Abstract

In the Australian rail safety context, the national rail safety regulator recommends that rail transport operators establish quantitative safety targets for major projects. Traditionally this is executed through the use of metrics like Fatality Weighted Injuries (FWI), Tolerable Hazard Rates (THR)/Tolerable Functional unsafe Failure Rates (TFFR), and Safety Integrity Levels to demonstrate the achievement of a quantitative safety target. In this paper we will explore whether these traditional metrics are still useful when working with disruptive technologies such as Artificial Intelligence (AI). We will explore the use of AI in a typical signalling system in the context of traditional quantitative analysis techniques and will conclude with a set of observations and recommendations that define the critical role of quantitative analysis in shaping future projects.

Keywords: Artificial Intelligence, Safety Analysis, THR/TFFR, SIL, FWI, SPI.

1 Introduction

The railway industry has long relied on quantitative safety targets to ensure the safety and reliability of its operations. Metrics such as FWI, THR/TFFR, and SIL have been the cornerstones of quantitative safety analysis and risk assessment in the railway industry. These metrics provide a proven mechanism to assess the risk of potential hazards.

In the Australian context, the national rail safety regulator mandates the establishment of quantitative safety targets for major rail projects. This regulatory framework has been effective in managing risks associated with conventional railway systems. However, the advent of disruptive technologies, particularly AI, is challenging the status quo. AI-based systems introduce a new level of complexity and unpredictability that traditional safety metrics may not fully capture.

This paper seeks to explore the relevance and applicability of traditional quantitative safety targets in the context of these emerging technologies. We aim to investigate whether metrics designed for conventional systems can adequately address the unique risks posed by AI and other disruptive technologies. We will work through the application of AI in a typical signalling system and consider how it might impact the use of traditional quantitative safety metrics.

Furthermore, we will evaluate the potential limitations of existing safety metrics and propose recommendations for adapting safety analysis frameworks to better accommodate the nuances of AI-driven systems. The goal is to ensure that the rail industry continues to maintain high safety standards while embracing innovative technologies that promise to enhance operational efficiency, passenger experience, and system safety.

2 Traditional Techniques

Quantitative safety targets have been the backbone of safety assurance in the railway industry for decades. These targets are designed to provide clear, measurable criteria for assessing and mitigating risks. The following sections outline some of the most commonly used traditional techniques: FWI, THR/TFFR, SIL.

2.1 Fatality Weighted Injuries (FWI)

FWI is a composite metric used to quantify the overall risk of harm within a rail system. It combines fatalities and injuries into a single measure, weighting them according to their severity. The formula typically assigns a higher weight to fatalities and a proportionally lower weight to injuries, allowing for a comprehensive assessment of safety performance. For example, the Australian Rail Risk Model (ARRM) (Safe Decisions, 2016) uses a metric where 1 fatality is equivalent to 10 serious injuries or to 200 minor injuries. For example, in train movement-related accidents, such as level crossing collisions, the FWI for all user groups (members of the public, passengers, and workers) is calculated to be 5.7 per year (Whalley, 2024). This figure is derived from an annual average of 19.72 minor injuries, 4.49 serious injuries, and 5.15 fatalities.

FWI is valuable because it provides a straightforward way to compare the safety performance of different systems or projects. By converting diverse safety incidents into a single metric, stakeholders can easily identify areas needing improvement and track progress over time. For example, Figure 1 shows an F-N curve comparing equivalent fatalities (FWI) from rail (Wikipedia, 2024) and road accidents in Australia (BITRE, 2024). In this context, 'F' represents the frequency of accidents per year with 'N' or more equivalent fatalities, and 'N' is the number of fatalities. The F-N curve helps decision-makers evaluate the societal risk associated with each mode of transportation or any other activity over a period (decades or centuries).

FWI data is also useful for understanding the historical likelihood of harm for various types of incidents, and it can serve as a basis for risk ranking when assessing the risk of a hazard.

While FWI remains a valuable lagging indicator, its effectiveness is limited when applied to disruptive technologies. This is because we lack sufficient historical data for these new technologies, which makes it difficult to predict future risks. In contrast, FWI is more reliable

when used with established technologies, where ample data is available.

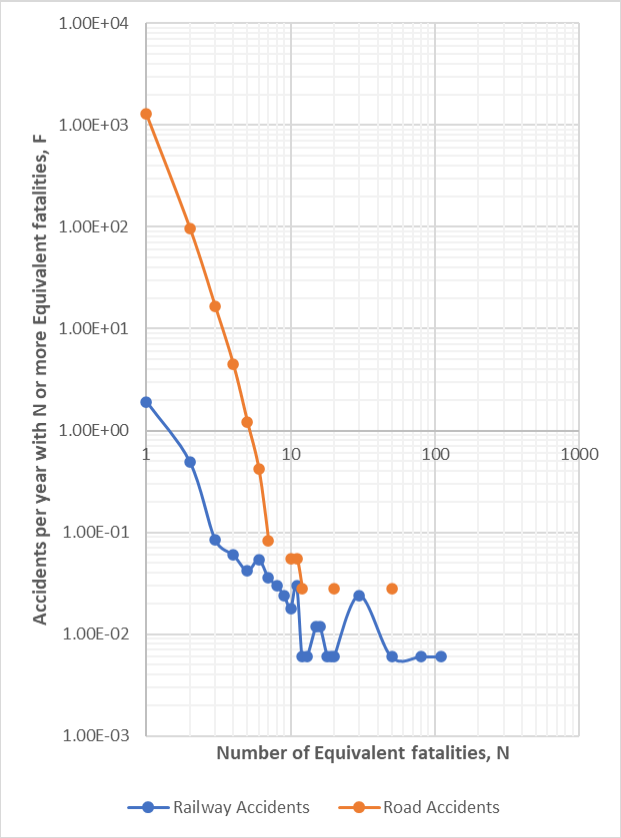

Figure 1: F-N Curve for Rail and Road Accidents in Australia

2.1.1 FWI - Sources

Since FWI is a lagging indicator, its data sources typically include:

1. **Accident and Incident Reports**: Detailed records of accidents and incidents are primary sources for FWI calculations. These reports often include the number of fatalities, serious injuries, and minor injuries, which are crucial for deriving FWI values.
2. **National Safety Databases**: Many countries maintain national databases that compile safety-related data across different transport sectors. For example, in Australia, organizations like the Office of the National Rail Safety Regulator (ONRSR) provide critical data for FWI analysis.
3. **Operational Data from Rail Operators**: Rail operators themselves are key contributors to FWI data. They maintain logs of safety incidents, which are then analysed to assess the safety performance of specific projects or systems.

