HSI2024
Human Systems Integration Conference
Jeju, Korea
August 27-29, 2024

Accident Case Study Analysis of Developmental Automated Driving System Collison

Sanjeev K. Appicharla
Adora Consulting Limited
Borehamwood, Herts, UK.
020 8387 4295
appicharlak@yahoo.co.uk

Abstract. The aim of the Case study is to produce a list of accident causal factors using the Cybernetic risk management model with the hybrid Swiss Cheese Model (SCM) and Management Oversight & Risk Tree (MORT) Methodology. The hybrid SCM/MORT methodology incorporates Jens Rasmussen's risk management framework (RMF) and is augmented by including the Heuristics & Biases approach to make the methodology capable of identifying latent failures conditions at all levels of the socio-technical system that control the system of interest (SOI). The desk top study included collection of information and data that is publicly available to represent all relevant viewpoints to ensure completeness. The results raise awareness of latent causal factors in the form of biases that have impact on Risk management, Decision making, Assurance and wider Human factors concerns that are relevant and applicable to Artificial Intelligence/ Machine Learning (AI/ML) domain. It is hoped the Case Study will contribute to reflection on the part of systems engineers to help them plan, design, develop and operate safer automated vehicle systems.

Summary Description of the Accident

The NTSB Report From the NTSB Report (NTSB, 2020), the following information is extracted: This information describes the event T and the accident SA1 as per the MORT Manual (Kingston, J et al, 2009a)(pp. 1).

"On March 18, 2018, at 9:58 p.m., an automated test vehicle, based on a modified 2017 Volvo XC90 sport utility vehicle (SUV), struck a female pedestrian walking across the northbound lanes of N. Mill Avenue in Tempe, Arizona. The SUV was operated by the Advanced Technologies Group of Uber Technologies, Inc., which had modified the vehicle with a proprietary developmental automated driving system (ADS). A female operator occupied the driver's seat of the SUV, which was being controlled by the ADS. The road was dry and was illuminated by street lighting (NTSB, 2020).

The SUV was completing the second loop on an established test route that included part of northbound N. Mill Avenue. The vehicle had been operating for about 19 minutes in autonomous mode—controlled by the ADS—when it approached the collision site in the right lane at a speed of 45 mph, as recorded by the ADS. About that time, the pedestrian began walking across N. Mill Avenue where there was no crosswalk, pushing a bicycle by her side. The ADS detected the pedestrian 5.6 seconds before impact. Although the ADS continued to track the pedestrian until the crash, it never accurately classified her as a pedestrian or predicted her path. By the time the ADS determined that a collision was imminent, the situation exceeded the response specifications of the ADS braking system. The system design precluded activation of emergency braking for collision mitigation, relying instead on the operator's intervention to avoid a collision or mitigate an impact. Video from the SUV's inward-facing camera shows that the operator was glancing away from the road for an extended period while the vehicle was approaching the pedestrian. Specifically, she was looking toward the bottom of the SUV's center console, where she had placed her cell phone at the start of the trip. The operator redirected her gaze to the road ahead about 1 second before impact. ADS data show that the operator began steering left 0.02 seconds before striking the pedestrian, at a speed of 39 mph. The pedestrian died in the crash. The vehicle operator was not injured.

The National Transportation Safety Board determine that the probable cause of the crash in Tempe, Arizona, was the failure of the vehicle operator to monitor the driving environment and the operation of the automated driving system because she was visually distracted throughout the trip by her personal cell phone. Contributing to the crash were the Uber Advanced Technologies Group's (1) inadequate safety risk assessment procedures, (2) ineffective oversight of vehicle operators, and (3) lack of adequate mechanisms for addressing operators' automation complacency—all a consequence of its inadequate safety culture. (NTSB, 2020).

Further factors contributing to the crash were (1) the impaired pedestrian's crossing of N. Mill Avenue outside a crosswalk, and (2) the Arizona Department of Transportation's insufficient oversight of automated vehicle testing. Further, to the above probable cause, and contributory factors, the NTSB investigation identified the following two safety issues: Uber Advanced Technologies Group's inadequate safety culture and the need for safety risk management requirements for testing automated vehicles on public roads. As a result of its investigation, the National Transportation Safety Board made recommendations to the National Highway Traffic Safety Administration, the state of Arizona, the American Association of Motor Vehicle Administrators, and the Uber Technologies, Inc., Advanced Technologies Group." In summary, the NTSB concluded 19 findings (NTSB, 2020) (clause 3.1) (pp.58). (1) driver licensing, experience, or knowledge of the ADS operation; (2) vehicle operator substance impairment or fatigue; or (3) mechanical condition of the vehicle did not contribute to the accident (NTSB, 2020) (Executive Summary). It is to be noted that (NTSB, 2020) Recommendations used the 1997 RMF implicitly (Read, G. J., et al, 2022) (see Figure 1).

Human system integration (HSI) and Accident Analysis in the AV sector (INCOSE, 2023) defines Systems Engineering (SE) as a transdisciplinary and integrative approach to enable the successful realization, use, and retirement of engineered systems, using systems principles and concepts, and scientific, technological, and management methods. Since the AI systems are defined as an engineered systems that do not understand; they need human design choices, engineering, and oversight (see section 5) IEC 22989 (2022) (ISO, 2022). (INCOSE HSI WG , 2023) defines Human Systems Integration (HSI) is a transdisciplinary sociotechnical and management approach of systems engineering (SE) used to ensure that system's technical, organizational, and human elements are appropriately addressed across the whole system lifecycle, service, or enterprise system. In the context of AI systems, it is to b noted that human-machine teaming is a term that denotes integration of human interaction with machine intelligence capabilities (ISO, 2022).

Accident analysis, to re- paraphrase, (Leveson.N.G, 2011) can yield valuable information about the circumstances leading to the accident and help understand why it happened so that future accidents can be prevented(pp.54). James Reason coined the term, Organizational Accidents, to denote accidents in which the elaborate defences were breached by unlikely combination of several distinct failures (Reason, J et al., 2006)(section 4.1). These failures are of two basic kinds: active failures and latent failures (ibid, see Figure 5). These are distinguished by whom they were committed and the time the failures they have taken to have an impact (Reason, J., 1993). These are defined in the section on the *Heuristics & Biases approach to risk management*. Lack of consideration of human, organisation and technical factors using a systems thinking approach (Rasmussen,J et al., 1994a) (Appicharla, S. K., 2006a) and lack of identification of all latent factors as per the 1997 Risk Management Framework (RMF) (Rasmussen. J, 1997), (Read, G. J., et al, 2022) (see Figure 1) in accident analysis are frequent omissions (Appicharla.S, 2023a). Accident analysis reveals several decision makers who from their local perspective strive to meet their production targets within their local constraints and made decisions that prepare the latent pathway to accident (Reason, J., 1993), (Rasmussen. J, 1997).

The following three limitations of accident studies must be recognised (Reason.J, 1990b). First, the main input to the accident study is the accident report (Read, G. J., et al, 2022). This information may be limited to what was available to the accident investigators commercially, even if the investigators are open minded. Therefore, additional information is needed to understand the socio-technical context. The second limitation is that of *bias*. The analyst may suffer from '*Hindsight bias'*. Observers of past events may exaggerate what other people or actual persons involved in the hazardous situation should have been able to anticipate in foresight. Another aspect of *Hindsight bias* is that people are unaware of the degree to which outcome knowledge influences their perceptions of past. As a result, they overestimate what they would have known had they possessed this knowledge (Reason.J, 1990b).(pp.17,215). *Hindsight bias* leads people to blame unhappy outcomes on folly rather than ignorance. *Outcome bias* leads people to confuse the quality of decision-making with the quality of outcomes. This leads to regretting sound decision with unlucky outcomes and feeling unwarranted pride in unsound decision with lucky outcomes (Fischhoff, B., & Kadvany, J., 2011)(pp. 124). However, it is seen that that *Hindsight bias* is over emphasised in (Leveson.N.G, 2011) and this emphasis may be seen prompting a viewpoint that is too narrow to effectively address *bias risks* (Schwartz, Reva, et al., 2022)(pp.11/77).(Leveson.N.G, 2011) (section 2.7) states when learning how to engineer safer systems then the emphasis in accident analysis needs to shift from *cause* (in terms of events or errors), which has a limiting, blame orientation, to understanding accidents in terms of *reasons*, that is, why the events and errors occurred. This distinction between *reasons* and *causes* in terms of *causality and unity of subject and objec*t are two ideas articulated by Schopenhauer (1813) (2006) (Appicharla S. K, 2010c). The accident analysis needs to consider the workflow from the Identification of the Socio-technical system (STS) to Work Domain Analysis to Activity Analysis, Individual actor decision analysis, and User-work coupling in a staged manner to understand what designers had in their mind (Rasmussen,J et al., 1994a).

The paper is structured into three major sections. The first section deals with the challenges to AV safety management system from systems engineering perspective. The second section discuss the safety assurance aspects. The third section presents the Accident Causation model with its application results.

Challenges to the AV Safety Management System

Following the workflow stated in the previous paragraph, we examine the sociotechnical system involved in the control of hazard source (Rasmussen. J, 1997). As per the UK Government paper (The DSIT, 2024), the release of ChatGPT is a sputnik moment for humanity – taking humanity by surprise with rapid and unexpected progress in a technology of its own creation. The paper notes with concern accelerating investment into and public adoption of advanced AI that the AI systems are becoming more powerful and consequential to human lives. The DSIT 2024 paper defines the term, "AI Safety", thus: "the understanding, prevention, and mitigation of harms from AI. These harms could be deliberate or accidental; caused to individuals, groups, organisations, nations or globally; and of many types, including but not limited to physical, psychological, social, or economic harms."

Standards at the industry body level set by standards development organisations seek to provide guidance on safety requirements. In the UK, the PAS 1881 (The BSI, 2022) defines the requirements for assuring the operational safety of automated vehicles trials The PAS 1881 is to be read in conjunction with the other PASs dealing with functional safety, safety of the intended functionality (SOTIF), cyber-security, and guidelines for developing and assessing the control system for AVs and Safety operators and the DfT's Code of practice for automated vehicle testing and trialling. These standards define the UK strategy based upon the concept that safety risks have been identified, managed, and reduced as low as reasonably practicable (ALARP) and to an acceptable level as noted in the PAS 1881: 2022(section 6.6). Both in risk assessments and accident analysis, hazard analysis is a key practice to apply PAS 1881 (The BSI, 2022)(section 5 (c)). The ISO,IEC, and the SAE also are creating standards for the automated vehicles. ISO/IEC standard TR 5469 (ISO, 2024) and

(AVSC, 2024) are examples of standards/Guidance by the standards development organisations such as IEC and SAE. This level is the level 5 of the 1997 RMF (Read, G. J., et al, 2022) (see Figure 1).

Badri et al. (2018) and Kadir et al. (2019) cited by (Grosse, E. H., et al, 2021) described omission of human factors/ergonomics aspects in the occupational health and safety, and in research relating to the fourth industrial revolution I4.0. Motivated by this information, (Grosse, E. H., et al, 2021) performed content analysis of research papers on the I4.0 revolution. And as a result, they introduced five "Key Concepts" that can provide a basis for understanding the interrelation of the fourth industrial revolution I4.0 and HF field. The concept is that Industry 4.0 systems are sociotechnical systems(STS). The Cyber Physical Systems increase the coupling between sub-systems of technical and social components leading to greater interactive complexity and thus, increasing the risk foreseen by (Perrow, C, 1984;1999). The second concept is that of consideration of HF/E aspects throughout the life cycle of the system. This is not done can be seen from the PAS standard 1881 discussed in the previous paragraph as well. The third aspect is Human-system interaction engages perceptual, cognitive, and motor systems. This concern is noted for high technology high hazard systems by (Rasmussen,J et al., 1994a) (chapter 3) in the form of provision of Ecological Information System. The fourth concept is related to psychosocial needs of human stakeholders in relation to the workplace. This theme was foreseen by the cognitive systems engineering experts on how to attain the coupling between the social and technical components in terms of social organisation by (Rasmussen,J et al., 1994a)(chapter 4). The fifth concept relates to the Rasmussen's claim of the tendency of firms to drift to unsafe states. This principle, however, has been well illustrated in a number of disaster scenarios (Rasmussen. J, 1997). The drift model shows how organisations may migrate towards unsafe states under the economic pressures, and unacceptable workloads as advanced by (Rasmussen,J et al., 1994a). Note that this drift towards unsafe boundary is different from concept & data drift in the AI context (ISO, 2022) (section 5.11). These themes are already familiar, but rarely adopted in practice according to researchers and practitioners in the SE, System Safety and the HF/E domain (Appicharla S. K, 2010c) , (Waterson, P.,et al, 2015), , (INCOSE HSI WG , 2023). In addition to these concepts, we need to explore few other concepts like Concept of Operations, Operational Concept as well. The evidence presented so far support the hypothesis of less than adequate adoption of the socio-technical systems concept. Further, as Human-AI Interactions increase *complexity* (Senge. P, 1990b), (Rasmussen. J, 1997) and scope for *biases due to human- AI systems interaction is increased* (NIST, 2023)(pp.40-41). And *biases due to uncertainty* (INCOSE, 2023) (section 1.4.2) is another factor. Therefore, the Safety Management System needs to consider inclusion of human organisational and technical (HOT) factors in hazards analysis to consider *bias risks* in its safety data in addition to the requirements noted in the PAS 1881 (The BSI, 2022). The question of *systemic and cognitive biases* is omitted in the definitions used for functional safety, SOTIF/operational safety by the PAS 1881. These discussions are relevant to levels 1 to 4 of the 1997 RMF (Read, G. J., et al, 2022) (see Figure 1).

In the context of risks to the automated vehicle trialling, lessons learnt in the aviation sector are invaluable.(Reason J, 2001) observed that for at least two decades Human and organisational factors dominated the risks to aviation sector So, an effective safety management system (SMS) must be capable of both identifying and controlling these 'softer' and subtler issues. However, regulators and aviation managers are competent in technical and operational backgrounds but unaware of organisational aspects . So, the problem he framed was how to recognise and evaluate, build, and operate effective SMSs?." To solve this problem (Reason J, 2001) promoted the idea of three components of culture, namely, Cognition, Commitment and Competence and integrate them together in a matrix form with Principles, Policies , Procedures, and Practices to help improve the safety management systems. These are discussed later in the Accident Causation Model section. However, at this juncture it is to be noted that balancing of organisational & management failures and front line errors is recommended by some researchers, who object to focus of attention on higher level factors (Accou, B., & Carpinelli, F, 2022), (Reason,.J et al., 2006).This is one of the latent factor(s) that prepares the latent pathway to the safety incidents is to be noted(see figure 5) (Reason,.J et al., 2006), (Appicharla.S, 2023a). Thus, we need to be careful in adopting the FAA model of safety model in view of the contributory factors observed in Boeing 737 Max 8 crashes (Appicharla.S, 2023a).

In Reason's classification of approaches to safety management systems there are three models to consider in an integrated manner : engineering, person and system approaches to human error (Reason J, 2001). And weakness of the engineering and person approaches need to be considered as well. Thus, it is safe to conclude that the functional safety PAS 1881: 2022 and related standards cater to engineering and person approach only. Therefore, a Systems Approach to Safety is needed (Appicharla.S, 2023a).