2.1.2 FWI - Example

The ARRM Risk Profile Report (Whalley, 2024) calculates FWI using data obtained from ONRSR Rail Safety Reports, ensuring that the risk assessment is based on reliable and comprehensive safety data. For instance, in a reported incident involving 2 fatalities, 5 serious injuries, 10 reportable minor injuries, and 15 non-reportable minor injuries, the FWI would be calculated as follows:

- Fatalities: 2 fatalities × weight of 1 = 2.0
- Serious Injuries: 5 serious injuries × weight of 1/10 = 0.5
- Reportable Minor Injuries: 10 reportable minor injuries × weight of 1/200 = 0.05
- Non-Reportable Minor Injuries: 15 non-reportable minor injuries × weight of 1/1000 = 0.015

The total FWI for this incident would be:

FWI = 2.0 + 0.5 + 0.05 + 0.015 = 2.565

This calculated FWI value provides a standardized measure of the incident's severity, contributing to the overall risk profile and helping stakeholders identify areas that may require safety improvements.

2.2 Tolerable Hazard Rates (THR) and Tolerable Functional unsafe Failure Rates (TFFR)

THR represent the maximum allowable frequency of hazardous events within a rail system. These rates are established based on historical data, regulatory guidelines, and industry good practices. THRs are used to ensure that the risk of specific hazards remains within acceptable limits, thus protecting passengers, employees, and the public. Noting that it is often a challenge within the railway industry for a THR to be assigned by the railway operators at a railway level; however, the establishment of a THR for a specific sub-system (e.g. an interlocking) is not uncommon. As described in industry standard EN50126-2 (CENELEC, 2017) a THR should be allocated down to a TFFR for each safety function performed by the system.

The use of THR/TFFR involves identifying potential hazards, estimating their likelihood, and determining whether these estimates fall within the predefined tolerable limits. This process requires detailed hazard analysis and risk assessment, often employing techniques such as Fault Tree Analysis (FTA) and Event Tree Analysis (ETA). An illustrative example is analysing the hazard of a 'Signalling system failing to detect the presence of a train.' This can be managed using FTA to decompose the hazard into its contributing factors and allocate THR down to a TFFR as described in Section 2.2.2.

2.2.1 THR/TFFR - Sources

THR/TFFRs are calculated using an FTA using the failure rate of each ‘leaf event’ of the fault tree. For equipment these are generally predicted using various reliability prediction methodologies such as MIL-STD-217F (MIL-STD-217F, 1991), Telcordia (Telcordia, 2013), or IEC 61709 (IEC, 2017) for individual electronic components (e.g. resistors, capacitors, transistors) and other industry databases such as NPRD (NPRD, 2016) and EPRD (EPRD, 2014). When the reliability of a human is considered in calculating the THR/TFFR, various other methods such as the Railway Action Reliability Assessment (RARA) are used (Pauley, 2023).

2.2.2 THR/TFFR - Example

A THR can be allocated down to TFFR as in the simple example show in Figure 2.

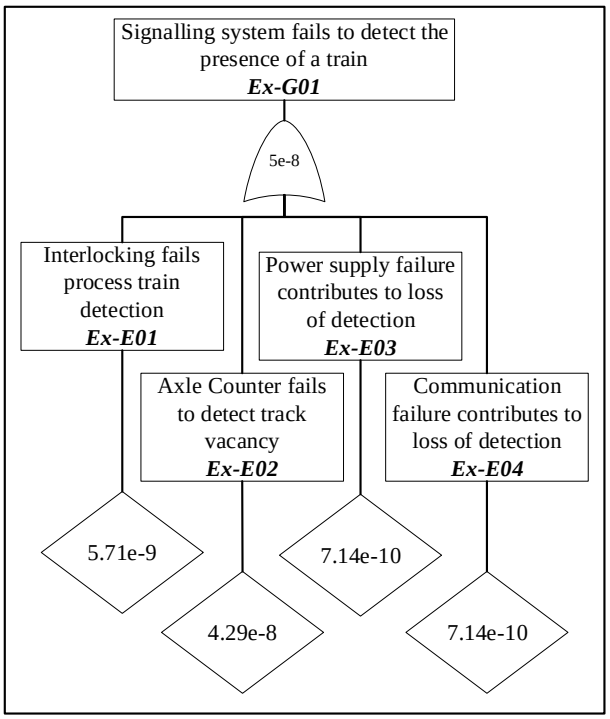

Figure 2: THR/TFFR allocation example

In this example, a THR of 5e-8 events per hour is allocated to this undesired event (across the full signalling system). The signalling system is then determined to consist of ~8 interlocking/object controllers associated with managing this function and 60 axle counter units. As the interlocking and axle counter systems fail safe and communicate with safety protocols, the power supply and communication systems are each allocated a single unit.

To achieve a THR of 5e-8, a TFFR of 7.14e-10 events per hour is allocated to each unit. This allocation is calculated as follows:

- Total number of units: 8 + 60 + 1 + 1 = 70 units
- TFFR per unit = THR / Number of units
- TFFR per unit = 5e-8 / 70 = 7.14e-10 events per hour.

Therefore, the THR allocation for each subsystem is:

- Interlocking (8 units): THR = 7.14e-10 x 8 = 5.71e-9 events per hour
- Axle counters (60 units): THR = 7.14e-10 x 60 = 4.29e-8 events per hour
- Power supply (1 unit): THR = 7.14e-10 events per hour
- Communication equipment (1 unit): THR = 7.14e-10 events per hour.

Using the sources described in Section 2.2.1 each of the undeveloped events will be developed further to demonstrate that the selected products will achieve the allocated TFFR.

2.3 Safety Integrity Levels (SIL)

Safety Integrity Levels (SIL) are a measure of the reliability and robustness of safety-related systems. SILs are defined within international standards, such as IEC 61508 (IEC, 2010), EN 50126 (CENELEC, 2017), EN 50128 (CENELEC, 2020) EN50716:2023 (CENELEC, 2023), and EN 50129 (CENELEC, 2018) and range from Basic Integrity (BI) (least stringent) to SIL 4 (most stringent). Each level corresponds to a specific probability of failure, with higher levels requiring more rigorous design, implementation, and testing procedures.

SIL assessments involve evaluating the entire lifecycle of a safety-related system, from initial design through to operation and maintenance. This includes identifying safety functions, performing risk assessments, and implementing appropriate safety measures to achieve the desired integrity level. Primarily SIL is used to provide a level of robustness to the quality processes used to develop complex electronic systems. However, the application of SIL is not always necessary. If a system can be fully understood and evaluated using techniques such as THR, a SIL allocation may not be required. This scenario typically applies to simpler systems with limited complexity and well-defined failure modes.