Inadequate Socio-technical Risk Perspective or Techno-solutionism bias

Techno-solutionism is a *bias* that arises due to exclusion of socio-technical context in the problem situation (Schwartz, Reva, et al., 2022). The 1997 RMF can be used to advance the concepts of HF/HSI into the research on AI safety as well (Rasmussen. J, 1997). Drawing upon the concepts advanced in the the 1997 RMF, a new analytical tool, the Sociotechnical Influences Space (SIS) is proposed to support organisations in taking a holistic approach to the incorporation of advanced technologies into workplaces and assist function allocation in mixed human-artificial agent teams. An application in the Australian defence context is presented (Brady, A., & Naikar, N, 2022)(see Figure 4). (Read, G. J., et al, 2022) used the 1997 RMF and related Actor Map and accident model called AcciMAP to study automated vehicle accidents. (Read, G. J., et al, 2022) cite the system thinking approaches like the Systems Theoretic Process Analysis method (STPA), Functional Analysis Resonance Method (FRAM), NETworked Hazard Analysis and Risk Management System (NETHARMS), and the Event Analysis of Systemic Teamwork Broken Links approach (EAST-BL) that can be applied in a pro-active manner. However, they note prospective analyses have not yet considered risks across the entire sociotechnical system, and this may provide a fruitful area for future research. Using a socio-technical approach to identify *AI biases* makes it possible to evaluate dynamic system and understand how *biases* impact each other and under what conditions the *biases* are attenuated or amplified. Adopting a socio- technical perspective can enable a broader understanding of AI impacts and the key decisions that happen throughout the AI lifecycle–such as whether technology is even a solution to a given task or problem (Schwartz, Reva, et al., 2022)(pp. 10/77). The Concept of Operations, an engineering concept, defined in the IEC 29148 standard (ISO, 2018) that can help envisage risk management in a STS context is another concept rarely discussed in the safety documentation.

Inadequate Systems Engineering Concepts : Concept of Operations (ConOps) all & the Operational Concept(OpsCon)

The trialling organisation must pay attention to the Concept of Operations (ConOps) and the Operational Concept(OpsCON) in addition to Operation design domain((ODD) to uncover hazards generated by the AV system. The details of ODD are discussed by (NTSB, 2020)(section 1.5.3). The ConOps describes the way the organization will operate to achieve its missions, goals, and objectives (INCOSE, 2023). The ConOps document is developed to illustrate by means of verbal and graphic statement(s), in broad outline, of an organization's assumptions or intent with regard to an operation or series of operations. The ConOps includes how an organization intends to employ available human and technological resources to achieve one or more outcomes (ISO, 2018). The ConOps document can help develop a shared understanding of the interactions between the front-line/remote operations staff, various technological systems, emergent properties involved at these interfaces between organisations, persons, and systems on their boundaries under various states of operations. Further, it enables a pro-active approach to risk management. The ConOps can give information in a graphical manner as well. Operational concept (OpsCon)—Describes the way the system will be used during operations, for what purpose, in its operational environment by its intended users and does not enable unintended users to negatively impact the intended use of the system nor allow unintended users from using the system in unintended ways (INCOSE, 2023)(pp.105).The ML Framework and related lifecycle process standards can provide inputs to the ConOps[1]as well. The workflow from the CoNOps to the ODD through the OpsCon and deployment in the form of simulation, test track running, and trials afford a visibility of the latent pathway to accident risk. ConOps etc when elaborated reveal insights into emergent properties of AV system and alignment of lack of there with societal goals as noted in the Sociotechnical Influences Space (SIS) tool (Brady, A., & Naikar, N, 2022).

Safety Assurance of Machine Learning Driven Automated Vehicles

Brief introduction to the developments in the AI/ML domain.

In the late1940s developments in control systems and communication engineering by led (Wiener, N, 1948), and (John McCarthy et al,1955)'s proposal for the Artificial Intelligence computers (Editorial, 2019) coupled with inventions of transistors and integrated circuits by Bell Lab scientist and engineers in the late 1950s led the third industrial revolution. Further developments in the 1980s in the fields of computer science and cognitive science led to the information technology and communication revolution added foundation to the fourth industrial revolution. Report by (Kagermann, H. et al, 2013) cited in (Grosse, E. H., et al, 2021) kicked started the fourth Industrial Revolution I4.0. Progression is enabled by the availability of large amounts of data and computation resources. ML methods include neural networks and deep learning(foot-note2). Further, the works by Bengio, Y., Hinton, G, and LeCun. Y.A, the 2018 ACM Turing Award winners are known as "Godfathers of AI" and "Godfathers of Deep Learning". Bengio, Y., Hinton, G, and LeCun. Y.A, the 2018 ACM Turing Award winners along with the works by (Vaswani, A., et al, 2017,version 1) fuelled the revolution further. The concepts of Deep Learning, Recurrent Neural Networks, learning algorithms that turned out to be very good at discovering intricate structures in high-dimensional data and applicable to many domains of science, business, and government etc. are described by (LeCun, Y., Bengio, Y., & Hinton, G., 2015). The textbook by Goodfellow et.al (Goodfellow I, et al , 2016), articles by Bengio (Bengio, Y., 2009), Hinton, G (Hinton, G. E.,et al, 2006), LeCun. Y.A (LeCun, Y., Bengio, Y., & Hinton, G., 2015), Vashwani,A. et al provide a bird's eye view of the subject matter under discussion. ISO/IEC 22989:2022 may be consulted for AI concepts and terminology (ISO, 2022). (Sir Roger Penrose, 2004) is one of the many books that can be suggested for further reading on the Quantum Mechanics and Relativity revolutions in physics or the AI revolution and by no means this paragraph can be construed to describe the entire history of revolutions in physics during the 20th century and first two decades of 21st century.

Inadequate System Safety and Operational Safety cases

Three approaches are recommended by the PAS 1881 (2022) (The BSI, 2022).First, Independent safety case review to determine if specific trial is safe to proceed. Second, Process review for a trialling organization to show that it has the appropriate safety management system and processes in place to produce a safety case. Third, Self-certification for individual trials if the trialling organization has achieved safety assurance through a process review. These can be used individually or combined to assure that the consistent and robust approach to managing the risks associated with automated vehicles, and to facilitate trials and services across testing locations and environments through the operational safety case (The BSI, 2022). In this context, terms of verificaion and validation used in the context of AI Trustworthiness (section 3.5) are to be distinguished from their usage in the AI system life cycle stages and processes (section 6.2.4) (ISO, 2022). A survey on AI Assurance Literature from 1985 till 2021 was published by (Batarseh, F. A.,et.al, 2021). They defined the AI Assurance Process thus: "A process that is applied at all stages of the AI engineering lifecycle ensuring

[1] The foundational standards for AI ISO/IEC 22989 and ISO/IEC 23053. https://jtc1info.org/wp-content/uploads/2022/06/03_08_Paul_Milan_Wei_The-foundational-standards-for-AI-20220525-ww-mp.pdf Accessed on 7th June 2024.

that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy, and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users". Traditional safety cases processes need to be augmented to help identify *bias risks* as *biases* remain endemic across technology processes and can lead to harmful impacts regardless of intent if the AI Assurance goals (Batarseh, F. A.,et.al, 2021). are to be met (Schwartz, Reva, et al., 2022) (pp.i/77). As noted earlier, the ConOps and other concepts should be added to the safety case documentation with particular emphaisis on identifying *biases* and their contribution to safety risks and the safety mangement system is to be enhanced (The BSI, 2022)(section 5).

Systemic and cognitive *biases* that form the iceberg hidden beneath the surface are to be recognised and addressed in the assurance activity (Schwartz, Reva, et al., 2022), (Starbuck W.H; Farjoun M, 2005)(pp.246-266). For example, an instance of *historical bias* amongst natural scientists can be recounted: In 1928 Nobel laureate Max Born expressed confidence that physics as it was known will be over in six months (Hawking, S, 2005) (pp.118). Further prevalence of *overconfidence and loss aversion biases* amongst corporate executives is noted by (Kahneman.D, 2012). These facts raise doubts on capability to achieving a safety case in accordance with the PAS 1881(The BSI, 2022). As decisions involving AI life cycle may introduce *systemic and cognitive biases* as the human roles interacting with the AI system and developing it are part of the system , and as each AI actor may process information according to different information processing strategies, therefore, *biases* must be identified and addressed in the safety management system (Rasmussen,J et al., 1994a) (pp.78), (NIST, 2023) (Appendix C), (Schwartz, Reva, et al., 2022). The ISO/IEC standard TR 5469 uses the definition of system without the use of note 3 used in the IEC 15228 standard and thus excludes human and organisational factors from inside the loop. Moreover, traditional safety strategies of fail safe etc. may not suffice (ISO, 2024). Further, there is an early warning regarding the lay public view point that the danger of machines posing threat to, or mastering humanity are no longer undeniable but must be recognised during the design and development stage (Wiener,N, 1960), (NIST, 2023).

(Bloomfield, R., & Rushby, J., 2021)argue that some of these assurance challenges are new, for example, autonomous systems with major functions driven by machine learning and AI, and ultra-rapid system development, while others are the familiar, persistent issues of the need for efficient, effective, and timely assurance. (Deng, Y et.al, 2023) argue that the evaluation of assurance cases is carried out with the help of human insight and experience, therefore, it is prone to errors in human error judgement. Moreover, these are based upon text-based documentation of 500 pages that is huge and there is a potential need for reasoning about these cases using automation and AI systems. To satisfy this demand, they presented Trustworthiness Derivation Trees to enhance assurance cases. (Hinton, G. E., 2010) discussed progress in machine learning shows that it is possible to learn in deep hierarchies without requiring any labelled data. Further, Hinton and Nobel laureate Kahneman (Bengio, Y., Hinton, G., et al, 2023) raised concerns over the upcoming advanced AI models. The authors note that in 2019, GPT-2 could not reliably count to ten. Only four years later, deep learning systems can write software, generate photo realistic scenes on demand, advise on intellectual topics, and combine language and image processing to steer robots. As AI developers scale these systems, unforeseen abilities sand behaviors emerge spontaneously, without explicit programming (Bengio, Y., Hinton, G., et al, 2023). *We have reached a stage where programmers are baffled due to the emergent behaviour that is harmful in nature and as it is the role of systems engineers to manage the unwanted emergent properties, and the engineering practices need to improve* (Wiener,N, 1960), (Schwartz, Reva, et al., 2022), (INCOSE, 2023). *The improved understanding by systems engineers so obtained through reflection when applied appropriately contributes to reducing technical debt* (INCOSE, 2023)(pp.7).

The Accident Causation Model

(Rasmussen. J, 1997) stated in relation to accident analysis from models of deviation from the rational norms perspective: “The combination of the two basic views that (1) accidents should be understood in terms of an energy related process and (2) hazard management therefore should be directed towards planning of the release route led Johnson (1973) to focus on the management as being responsible for the planning of the context within which accidents unfold, that is, he stressed the role of ‘less than adequate’ management decisions and developed MORT - the ‘Management Oversight and Risk Tree’ tool for accident analysis. Later, Reason (1990) has focused analysis on management errors and organisational factors, such as ‘resident pathogens’ making organisations vulnerable to accidents”. Synthesizing the work on Swiss Cheese Model (SCM) of accident causation and the Management Oversight and Risk Tree (MORT), the author developed a Cybernetic risk model and uses it in conjunction with the Systems engineering process (INCOSE, 2023). The model assumes the knowledge base of controls system theory and the “*Heuristics and Biases*” approach underlying the SCM and the MORT Audit Model. This is described in (Appicharla.S, 2023a), (Kingston, J et al, 2009a). The author has obtained permission vide from the Intellectual Property holder of the MORT documentation vide private correspondence to include human factors into the MORT Tree. The additional set of questions that arise from this inclusion are drawn from the 3Cs and 4Ps matrix advanced by (Reason J, 2001) and themes discussed by (Read, G. J., et al, 2022), Table 2, (Schwartz, Reva, et al., 2022)

The SIRI Cybernetic Accident Risk Management Model

The risk model (see Figure 1) uses the hybrid Swiss Cheese Model (SCM) and (MORT) terminology under the lens of 1997 RMF, control system theory and *Heuristics & biases* approach to explain accident causation. The control systems theoretic model explains how the unsafe outcomes are result of inadequate interaction of five categorical factors relating to compliance with system engineering and related safety/human factors integration standards, business policy and integration of risk related policies, management policy and its implementation, risk management (policy and its implementation) and *biases* in thinking at various levels of STS acting as disturbances.

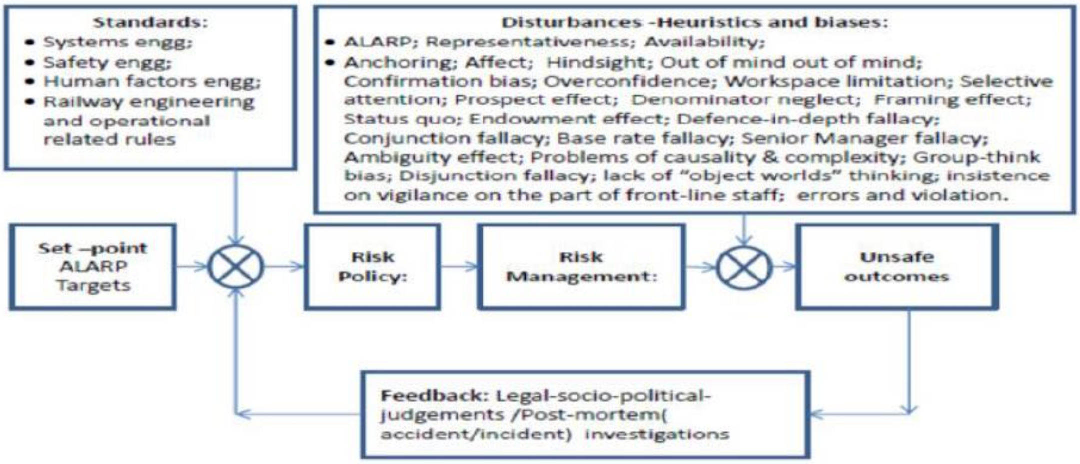

Standards:
• Systems engg;
• Safety engg;
• Human factors engg;
• Railway engineering and operational related rules
Disturbances -Heuristics and biases:
• ALARP; Representativeness; Availability;
• Anchoring; Affect; Hindsight; Out of mind out of mind; Confirmation bias; Overconfidence; Workspace limitation; Selective attention; Prospect effect; Denominator neglect; Framing effect; Status quo; Endowment effect; Defence-in-depth fallacy; Conjunction fallacy; Base rate fallacy; Senior Manager fallacy; Ambiguity effect; Problems of causality & complexity; Group-think bias; Disjunction fallacy; lack of “object worlds” thinking; insistence on vigilance on the part of front-line staff; errors and violation.
Set –point ALARP Targets
Risk Policy:
Risk Management:
Unsafe outcomes
Feedback: Legal-socio-political-judgements /Post-mortem(accident/incident) investigations

Figure 1. The SIRI Cybernetic Accident Risk Management Model (Appicharla.S, 2023a)

Control systems engineers may recognise Figure 1 from its similarity to data fusion filter algorithm in the form of Kalman filter (Lidl ,R; Pilz,G, 2004), (Rasmussen,J et al., 1994a). With reference to the above SIRI Cybernetic Risk Management Model (Appicharla.S, 2023a), the deviations from the best practices are high level latent failure contributors to the unsafe outcomes.