However, it is important to note that complex electronic systems, especially those with software components, generally do necessitate a SIL allocation. This is because:

1. Software complexity: Software-based systems often have numerous potential failure modes that are difficult to predict and evaluate comprehensively using simpler techniques.
2. Systematic failures: Complex systems are more prone to systematic failures, which SIL processes help to mitigate.
3. Reliability requirements: Higher reliability requirements for safety-critical functions often necessitate the rigorous processes associated with SIL.
4. Regulatory compliance: Require SIL allocation for complex safety-related systems to meet safety standards and regulations.

In practice, the decision to apply SIL should be based on a thorough assessment of the system's complexity, criticality, and the adequacy of alternative evaluation methods (e.g. compliance with alternative standards). For complex electronic systems with software components, SIL allocation remains a crucial part of ensuring safety and reliability.

Note that EN50716 (CENELEC, 2023), which has recently been released, now addresses the use of AI in SIL rated systems and marks it as “Not Recommended” (see Table A.3).

3 Disruptive Technology - AI

The integration of AI into the railway sector marks a significant shift in how safety and operational efficiency are managed. AI, defined as ‘a computerized system that is able to perform physical tasks and cognitive functions, solve various problems, or make decisions without explicit human instructions’ (Kaplan and Haenlein, 2019), presents both opportunities and challenges for the industry. Its potential to optimize railway operations, enhance safety and security, and improve customer service is increasingly recognized (Burroughs, 2019).

The transformative impact of AI on capacity management, maintenance, and passenger flow prediction is already evident. Notable examples include Toshiba's train timetabling AI in the UK and SNCF's predictive

maintenance systems in Paris (Fragnelli and Sanguineti, 2014; SNCF, 2020). These systems showcase the practical applications of AI in optimizing and streamlining various aspects of railway operations.

Between 2010 and 2022, 141 studies have explored the potential applications of AI in railways (Ruifan Tang et al., 2022). Key areas identified include:

1. **Maintenance and Inspection**: AI is used for defect detection, fault diagnosis, failure prediction, and planning maintenance activities. Autonomous maintenance systems leverage AI to identify issues before they lead to significant disruptions, enhancing the reliability and safety of railway operations.
2. **Safety and Security**: AI contributes to risk management, accident analysis, anomaly detection, and disruption management. It plays a crucial role in safety-critical applications, such as collision avoidance systems (Wohlfeil, 2011) and critical software for train control, requiring rigorous safety analyses to ensure these systems perform reliably.
3. **Autonomous Driving and Control**: AI technologies optimize energy use, enhance intelligent train control, and manage train trajectories.
4. **Traffic Planning and Management**: AI aids in rescheduling, delay prediction, capacity management, and train timetabling. It also assists in complex tasks like shunting, routing, stop planning, and track design, helping to optimize the overall flow of railway traffic and reduce bottlenecks.
5. **Passenger Mobility**: AI's capability to predict passenger flow helps manage crowds and improve the passenger experience. Accurate predictions enable better resource allocation and service planning, enhancing overall efficiency and safety.

Of the 141 studies, Ruifan Tang et al., 2022 highlight that 57% focused on maintenance and inspection, 25% on traffic planning and management, 8% on safety and security, 5% on autonomous driving and control, and another 5% on passenger mobility.

The maintenance and inspection domain is notably advanced in AI applications compared to others (Ruifan Tang et al., 2022). However, real-life AI integration in safety-critical areas like autonomous driving and traffic rescheduling remains limited due to challenges such as non-stable AI behaviour, opacity of some AI approaches, and data scarcity. Most AI applications in safety, security, and real-time scheduling are still theoretical, with few practical implementations or commercialization cases listed in the following section.

3.1 Adoption of AI in Railway

The Dubai Roads and Transport Authority (RTA) is trialling AI and simulators to reduce passenger waiting times and manage crowds during rush hour at busy metro stations. This initiative aims to develop a smart, interactive system that adjusts transit timings based on real-time demand patterns to prevent overcrowding. By analysing data from nol cards (similar to Myki cards in Victoria) and metro demand algorithms, the AI model simulates train journeys and proposes specific boarding periods. This approach has successfully reduced congestion by 40-60% and decreased waiting times to 30 minutes. The results are currently under review, with potential for full implementation to enhance the customer experience (ITP Staff, 2021).

In Copenhagen, the S-bane network is being upgraded to become the world's longest automated urban railway using Siemens' GoA4 technology, which enables fully driverless train operations. This AI-driven system will enhance efficiency, increase train frequency, and maintain high punctuality by automating tasks like speed control, braking, and incident management. The upgrade is set to be completed by 2030, ensuring uninterrupted service during the transition to driverless trains (CDOTrends editors, 2024).

While the Dubai Metro's AI trial is not safety-critical, it demonstrates RTA's exploration of AI's potential to address operational challenges, although no further progress has been reported since 2021. On the other hand, the Copenhagen Metro's AI application is clearly safety-critical, involving signalling systems set to be operational by 2030. However, this use of AI has only been reported in the media (CDOTrends editors, 2024), and Siemens' official website does not confirm the same (Siemens Mobility GmbH., 2024), leaving some uncertainty about the actual deployment of AI in the railway industry.

3.2 Safety Assurance Challenges

The limited application of AI in safety-critical systems reflects a lack of confidence in its behavioural reliability. However, AI integration in safety-critical applications such as intelligent train control, collision avoidance, and the development of critical software systems is being explored, highlighting the need for rigorous safety analyses. This necessitates the development of processes or techniques to build confidence in AI's use within safety-critical contexts.

Traditional safety assurance techniques, such as Fatality Weighted Index (FWI) and Tolerable Hazard Rate (THR)/Tolerable Failure Frequency Rate (TFFR), which are discussed in Section 2, offer some insights but are not entirely sufficient for AI-driven systems. While FWI remains valuable as a lagging indicator for monitoring the safety performance of AI-powered rail systems during operation, it does not provide the necessary confidence for introducing AI into safety-critical environments.

THR/TFFR, typically calculated for the hardware components of electrical, electronics, and programmable electronics systems, may still be relevant for AI-powered systems, considering AI is software that operates on such hardware. However, unlike traditional software, AI exhibits non-deterministic behaviour, adaptability, and the potential for unforeseen failure modes, making it challenging to apply conventional safety assurance methods like Software SIL defined in EN 50128 (CENELEC, 2020).