The Model describes how the unsafe outcomes are a result of:

- Less than adequate (LTA) development/application of standards related to systems engineering, safety engineering, human factors, and domain related standards (Appicharla S. K, 2010c), (Fang, X., & Johnson, N, 2019), (Redmill, F, 2000). Underdeveloped software testing standards is one of them (NIST, 2023)(pp.38). Compliance with necessary standards is a part of assurance criteria but may not cover all hazards (The BSI, 2022) (Starbuck W.H; Farjoun M, 2005).
- LTA responses to the disturbances in the form of *Heuristics and Biases* (Reason.J, 1990b), (Appicharla.S, 2023a). *Human-AI Interactions increase the scope for biases and biases risks are an important category* (Schwartz, Reva, et al., 2022). The *Heuristics and Biases* (H &B) approach to Risk Management sub-section presents these in more detail.
- LTA Business policy and its implementation and its integration with risk related policies (Kingston, J et al, 2009a), (Rasmussen,J et al., 1994a). Use of pre-trained models can also increase levels of statistical uncertainty and cause issues with *bias* management, scientific validity, and reproducibility (NIST, 2023) (pp.38). Business policy needs to pay attention to concerns such as audit, ethical, duty of care, safety risk legal and community responsibilities as demonstrated in the Boeing 737 Max -8 accident analysis (Appicharla.S, 2023a).These are recommended in the AV sector as well (Koopman, P., & Widen, W. H, 2023).
- LTA Risk management practices of System Definition, Hazard Identification, Risk Analysis and Implementing Risk Controls Options, and Assurance Management (Reason.J, 1990b), (Rasmussen. J, 1997). There is a higher degree of difficulty in predicting failure modes for emergent properties of large-scale pre-trained models (NIST, 2023)(pp.38).
- Finally, LTA philosophy of ALARP decision making, oversight process/ LTA business/mission analysis. Affect and other heuristics may have an adverse impact on ALARP decision making by injecting *biases* (INCOSE, 2023)(pp. 185), (Langdalen, H et al, 2020), (Ale, B. J. M.; D. N. D. Hartford, D. Slater, 2015).

Heuristics and Biases (H &B) approach to Risk Management

Within the Cognitive System Engineering (CSE) discipline (Rasmussen,J et al., 1994a), the Systems and human reliability discipline (Reason.J, 1990b) and Organisational research discipline (Starbuck W.H; Farjoun M, 2005), (Kahneman.D, 2012), and (Simon H.A , 1997), it is accepted that decision makers are prone to use simple procedures that helps to find adequate, though often imperfect, answers to difficult questions. Technically speaking, these mental short-cuts or simple procedures are known as Heuristics (Kahneman.D, 2012)(pp.98), (Schwartz, Reva, et al., 2022). H & B approach is described in 146 pages (A5 size) in “Thinking Fast & Slow” (Kahneman.D, 2012). In the woefully inadequate short space, the author attempts to describe the three main heuristics that are employed in making judgments under uncertainty. First, *‘Representativeness heuristics’ (RH),* which is usually employed when people are asked to judge the probability that an object or event A belongs to class or process B. *Insensitivity to prior probability of outcomes (base rate neglect), insensitivity to sample size, misconceptions of chance, insensitivity to predictability, illusion of validity, misconceptions of regression* are some of the *biases* due to the representativeness heuristic. (Tuccio W.A PhD, 2011). showed that at least 19 aviation accidents over a ten-year period can be attributed to the “H& B” school of thought in decision making field and recommended that pilot training should adopt these concepts. This open access article is a must read to understand heuristics from accident analysis perspective. *Over-simplification of causality* is attributed to the representativeness and availability heuristics (Reason.J, 1990b)(pp.91).

The second one, ‘*Availability heuristic’* (AH), related to memory, is the ease with which outcomes can be brought to mind (recalled and visualised) increases their subjective salience and perceived likelihood (probability) of occurrence. Infrequent and high-impact events can often be easily brought to mind, leading people to overestimate the likelihood of such an event. *Biases due to retrievability of instances, biases due to effectiveness of search set, biases of imaginability, out of sight and out of mind bias and illusory correlation are* due to the availability heuristic (Kahneman.D, 2012). Another example of the *"Out of sight out of mind bias"* that arises from Availability heuristic is the omission of commonly known factors (see Kahneman et.al, 1982) from the risk analysis (fault tree analysis of car starting scenario) (See Fischoff and co-authors,1978) cited by (Reason.J, 1990b) (pp. 89). Omission of contributions from latent failure conditions in risk assessment(s) is an example of *"Out of sight out of mind bias"* in the case of RSSB Safety Risk Model (RSSB, 2020).

The third heuristic is the '*Adjustment from an anchor* '(AAH): subjects and people in real world situations are often unduly influenced by outside suggestions. People can be influenced even against their intentions when they know that the suggestion is made by someone who is not an expert in experimental situations, and real-world situation, subjects and decision takers make estimates by starting from an initial value that is adjusted to yield the final answer; *underestimate the probability of disjunctive event ('Either Or' gate event in fault tree analysis of a complex system; overestimate the probability of conjunctive (AND gate) events are some biases* from this heuristic (Kahneman.D, 2012). *Under-estimation of safety risk posed by MCAS system is one bias* noticed in the case of Boeing 737 Max -8 crashes and several others are identified (Appicharla.S, 2023a). Over-estimation of risk by RSSB Safety Risk Model is another example. (Evans A.E, 2020) provides the order of magnitude by which the over estimation of risk is done. (Tuccio W.A PhD, 2011) discuss the role of this heuristics in the case of Southwest 1455 aircraft suffered. Lawton and Ward (2005) cited by (Appicharla.S, 2023a) argued that the *net result of a systems-based analysis is a more comprehensive understanding of the crash, in order to provide a more effective strategy for preventing future crashes by addressing all levels of factors and the critical interactions among them.* In the introduction of their work (Rasmussen,J et al., 1994a) (pp. 14) refer to management theories by Stafford Beer (1966) and (Senge. P, 1990b) where the concept of control function is used. *The concept of learning organization is still valid to learn all causal factors in accident analysis.*

Definition: (Reason, J, 1993) defined **Active failures** as unsafe acts committed by those at the "sharp end" of the system ((pilots, train drivers, control room operators, maintenance crews, and the like They are the people who are at the human -system interface whose actions can do, and sometimes have immediate consequences. These may be called acts of omissions or commissions on the part of front-line operatives. **Definition**: (Reason, J, 1993) defined **Latent failures** that are usually fallible decisions taken at the higher-level echelons of the organisation whose damaging or adverse consequences may lie dormant within the system for a long time, only becoming evident when they combine with local triggering factors (i.e., active failures, technical failures, atypical system states etc.,) to breach system's defenses. (Reason, J, 1993) noted that apart from the life-cycle errors in the system development and design processes that may occur, there would be cultural factors of competence, commitment, and cognizance that are impacted by the quality of decision making. *In the context of AI system these may be called bias risks* (Schwartz, Reva, et al., 2022).

- *Competence factor* deals with organisational capability to meet the safety goals. Elements of such competence are related to the organisation processes and standards for systems engineering process and their application (INCOSE, 2023), (Reason J, 2001).
- *Commitment* to *safety* relates to the motivation and resources for the pursuit of the safety goals in terms of either meeting regulatory targets or pursue leadership status in overcoming the hazards inherent in design and operations. Safety Management Policy together with the ways and means to pursue the safety objectives define the motives. Most importantly, capability and commitment must be tailored to cognizance of hazards (Appicharla.S, 2023a), (Starbuck W.H; Farjoun M, 2005)(pp.63). (Reason J, 2001).
- *Cognizance of hazards* must include managerial attention to latent failure conditions contribution by means of human and organisational factors to accidents. Senior managers must look beyond the active failures to understand the resident pathogens in organisation and management practices (Appicharla.S, 2023a), (Reason J, 2001).