To effectively manage the safety of AI-driven systems, there is a need to develop new safety targets and frameworks that can accommodate the unique properties of AI. This includes addressing issues related to the

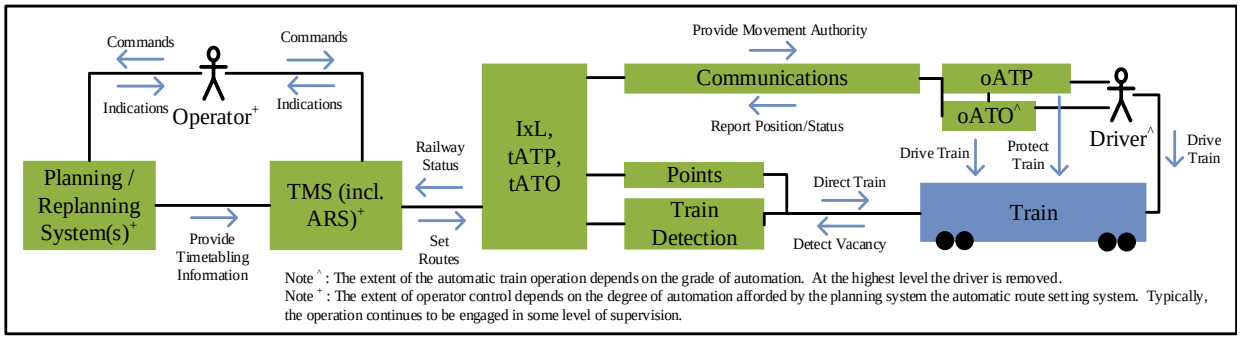

Figure 3: 'Typical' Signalling System

transparency of AI decision-making processes, the ability to predict and mitigate AI system failures, and ensuring that AI systems can be safely integrated with existing railway infrastructure and operations. As the industry continues to adopt AI technologies, defining appropriate safety targets and standards will be crucial in ensuring the safe and reliable operation of rail systems.

4 Case Study – Typical Signalling System

Considering the active adoption of AI into rail industry as briefly described in Section 3 there is a need to consider whether our traditional techniques of quantifying risk described in Section 2 are suitable for the task. This section will explore the various potential uses of AI within a typical signalling system and whether the traditional quantitative techniques still have a place in assessing AI.

4.1 Typical' Signalling System

The 'typical' signalling system to be considered is described in Figure 3. It consists of a typical system consisting of the following key sub-systems:

1. Planning and Replanning System: that plans how the available trains can achieve the timetable accounting for disruptions in the rail operations.
2. Traffic Management System (TMS) including Automatic Route Setting (ARS): which uses inputs from the planning systems and direction from the operator to generate appropriate route requests for the trains to operate to the target timetable.
3. Interlocking (IxL): which checks the route requests from the TMS/ARS to confirm they are safe (i.e. no conflicts between train movements) prior to setting the route and sending a movement authority to a train.
4. Trackside / Onboard Automatic Train Protection System (tATP/oATP): that monitors train position and prevents a train from exceeding its given movement authority.
5. Trackside / Onboard Automatic Train Operation (tATO/oATO): that depending on its grade of automation drives the train to achieve the given timetable within its set movement authority.

From experience each of these systems are typically assigned a THR/TFFR and SIL as described in Table 1.

Table 1: THR/TFFR and SIL Allocation to System

System	THR/TFFR Allocation [h^{-1}]	SIL Allocation	Comment
1	-	-	Not generally assigned a SIL. If a SIL were to be assigned it would be Basic Integrity
2	$10^{-7} \leq$ TFFR < 10^{-4}	Basic Integrity - SIL2	Depends on the specific function
3	$10^{-9} \leq$ TFFR < 10^{-8}	SIL4	
4	$10^{-9} \leq$ TFFR < 10^{-8}	SIL4	
5	$10^{-7} \leq$ TFFR < 10^{-6}	SIL2	Noting some functions may be considered Basic Integrity.

4.2 Uses of AI

As seen in Section 3 there are a number of ways AI is being used in the rail industry. For the purposes of this paper these will be split into four levels. These levels are:

A. Development Aid
B. Non-Safety
C. Safety Related and
D. Safety Critical.

In addition to these four levels each can be considered based on the level of AI that could potentially be applied, which for this paper we will limit to non-learning (i.e. reactive machine AI) and learning (i.e. limited memory AI).

Table 2 provides some examples of what might be considered in these levels.

Table 2: Examples

Level	Example	Description
A	Reviews	Use of AI to review documents, code, etc.
A	Test Environment	Use of AI to generate scenarios, automate testing and support other verification and validation activities
A	Code Generation	Use of AI to generate code from some level of specification
B	Traffic Planning and Management	Use of AI to aid in rescheduling, delay prediction, capacity management, and train timetabling
C	Automatic Route Setting	Use of AI in the automatic routing of trains based on the planned timetable

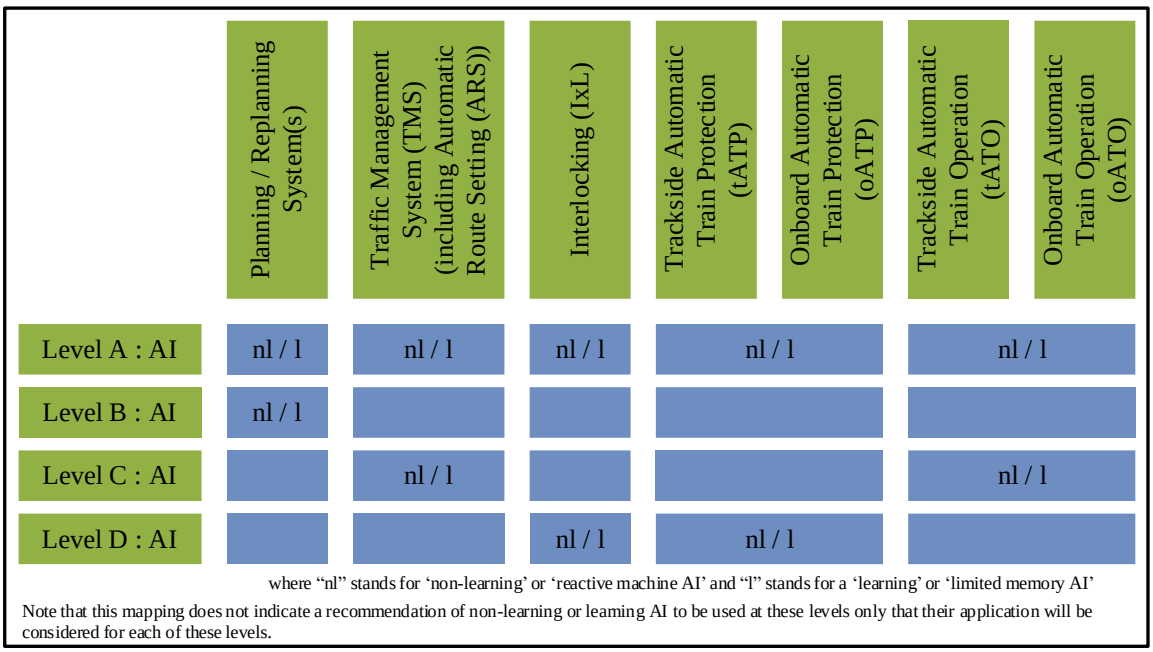

Figure 4: Mapping

Level	Example	Description
C	Automatic Train Operation	Use of AI to optimize energy use, enhance intelligent train control, and manage train trajectories
D	Interlocking	Use of AI to determine if it is safe to send a movement authority to a train
D	Automatic Train Protection	Use of AI to monitor and prevent a train from exceeding its movement authority and prevent collisions

The potential application of both learning (l) and non-learning (nl) AI will now be considered across the five sub-systems identified in Section 4.1 as depicted in Figure 4.