Accident Analysis Results

From the scrutiny of (NTSB, 2020) the 1997 RMF (Rasmussen. J, 1997) (Appicharla S. K, 2010c)is constructed in Table 1. Further, the following equations are generated as well (Appicharla, S. K., 2006a). A system hazard either can arise from dysfunctional interactions between the system components or less than adequate lifecycle factors or other less than adequate control actions or management actions from the various stakeholder(s) involved. These are listed in the below. Those who are not familiar with HAZOP style procedure of using Gude words and parameters (Appicharla, S. K., 2006a) for hazard identification may skip these equations. However, these will be of use to AI stakeholders who perform design work and carry out safety assessments/audit or review them (ISO, 2022), (Schwartz, Reva, et al., 2022)(pp.34/77). (INCOSE, 2023) discusses the relation between emergence, accidents, and hazards (pp. 185). In the case of AV systems, the concept of Trustworthiness of AI systems becomes the emergent property that needs to be managed of the AV system that uses data-driven training AI model and optimization methods. The unwanted emergent property that needs to be managed is the accident or unreasonable risk. A typical example of such accident risk during the dynamic driving task (DDT) is presented in the form of following equations (The BSI, 2022)(clause 3.5 & 5).

$\text{No \_ ADS Object and Event Detection \& Collision event} = \text{Hazardous Event(HE)}$ eqn (1).
$HE + Adequate\ \_Braking\_Action = Protection$ eqn (2).
$HE + Less\ than\ adequate\ (LTA) Braking\_Action = Accident$ eqn (3).

The above equations can be used to relate safety property to Trustworthiness. The above equations (1) to (3) can help formulate a Bayesian risk assessment with appropriate data input is to be noted (Kahneman.D, 2012) (pp. 166). The MORT tree can also be converted to Bayesian Belief Network, if needed. From the inspection of the above Table 1, and the MORT Fault Tree, we can construct our fault tree (The NRI Foundation, 2009). The nodes of the MORT Fault tree are described. The MORT Fault Tree and User Manual are accessible freely. From the inspection of the above Table 1, and the standard MORT Fault Tree, we can construct our AV fault tree as under as per the three-step method (The NRI Foundation, 2009).

The following results enable a comparison to be drawn with the application of AcciMAP by (Read, G. J., et al, 2022), the INCOSE Case study (INCOSE, 2023) (section 6.5) . Evidence of automation complacency found in various settings by (NTSB, 2020) is simply an output of fallible human cognition activated by *similarity heuristic* (Reason, J et al., 2006) The differences between the analytical RMF and the empirical SCM approach are reconciled with the two system view of human perception and cogntion called Thinking Fast and Slow (Kahneman.D, 2012). The two methods are harmonised through the concept of systems thinking and its practices advanced by Stafford Beer and (Senge. P, 1990b) who were cited by (Rasmussen,J et al., 1994a). Decision making strategy from the risk management perspective of eliminating affordance for harm posed to vulnerable persons need to be considered as well (Reason, J et al., 2006).

SB1. Potentially Harmful Energy Flow or Environmental Condition a2. Functional Energy /b3. Control of Use LTA (Kingston, J et al, 2009a)(pp.1-3). The vigilance failure of the vehicle operator to monitor the driving environment and the operation of the automated driving system because she was visually distracted throughout the trip by her personal cell phone (NTSB, 2020) (pp.V). As noted in the equation (3), the protection system was operator's vigilance because the Uber ATG Concept of Operations did not envisage the ADAS to brake (NTSB, 2020) (footnote 11) and the ADS braking system was found to be less than adequate(ibid) (section 1.5.5.3 Hazard Avoidance and Emergency Braking). (NTSB, 2020)2.2.2.1 Operator's Actions discusses the concern of Automation Complacency. However, hypothetically, it may be argued that failure of ML programs may be masked by the Automatic Emergency Braking if the NTSB recommendations for the same are accepted (Koopman, P., & Widen, W. H, 2023). *Observation 1:* Reliance upon perception that a human (expert or otherwise) can effectively and objectively oversee the use of algorithmic decision systems(ADS) is a problematic assumption (Schwartz, Reva, et al., 2022)(pp.34/77). By removing the second operator, ATG also removed a layer of safety redundancy (NTSB, 2020) (2.2.2.2 Uber ATG Oversight of Vehicle Operators). The ConOps and the OpsCon documentation were not prepared (ISO, 2018).

SB2. Vulnerable People or Objects & SA.2 Stabilization and Restoration. a1. Non-functional Energy' b3. Control of exposure LTA (Kingston, J et al, 2009a)(pp.4). The NTSB concludes that the pedestrian's unsafe behaviour in crossing the street in front of the approaching vehicle at night and at a location without a crosswalk violated Arizona statutes and was possibly due to diminished perception and judgment resulting from drug use (NTSB, 2020)(section 2.1.2). *Observation 2*: The pedestrian did not have the right of way as per the state regulations (NTSB, 2020) (Clause 3.1.3). The pedestrian's violation in an unsafe act in the presence of potential hazard of traffic collision and is an active failure (Reason, J et al., 2006). SA2. Stabilisation and Restoration adequate: The emergency response to the crash was timely and adequate (NTSB, 2020)(pp.2).

Table 1: Swiss Cheese Model (Reason, J et al., 2006) , EBTA/MORT (The NRI Foundation, 2009)

B1- System Hazard.	SB2-Vulnerable targets of victims	SB3-Less Than Adequate (LTA) Barriers or risk controls
Vehicle collision with another vehicle or pedestrian (safety critical deviation).	Safety operator and pedestrian	The US Department of Transportation Federal Automated Vehicles Policy LTA
		The NHTSA Federal Motor Vehicle Safety Standards LTA
		The Arizona Department of Transportation's oversight of Automated Vehicle testing LTA
		The American association of Motor Vehicles Regulations
		Uber ATG Design Regulations
		Uber ATG System Engineering and Software Safety Team Processes
		Operator Training & Behaviour Monitoring
		Road users Training & Behaviour Monitoring

SB3 -Less than adequate barriers and controls SC1-5: Control of work and process LTA; SD1 Technical Information Systems LTA(pp.5): a1. Technical Information LTA: b1. Knowledge LTA: (pp.5).c1. *Observation 3*: (Fenn, J.,et al, 2023) state that Deep neural networks (DNNs) are used for object detection as part of a vision-based perception system that typically process sequences of input images and produce *bounding boxes* that spatially localise and highlight the detected objects of interest on each image. In use cases such as collision avoidance, the safety contribution of object detection to system hazards can be characterised by false negative detections (i.e., an object posing a collision hazard exists in the image, but the DNN does not recognise it) and inaccurate localizations (i.e., an object posing a collision hazard that exists in the image is correctly recognised, but the bounding box produced either partially covers it, or does not cover it). The equation (1) in the causation model describes this hazard. *Using Hindsight bias*, it is argued the NTSB failed to establish the failure of DNN as a contributory cause. It may be argued that systems engineers with specialisation in safety may downplay the potential benefits of automated driving while 1.2 million people die each year in traffic due to human error(94% ±2.2% to 95% confidence limits as per the NHTSA[2] by presenting a narrative which suggests that automated driving systems are dangerous and undesirable (de Winter, J. C., 2019). *Over-simplification of causality bias* (Reason.J, 1990b) & *techno-solutionism* bias (Schwartz, Reva, et al., 2022) are counter arguments to be noted. Net reduction in road deaths is welcome but other social gaols of trustworthiness etc are applicable as well (Koopman, P., & Widen, W. H, 2023).From an organisational expectations viewpoint (Starbuck W.H; Farjoun M, 2005) (pp.275) safety may be viewed as a part of quality assurance and after-the- fact auditing process.

Table 2: MORT (The NRI Foundation, 2009) SB3 -Less than adequate (LTA) barriers and controls

MORT Code /HYPOTHESIS	NTSB Evidence (NTSB, 2020)	Argument
SC1-5: Control of work and process LTA. Note: Levels 1 to 4 of the 1997 RMF are covered in this table (Read, G. J., et al, 2022).	Section 2.2.1: Uber ATG Safety Risk Management LTA Section 2.2.3.1: Precrash Safety Plan and Safety Culture Framework LTA (see foot note 4 for NHTSA's "Automated Driving Systems 2.0 Voluntary Guidance,"(pp.7)) (OEDR) requirement.	*Observation 4:* (NTSB, 2020) did not investigate any ML algorithms for wrong classification of objects detected and did not state if failure to detect an obstacle was a result of omissions in the training data. Nor did the NTSB request to the ATG to provide such information. In control theoretic terms, whether the error in the classification output due to the feedback or feed-forward operation in relation to deep learning or the data fusion filter (such as Kalman filter) or why classification flickered is not stated (Lidl ,R; Pilz,G, 2004)(section 39). (NTSB, 2020) Object and Event Detection and Response (OEDR) and related documentation LTA , and this factor contributed to the risk. In the pre-crash scenario, there was no safety management system considered by the ATG (NTSB, 2020)(2.2.3 Uber ATG Safety Policies). *Observation 5*: Configurations discussed by (Fenn, J.,et al, 2023) are of no help if the basic use cases of OEDR fail to provide data to the ADS or its monitor. The basic source of true information should be available. *Oversimplification of causality bias* due to representativeness and availability heuristics is concluded (Schwartz, Reva, et al., 2022).*Under- estimation of risk* due to Anchoring heuristic is concluded (Schwartz, Reva, et al., 2022). See observation #7 as well. *Opaque nature of safety risk* is to be considered (Reason.J, 1990b)(pp.179).
SD1 Technical Information Systems LTA(PP.5): a1. Technical Information LTA: b1. Knowledge LTA: (pp.5).	(section 2.2.1.1 Precrash ADS Functionality)LTA (section 1.5.5.3 Hazard Avoidance and Emergency Braking)LTA	*Observation 6:* The ConOps & OpsCon implicit in the Uber ATG's Planning for the Testing trials did not include redundant safety systems such as FCW and AEB (NTSB, 2020). ADS failure detection LTA/functional design LTA. The information about the Uber ATG and assessment of its safety culture was obtained through an independent assessment team (see footnote 58)(ibid). The independent team of safety assessors suffered from representativeness heuristic and suffered from *omission bias* due to neglect of ML algorithms (Schwartz, Reva, et al., 2022). *Observation 7:*

[2] https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115

c1. Based upon existing knowledge. d1. Application of Codes and Manuals, LTA? d2. List of Experts LTA. d3.ML algorithms/ Local Knowledge LTA.	(section 1.6.3 1.6.3 Interaction with Uber ATG Automated Driving System) LTA (see foot note 4 for NHTSA's Voluntary Guidance for ADS fallback requirements"(pp.8))	*Confirmation bias* in the NHTSA safety measure of ODD testing plan but excluding ConOps & OpsCon and Systems thinking practice inclusion of HOT factors in the loop. The LTA safety culture (NTSB, 2020) and the NHTSA Guidance on system safety (pp. 5) that the Entities to follow a robust design and validation process based on a systems-engineering approach with the goal of designing ADSs free of unreasonable safety risks cannot be compiled with due to *resident pathogens in current systems engineering approach* (Appicharla.S, 2023a). PAS 1881 approach is considered LTA for the same reason (The BSI, 2022). (INCOSE, 2023) (pp. 185) AI Trustworthiness as en emergent property to be included in the ConOps. LTA systems engineering expertise due to *out of sight out of mind bias* (Reason.J, 1990b)(pp.89)

Table3: MORT (The NRI Foundation, 2009), M. Management System Factors LTA

MORT Code /Claim	NTSB Evidence (NTSB, 2020)	Argument
MA3. Risk Management System LTA.(see the NHTSA Guidance footnote 4) .Note: Levels 5, 6 of the 1997 RMF are covered in this table (Read, G. J., et al, 2022).	2.3.2.1 Safety Standards and Automated Vehicle Guidance	*Observation 8:* The NHTSA Risk policy(pp.16) is like the FAA Boeing Organization Design Authority in terms of self-certification (Appicharla.S, 2023a). Operations beyond ODD and less than adequate risk mitigation pertaining to beyond ODD operations are concerns that are observed by NTSB [3]. Voluntary safety self-assessments (VSSA) LTA; Evaluation and Approval of these Assessments by NHTSA LTA. (2.3.2.2 Recommendations) (NTSB, 2020). See observation 9 as well. Despite the availability of plethora of safety standards published later to the fatal collision Uber ATG accident
MB1. Risk Management Policy LTA	2.3.4 State Approach: Legislating Automated Vehicle Testing	*Observation 9*: The state level requirements for testing are varying and no common standards exist. The functional safety approach is less than adequate 4 (ISO, 2024). The unprecedented complexity and LTA regulatory awareness together with asking for co-operation (see pp. 15) with authorities and with resident pathogens in the current systems engineering approaches makes the state level requirements LTA. This is due to *Omission bias* due to availability heuristic on the part of policy makers(Kahneman.D, 2012).
MB3. Risk Analysis Process LTA a1. Concepts and Requirements LTA:	2.3.1 Terminology of Automation	*Observation 10*: Balancing innovation and safety as emergent properties need an inherent control mechanism embedded in the design and development system such that programmers are not baffled (Wiener,N, 1960). Cognitive systems engineering approach based

[3] https://www.ntsb.gov/Advocacy/safety-topics/Documents/2021-Comments-to-NHTSA-Framework-for-ADS-Safety-ANPRM.pdf

[4] https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf cited in (Schwartz, Reva, et al., 2022).

b5. Specification of Requirements LTA: c9. Stakeholder/customer requirement c10. Statutory codes and regulations /c11. Requirements of other National and International codes and standards/c12. Local Codes and Bylaws/c13. Internal Standards. MB3. Risk Analysis/Assessment Process LTAa2. Design and Development LTA:b8. Energy Control LTA:c20. Automatic Controls LTA:c24. Controls and Barriers TA:b9. Human Factors (Ergonomics) Review LTA:	2.3.1.1 Advanced Driver Assistance System. 2.3.1.2 Automated Driving System. 2.2.2 Operator Supervision of Vehicle Automation 2.2.2.1 Operator's Actions. 2.2.2.2 Uber ATG Oversight of Vehicle Operators. 2.3.3 Industry Efforts	on systems thinking is recommended (Rasmussen,J et al., 1994a). However, bi*ases* cannot be controlled to be recognised as well (Schwartz, Reva, et al., 2022). Current HF/E approaches like (Wilson, J.R, 2014) do not include *biases* in the risk management. Thus*, omission bias is concluded* (Reason.J, 1990b)(pp.89), (Appicharla.S, 2023a).*Observation 11:* LTA awareness of developers multiply the extant HF problems warranting the learning organisation approach (Senge. P, 1990b). Traditional moral experiment of trolley problem does not solve the moral decision-making problem either (Dubljević, V. et al, 2023). *Observation 12:* "HAVs" as currently envisioned use technology that is inherently incompatible with legacy safety standards approaches" (Koopman, P. et al., 2019). Therefore, the AI standard TR 5469 is not of much use (ISO, 2024). *Techno-solutionism bia*s is concluded (Schwartz, Reva, et al., 2022). Further, interactions between through the levels of automation adds to complexity and it needs to be addressed. *Observation 13*: Use of data recording by NHTSA is a good measure (pp. 14). *Observation 14: Cognizance of hazards must include managerial attention to latent failure conditions contribution by means of human and organisational factors to accidents (Appicharla.S, 2023a).* ALARP decision making leads to bias risks in the absence of common safety standards, mandatory safety assessments and approvals for test permits (Koopman, P., & Widen, W. H, 2023), (NTSB, 2020), (Schwartz, Reva, et al., 2022). *LTA commitment, competence to manage and cognition of AV collision hazard is LTA* (Starbuck W.H; Farjoun M, 2005)*(pp.226)*

Conclusions

The paper identified *heuristics* and *biases* at all levels of socio-technical system that contributed to the fatal AV collision. The hybrid Swiss Cheese Model (SCM) and Management Oversight & Risk Tree (MORT) Methodology to identify *bias risks* was presented and its application was demonstrated. The NTSB Recommendations do not address risks inherent in the STS context is concluded.