4.3 Use of AI in Development Aids

The use of AI as an aid in the development of a safety system will need to be assessed on a case-by-case basis. As with any tool used, the reliance on that tool needs to be assessed for its impact on the safety of the system being developed. For example, in the development of a traditional safety related/critical software-based system the choice of the language used and the compiler selected needs to be considered with care and justified. This is typically done by reviewing the pedigree of the language and compiler, following advice given in standards such as EN50128 (CENELEC, 2020) (note Table A.15), assessing the risks the tool could contribute to, and ensuring controls are put in place to prevent/mitigate those risks. The same assessment would be needed for the use of AI as a development aid.

First let us consider the use of AI as a tool in reviews. On a recent project we have found that using a combination of scripting and non-learning AI to review through 1000s of pages of design documents to consolidate a meaningful list of assumptions made in those documents was quite an effective tool. Now in doing this activity there is always the question as to whether the AI actually captured all the items of concern and did not hallucinate (i.e. make things up). A spot check revealed that the AI was able to process the thousands of pages in just a few hours, a task that would have taken an intern 2 to 3 weeks to complete. In this case, this review was in addition to the traditional processes of design reviews with subject matter experts and stakeholders and so did not play a significant role in assessing the safety of the system; however, it does illustrate the potential usefulness of AI to augment the reviews done by engineers in development of safety critical systems. It can take a task that was previously unreasonable to perform due to the required manpower and turn it into something that could be done relatively quickly and efficiently. Emphasizing the output still needed a review by an engineer for a useful conclusion. Effectively, at this point in time the use of AI in review could be considered analogous to having the work performed by a highly productive but slightly dim intern. As such the outputs need to be reviewed and checked for validity by the responsible and competent engineer. So, while AI can provide greater efficiencies in some areas it is unlikely to be able replace competent engineers until confidence can be gained that the AI can achieve some level of competency in the specific task as might be done in image recognition tasks where the non-learning AI is proven through extensive to achieve a Safety Performance Indicator (SPI) (Koopman, 2022).

Second, consider the use of AI in a test environment. AI might be used to develop test cases for a bit of software or system, or it might be used to generate multiple scenarios to test a system. If this is the case, again the output will need be reviewed by a competent engineer to determine the usefulness of the scenarios. If used well the AI might be able to quickly provide a comprehensive suite of test cases; or, if used poorly it may give a useless bit of gibberish.

Third, consider the use of AI in code generation. Already we hear of many people who go to Chat GPT and ask it for code to perform a task in MS Excel and code is

Table 3: Development Aid Assessments

Aid	Example Failure Modes	Example Controls	Discussion
Requirements Management Tool	• Corruption of requirements (incorrect, invalid, incomplete, missing…) • Corrupted linking (contributing to missing requirements)	• Formal Reviews • Requirement Baselines exported • Limited automation • Vertical slice analysis • Verification and Validation activities	In the future AI is likely to help analyse and draft new requirements. This will introduce new avenues for the tool to contribute to the failure mode and increases the need for full formal reviews.
Change Management Tool	• Corruption of change requests (incorrect, invalid, incomplete, missing, …) • Incorrect workflows (delays in processing)	• Reviews by the Change Control Board • Corruption unlikely to produce meaningful information • Change Request goes through multiple layers of check (impact analysis, developer review, implementation, verification and validation) • External monitoring	In the future it is conceivable that AI will be used to raise, triage, and even assess change requests. This means it may be harder to catch corruption caused by the AI which may generate reasonable but possibly incorrect information.
Configuration Management Tools	• File Corruption (spurious, incomplete, incorrect, invalid, missing,.) • Incorrect document/software inclusion in a release	• Formal Reviews • Corruption unlikely to produce meaningful information • Configuration Audits/Reviews	AI may be used to support reviews and audits in configuration management. Care should be taken so that the AI does not corrupt or incorrectly interpret the data captured.
Compiler Tools	• Incorrect Code compiled • Additional code is added • Some code fails to be compiled	• Validation of compiled code on target hardware • Multi-level testing • Review known issues and bug with tool to evaluate if impacts safety	While compilers are unlikely to include AI, there are plenty of tools available that use AI to detect bugs, review code, document code, refactor code and, generate code. They can also explain code which contributes to the example failure modes and require diligence to review / check prior to use and may require some proving in and of themselves.
Software Test Tools	• Incorrect/Erroneous operation – fail to detect errors in code or spurious error detection	• Test Specification and Results are subject to review • Multi-level testing • Review known issues and bug with tool to evaluate if impacts safety • Code reviews • Unit, component, module testing	There are already tools available using AI to test software. While this when used in addition to (or prior to) traditional methods may add significant benefit. Care needs to be taken to monitor the reliance made on one test method.

generated to do that task. It may not be the most elegant or efficient code, but it works and is a great enabler for the engineer doing the work. Again, the engineer will need to review and understand the code or at the very least review the output to ensure the data produced is correct, but it is not a big step from using AI to assist in developing MS Excel macros to generate safety critical software. If done rigorously and with checks and balances, an argument may be made to justify its use; however, careful control would be required to avoid undesired outcomes.

So how do we traditionally control the risk of development aids introducing errors into our safety critical systems? It is by assessing the risk any particular aid may introduce and placing controls around it to prevent/mitigate the risk. Using AI as a development aid would be no different. Table 3 explores some typical tools, the risk they pose, and controls put in place to prevent/mitigate those risks (noting that AI could play a part in any or all of these tools). What we have considered thus far has been focused on non-learning AI; however, similar can be said for the use of a learning AI though greater care needs to be taken to ensure that the output doesn't negatively change over time. A learning AI has the potential to improve its performance but at the same time it runs a higher risk of losing its performance as well. Which at this time comes back to treating the use of AI as a highly productive but potentially

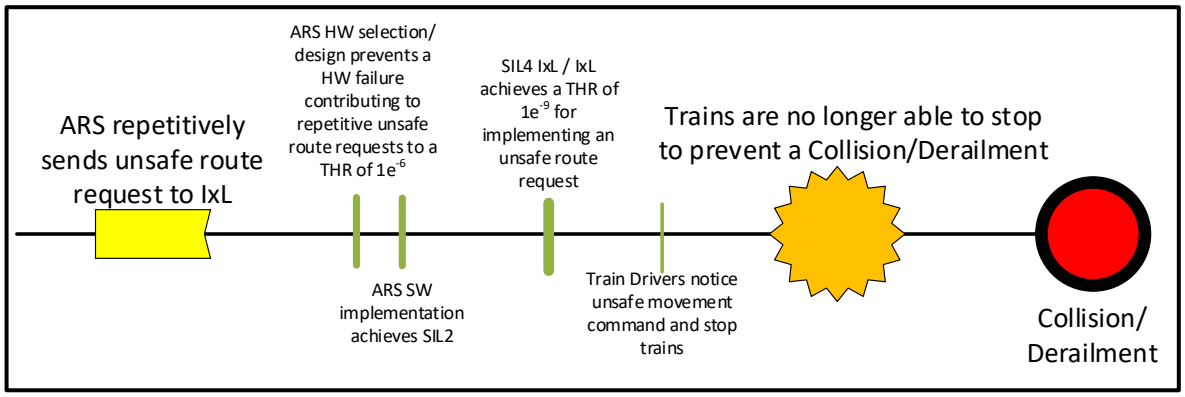

Figure 5: Simplified ARS Threat Barrier Diagram

dim intern. Everything it produces needs to be checked. Noting the extent of that check will depend on the criticality of the task assigned.

4.4 Use of AI in Non-Safety Systems

In the railway context considered in Figure 3 the example of a non-safety system is a planning/replanning system. As discussed in Section 3 the railway industry is already adopting the use of AI in these type systems. From a safety analysis perspective given that these systems have been deemed as non-safety traditionally they are not assigned a target FWI, THR, or SIL. Provided that such systems remain isolated from real world by safety related and safety critical systems any failure modes should be prevented from contributing to a hazard.

However, assuring the use of AI in a non-safety system to ensure a high-quality system is a possible test bench for exploring how non-learning and learning AI might be assured for a safety related or safety critical system. For example, in a re-planning AI system where the AI reviews the status of the railway, predicts train paths and re-schedules trains to achieve optimal paths that best adhere to the original timetable. While the AI must consider interlocking logic to propose feasible solutions, it is important to note that the actual implementation of these plans is carried out by TMS/ARS as shown in Figure 3. The TMS/ARS, rather than the interlocking system, is primarily responsible for preventing gridlock situations.

The AI planning system, in conjunction with the TMS/ARS, can be designed to incorporate interlocking logic to avoid proposing infeasible routes. This approach would help prevent train route queuing for following trains. However, it is crucial to maintain a clear distinction between the planning function and the safety-critical interlocking function.

To evaluate the effectiveness and safety of such an AI system, we could monitor, review, and test its performance. This process could involve assessing how often the AI suggests timetabling solutions that are both optimal and safe, essentially creating a safety performance indicator for the planning system. This approach could serve as a valuable test bench for exploring how both non-learning and learning AI might be assured for safety-related or safety-critical systems in the future, while maintaining the critical safety functions within the appropriate systems.

4.5 Use of AI in Safety-Related Systems

The ‘safety-related’ sub-systems considered in Figure 4 include a Traffic Management System (TMS) (or Train Control System) and an Automatic Train Operation (ATO) system. Depending on the function being performed, elements of these systems may be considered as ‘non-safety’ (i.e. Basic Integrity) and the observations of Section 4.4 apply. This is often the case for functions such as standard route setting controls in a train control system which are fully protected by the ‘safety-critical’ interlocking. Whereas other functions such as an axle counter bypass or vital blocking may be considered ‘safety-related’ and assigned a safety integrity level of SIL1 or SIL2. A similar pattern can be found in ATO systems where functions are fully protected by a corresponding ATP.

As noted in Section 3, a level of AI is already being used in ATO systems to improve the efficiency of train movements; however, again the safety argument associated with such a use is to consider those functions as ‘non-safety’ and either rely on the oversight of a driver or a higher integrity ATP. This is the same where AI is currently being introduced into a TMS.

However, let us consider a case where the TMS implements an AI powered ARS. A hazard – ARS repetitively sends unsafe route request to Interlocking (IxL) – that any ARS faces is depicted in Figure 5.

The concern is that if an ARS continues to send unsafe route requests, then based on pure probability eventually the ‘safety-critical’ high integrity IxL will accept one of those unsafe routes and allow it through. As shown in Figure 5 one way this has been handled is to implement the ARS to SIL2 and the design the hardware so that the THR of sending repetitive request is less than $1e^{-6}$ events per hour due to a hardware failure. But what if the ARS is powered by AI? While a significant amount of the processes and procedures in EN50128 (CENELEC, 2020) will continue to be applied, certain requirements be unable to be achieved. For example, Clause 7.5.4.3 “The Software Source Code shall be readable, understandable and testable” will not be achievable. One of the key elements of ‘AI’ based on techniques such as ‘deep learning’ is that how it ‘works’ is hidden and as such an AI could not be developed under the current EN50128 framework. However, it may be possible to develop the software that calls the ‘AI’ function to the standard and assign a SPI

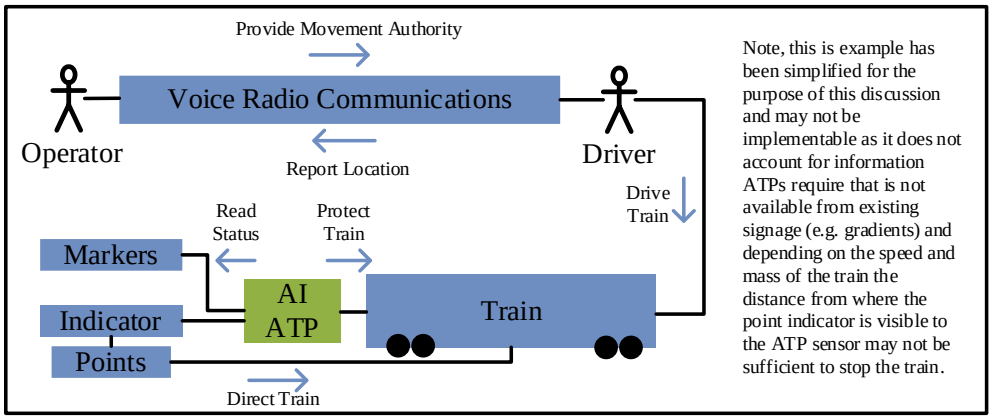

Figure 6: AI ATP Example

(Koopman, 2022) to the 'AI' function. For example, a SPI "the ARS AI shall send an unsafe route command less than once per 1000 hours of operation (i.e. $1e^{-3}$)" may be assigned. The challenge then becomes formulating a test that would be sufficient to give statistical confidence in the AI achieving this SPI. Assigning a SPI that is demonstrated prior to deployment assumes a 'non-learning' AI. If a 'learning' AI was implemented, then the SPI would need to be continually monitored to ensure the SPI is achieved.

An alternative would be to independently (logic independent from the SIL4 IxL) create a second SIL2 interlocking in the TMS that all 'AI ARS' requests are passed through as has been done in the past with traditionally developed TMSs. Another alternative could be to have a SIL2 function in the TMS independently monitor for a violation (i.e. route rejection by the IxL) and when more than one has been detected in 24 hours the ARS is disabled. Noting that disabling the ARS simply shifts the risk back to the human train controllers which may not be the safest outcome. Each of these alternatives leads to a system architecture where the 'AI ARS' is no longer allocated a 'safety' function and then can be treated as a 'non-safety' system. However, both come with their own complications which has led industry to avoid such solutions for traditionally developed systems.

4.6 Use of AI in Safety Critical Systems

We have considered how we might apply traditional quantitative methods to the use of AI in development aids (Section 4.3), non-safety systems (Section 4.4), and safety-related systems (Section 4.5). Now we turn to 'safety-critical' systems. As described in Table 1, Automatic Train Protection (ATP) and Interlocking (IxL) are traditionally considered as 'safety-critical' sub-systems. The authors are not aware of any ATPs or IXLs currently implementing AI and in the traditional sense of an ATP or IXL it is difficult to envision a benefit of implementing a traditional ATP or IXL using AI. However, it is plausible to consider an 'ATP' being proposed for use on a simple Train Order Working (TOW) railway that uses image recognition to read the existing mileage markers for positioning and the status of point indicators to bring the train to a stop prior to the points if in an incorrect position. Figure 6 shows such a system.

This AI ATP example is similar to that discussed in Wohlfeil, 2011. Depending on the railway it could be argued that implementing such a simple AI ATP could 'improve' the existing level of safety and as such may be implemented without demonstrating the achievement of any safety target. However, if such an AI ATP was to be relied upon as 'safety-critical' then an objective measure of safety performance is needed. As done with existing ATPs the physical hardware can be evaluated to show that it achieves a THR of 1e-9 events/hour and any traditional software surrounding the AI functionality in the ATP could be developed to SIL4. The AI would need to be assigned a suitable Safety Performance Indicator (SPI) (see Section 7.1.4 of Koopman, 2022) or some other metric that can be measured and monitored to determine the ongoing performance of the AI (especially if the AI continues to learn).

As described in Koopman, 2022 an SPI can be either a leading or lagging indicator. In this context a leading indicator may be the number of times the AI fails to recognise a point indicator or a marker board, and a lagging indicator might be the number of times the AI fails to enforce a movement authority or prevent the train from incorrectly traversing a set of points. These might be formulated as:

- Leading: The ATP AI shall have a false negative rate of correctly interpreting a marker board no worse than 0.001% over any 1/3rd of a second time interval.
- Lagging: The ATP AI shall fail to enforce a movement authority no less than 1 time per 100 million train kilometres travelled.

What these SPIs provide is a mechanism to measure the performance of the ATP AI against. The leading indicator example is measurable and testable in simulation and in real world testing. For example, a system could be installed (but not connected) on a train running a specific track and the accuracy of the detection could be monitored against a known database of the markerboards along the line. Likewise, another SPI might be the number of times the ATP AI has a false positive and detects a marker board that does not exist.

It is relevant to note that both these SPI's can be derived from or correlated to our traditional metrics of THR and

FWI. For example, an SPI of 0.001% over 1/3rd of a second correlates to a TFFR of ~9.26e-10 events per hour and failing to enforce an authority per train kilometres travelled can be correlated to how many fatalities or injuries it contributed to.

One of the challenges with a leading SPI is determining how to gather sufficient statistically relevant data to be able to make a substantiated claim that the system achieves the target prior to putting the system into production. This, however, is not an entirely new challenge. In traditional ATPs the validation of the selected braking curve algorithms used face a similar challenge, especially for freight consists of which may have vary in weights and characteristics from one trip to another. So, while it is a challenge it is surmountable.

The safety performance of the system should be subjected to continuous monitoring and evaluation. In traditional systems this is done via incident reports of events and root cause investigations to determine contributing factors and whether the failure necessitates a recall or product safety notice. For a leading SPI like the example provided, this could potentially be actively monitored in the field. For example, if the ATP system has several cameras, any disagreements can be flagged and evaluated as to whether there was a violation. Continual monitoring is especially crucial if the AI system is continuing to learn. Just as human operators are required to do refresher training so should a learning AI system be required to undergo re-certification to ensure it continues to perform safely.

Another challenge to consider with the application of AI into a system such as an ATP is the number of potential false positives that may occur. For example, it is conceivable that a future ATP may include an AI obstacle detection function using cameras to view the track ahead and recognise obstacles to bring the train to a stop before it hits the obstacle (cameras mounted on the vehicle or trackside, communicating with the ATP). What if such a system is introduced and it regularly stops the train when there is no obstacle? While a stopped train is typically considered to be in a safe state there are areas on a rail network where it may be unsafe to stop. On a freight or coal network these may be long/steep grades that the trains cannot restart on if they come to a stop, or it may be a viaduct, bridge, or tunnel for a passenger train. Therefore, it is important to not only assess the correct function of the AI system but also consider the additional unintended risks that it may introduce.

5 Conclusion

In conclusion our traditional quantitative safety metrics will continue to have a place in railway safety analysis amidst the introduction of disruptive technologies. Traditional techniques will continue to:

1. Be required to monitor the safety of the rail industry (i.e. measuring performance of the railway in FWI will continue to be important as disruptive technologies are introduced).
2. Be used to establish the tolerable risk targets (THR) regardless of whether a system uses disruptive technology or not.
3. Be used to assure traditional systems that surround and interact with AI powered systems have been developed to an appropriate level of rigour (SIL).

Having concluded that the traditional metrics are still useful does not mean that they are completely up to the task of assessing disruptive technology such as AI. New metrics and guidance will be required.

The following observations attempt to provide key observations and recommendations to assist in tackling this challenge.

1. When using AI as a development aid, the AI should be treated as a highly productive but potentially inexperienced assistant. Everything it produces needs to be checked. The more critical the item the more thorough checking is required. AI can make a good engineer more capable and efficient but it inherits the risk of over reliance hence suitable cross checking and commensurate review is required in proportion to the criticality of the activity.
2. Before using AI, or any tool, as a development aid it is critical to first understand what risks it poses to the safety system you are developing. AI introduces new failure modes that need to be considered and may introduce unintended consequences if not adequately considered.
3. It is recommended that standards continue to be updated to provide guidance on the use of AI in software development (for example as EN50716 (CENELEC, 2023) has been). This applies to both the use of AI in software development environment and also the techniques and methods required to provide a level of assurance when using AI as part of the system developed. Even if the current stance is at ‘not recommended’ as done in EN50716 (CENELEC, 2023).
4. The industry should continue to explore and let disruptive technologies mature in the ‘non-safety’ space. Clever system architectures can transform potentially ‘safety-related’ AI system into ‘non-safety’ components by incorporating deterministic safety controls, which can be evaluated using current techniques and measures.
5. When using AI in safety-related or safety-critical applications it is recommended that additional metrics such as an SPI is specified and established with a measurable and demonstratable target. This target can be correlated to existing metrics such as the systems THR or TFFR and when demonstrated integrated with the traditional safety approaches.
6. Note that new metrics, such as an SPI, brings with it additional challenges of how to test it to achieve a statistically relevant result and introduces a need for more active monitoring that the AI system continues to achieve the SPI. This is especially true for learning AI systems.

6 References

BITRE (2024): Australian Road Deaths Database (ARDD). Available at:

https://www.bitre.gov.au/statistics/safety/fatal_road_crash_database [Accessed 22 August 2024].

Burroughs, D. (2019): The future of intelligence is artificial. Available at: https://www.railjournal.com/in_depth/future-intelligence-artificial [Accessed 9 August 2024].

CDOTrends editors, (2024): Copenhagen's Trains Ditch the Driver. Website, 02 May 2024. Available at: https://www.cdotrends.com/story/3967/copenhagens-trains-ditch-driver?refresh=auto [Accessed 23-Aug-2024].

CENELEC (2017): Railway Applications - The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS) - Part 2: Systems Approach to Safety. EN50126-2:2017.

CENELEC (2018): Railway applications – Communication, signalling and processing systems – Safety related electronic systems for signalling. EN50129:2018.

CENELEC (2020): Railway applications – Communication, Signalling and processing systems – Software for railway control and protection systems. EN50128:2011+A1+A2:2020.

CENELEC (2023): Railway Applications - Requirements for software development. EN 50716:2023

EPRD (2014): Electronic Parts Reliability Data (EPRD-2014). Reliability Information Analysis Center (RIAC).

Filipsson, F. (2024): The Role of AI in Autonomous Public Transport. August 18, 2024. Available at: https://redresscompliance.com/ai-in-autonomous-public-transport/ [Accessed 23 Aug 2024].

Fragnelli, V. and Sanguineti, S. (2014): A game theoretic model for re-optimizing a railway timetable. European Transport Research Review 6(2):113-125. http://dx.doi.org/10.1007/s12544-013-0116-y.

Gibert, X., Patel, V.M. and Chellappa, R. (2017): Deep multitask learning for railway track inspection. IEEE Transactions on Intelligent Transportation Systems 18(1):153-164. http://dx.doi.org/10.1109/TITS.2016.2568758.

Gibson, H. (2012): Railway Action Reliability Assessment user manual - A technique for the quantification of human error in the rail industry. T270. RSSB. London.

IEC (2010): IEC 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems. International Electrotechnical Commission.

IEC (2017): IEC 61709, Electric components – Reliability – Reference conditions for failure rates and stress models for conversion. International Electrotechnical Commission (IEC).

ITP Staff (2021): Dubai RTA turns to AI to improve Metro services. Edge Middle East. Available at: https://www.edgemiddleeast.com/innovation/emergent-tech/95408-dubai-rta-turns-to-ai-to-improve-metro-services [Accessed 23 August 2024].

Kaplan, A. and Haenlein, M. (2019): Siri, siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons 62(1):15-25. http://dx.doi.org/10.1016/j.bushor.2018.08.004.

Koopman, P. (2022): How Safe is safe Enough? – Measuring and prediction Autonomous Vehicle Safety. 1st Edition. Carnegie Mellon University.

MIL-STD-217F (1991): Reliability Prediction of Electronic Equipment. United States Department of Defense, Military Standard.

NPRD (2016): Nonelectronic Parts Reliability Data (NPRD-2016). Reliability Information Analysis Center (RIAC).

ONRSR (2020): ONRSR Guideline - Major Projects #A415539 Version 2.0.

Pauley, K and Gabler, A. (2023): Quantifying Human Reliability in Safety Analysis – How Useful is it?. Australian Safety Critical Conference 2023.

Ruifan Tang, De Donato, L., Bešinović, N., Flammini, F., Goverde, R.M.P., Lin, Z., Liu, R., Tang, T., Vittorini, V. and Wang, Z. (2022): A literature review of Artificial Intelligence applications in railway systems. Transportation Research Part C: Emerging Technologies 134:103679. https://doi.org/10.1016/j.trc.2022.103679.

Safe Decisions (2016): A framework for considering safety when making decisions in the Australian Rail Industry. Version 1.0, 17 March 2016.

Siemens Mobility GmbH. (2024): Driverless train operations: Siemens Mobility upgrades signaling for entire S-bane network in Copenhagen, Denmark. Retrieved from https://press.siemens.com/global/en/pressrelease/driverless-train-operations-siemens-mobility-upgrades-signaling-entire-s-bane-network (Accessed on 23-08-2024).

Telcordia (2013): Reliability Prediction Procedure for Electronic Equipment (Telcordia SR-332). Issue 4. Telcordia Technologies, Inc.

Weits, E., Munck, S. and Eigenraam, A. (2019): Deriving THR and SIL from National Safety Targets, accounting for scale and exposure. In: Proceedings of the 2nd International Railway Symposium Aachen 2019.

Whalley, A. (2024): Australian Rail Risk Model - Risk Profile Report. Document No.: ARRM-0021-REP, Revision No.: 07.00. Verified and Approved by Robinson, N. Prepared by RGB Assurance Pty Ltd, ABN: 84-145-897-418, 236 Montague Road, West End QLD 4101.

Wikipedia (2024): List of railway accidents in Australia. Available at: https://en.wikipedia.org/wiki/List_of_railway_accidents_in_Australia [Accessed 12 January 2024].

Wohlfeil, J. (2011): Vision-based rail track and switch recognition for self-localization of trains in a rail network. In 2011 IEEE Intelligent Vehicles Symposium (IV), pp. 1025–1030. http://dx.doi.org/10.1109/IVS.2011.5940466.