Acknowledgements. The author expresses gratitude to the reviewers, family and the NTSB as well.

References

Accou, B., & Carpinelli, F. (2022). Systematically investigating human and organisational factors in complex socio-technical systems by using the "SAfety FRactal ANalysis" method. Applied ergonomics, 100(103662), 9. Retrieved March 6th, 2024, from https://www.sciencedirect.com/science/article/abs/pii/S0003687021003094

Ale, B. J. M.; D. N. D. Hartford, D. Slater. (2015). ALARP and CBA all in the same game. Safety science, 76, 90-100. Retrieved April 21st, 2021, from https://www.sciencedirect.com/science/article/abs/pii/S0925753515000405

Appicharla S. K. (2010c, October 19th). 5th IET International Conference on System Safety ,"System for investigation of railway interfaces [SIRI]. Retrieved December 28th, 2023, from IEEE Explore: https://ieeexplore.ieee.org/document/571235 1

——— (2006a). 1st IET International Conference on System Safety ; System for Investigation of Railway Interfaces. Retrieved March 1st, 2024, from https://bit.ly/4bFFZrf

——— (2023a, January 30th). The Boeing 737 MAX 8 Crashes, System-based Approach to Safety —A Different Perspective. Retrieved February 6th, 2023, from https://scsc.uk: https://scsc.uk/journal/index.php/scsj/article/view/18

(AVSC, 2024) *Automated Vehicle Safety Consortium (AVSC) Best Practice For Core Automated Vehicle Safety Information AVSC-D-02-2024*. Retrieved June 10th, 2024, from / https://www.sae.org/standards/content/avsc-d-02-2024/

Batarseh, F. A.,et.al. (2021). A survey on artificial intelligence assurance. Journal of Big Data, 8(1), 30. Retrieved February 16th, 2024, from https://doi.org/10.1186/s40537-021-00445-7

Bengio, Y. (2009). Learning Deep Architectures for AI. Foundation and Trends R© in Machine Learning,, 2(1), 1–127. Retrieved February 18th, 2024, from https://www.nowpublishers.com/article/DownloadSummary/MAL-00 6

Bengio, Y., Hinton, G., et al. (2023, November 12th). Managing ai risks in an era of rapid progress. Retrieved February 18th, 2024, from arXiv: https://arxiv.org/pdf/2310.17688.pdf

Bloomfield, R., & Rushby, J. (2021, January 14th). Assurance 2.0: A manifesto. Retrieved February 17th, 2024, from https://doi.org/10.48550/arXiv.2004.10474

Brady, A., & Naikar, N. (2022). Development of Rasmussen’s risk management framework for analysing multi-level sociotechnical influences in the design of envisioned work systems. Ergonomics, 65(3), 485-518. Retrieved February 23rd, 2024, from https://bit.ly/3Q1eP2i

Deng, Y et.al. (2023, September 23rd). Trusta: Reasoning about Assurance Cases with Formal Methods and Large Language Models. Retrieved February 17th, 2024, from https://arxiv.org/abs/2309.12941

Dubljević, V. et al. (2023, October 30th). Moral judgment in realistic traffic scenarios: moving beyond the trolley paradigm for ethics of autonomous vehicles. Retrieved from AI & Society: Knowledge, Culture and Communication: https://doi.org/10.1007/s00146-023-01813-y

Editorial. (2019, Septemeber 11th). Return of cybernetics. Nature Machine Intelligence, 1(385), 1. Retrieved February 26th, 2024, from https://www.nature.com/articles/s42256-019-0100-x

Evans A.E. (2020, May). FATAL TRAIN ACCIDENTS ON BRITAIN’S MAIN LINE RAILWAYS. Retrieved December 11th, 2023, from https://bit.ly/3UWLTwM

Fang, X., & Johnson, N. (2019). Three reasons why: Framing the challenges of assuring ai. In A. T. Romanovsky, Computer Safety, Reliability, and Security. SAFECOMP 2019. Lecture Notes in Computer Science (pp. 281-287). Toulouse, France. Retrieved February 17th, 2024, from https://link.springer.com/chapter/10.1007/978-3-030-26250-1_22

Fenn, J.,et al. (2023). SCSC-179: The Future of Safe Systems-Architecting Safer Autonomous Aviation Systems. In M. Parsons (Ed.), Proceedings of the Thirty First Safety-Critical Systems Symposium (pp. 163-188). York: the Safety-Critical Systems Club 2023. Retrieved February 25th, 2024, from safety critical systems club: https://scsc.uk/scsc-179

Fischhoff, B., & Kadvany, J. (2011). Risk: A Very Short Introduction. New York: Oxford University Press. Retrieved May 24th, 2021, from https://bit.ly/3Mkj3Rt

Goodfellow I, et al . (2016). Deep Learning (Adaptive Computation and Machine Learning series). Retrieved February 18th, 2024, from MIT Press: https://www.deeplearningbook.org/

Grosse, E. H., et al. (2021, March). Industry 4.0 and the human factor–A systems framework and analysis methodology for successful development. International journal of production economics,, 233(107992), 1-16.

Hawking, S. (2005). A briefer History of Time. London: Transworld Publishers.

Hinton, G. E. (2010). Learning to represent visual input. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1537), 177-184. Retrieved February 18th, 2024, from https://bit.ly/4bT0MHo

Hinton, G. E.,et al. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554. Retrieved February 18th, 2024, from bit.ly/3KjMfZs

INCOSE HSI Working Group (WG. (2023). Human Systems Integration,V1.2. San Diego,UA: INCOSE. Retrieved December 1st, 2023, from https://portal.incose.org/documents/library

INCOSE (2023). International Council on Systems Engineering Handbook, 5th Edition(D. D. WALDEN, Ed.) Hoboken, NJ 07030, USA: John Wiley & Sons Ltd. Retrieved October 13th, 2023, from https://bit.ly/3IgJeYQ

ISO/IEC TR 5469:2024(en) Functional safety and AI systems. Retrieved February 19th, 2024, from https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:5469:ed-1:v1:en

ISO (2018, November). ISO/IEC/IEEE 29148:2018(en) Systems and software engineering — Life cycle processes — Requirements engineering(Under Review). Retrieved February 23rd, 2024, from https://www.iso.org/standard/72089.html

Kagermann, H. et al. (2013, April 8th). Recommendations for implementing the strategic initiative INDUSTRIE 4.0.Securing the Future of German Manufacturing Industry, Final Report of the Industrie 4.0 Working Group. Retrieved February 26th, 2024, from acatech - National Academy of Science and Engineering: https://bit.ly/42PI8gl

Kahneman.D. (2012). Thinking Fast and Slow. London: Penguin Group.

Kingston, J et al. (2009a, December 20th). The Management Oversight and Risk Tree User Maual and Chart. Retrieved May 7th, 2022, from https://bit.ly/3vTTWzi

Koopman, P. et al. (2019). A safety standard approach for fully autonomous vehicles. Computer Safety, Reliability, and Security: SAFECOMP 2019 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE (pp. 326-332)). Turku, Finland: Springer International Publishing. Retrieved February 28th, 2024, from https://bit.ly/3VkcVQb

Koopman, P., & Widen, W. H. (2023, November 30th). Breaking the Tyranny of Net Risk Metrics for Automated Vehicle Safety. Retrieved February 28th, 2024, from SSRN : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4634179

Langdalen, H et al. (2020). On the importance of systems thinking when using the ALARP principle for risk management.". Reliability Engineering & System Safety, 204, 8. Retrieved April 21st, 2021, from https://bit.ly/3Lvr4oR

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May 27th). Deep learning. Nature, 521, pages436–444. Retrieved February 18tg, 2024, from https://doi.org/10.1038/nature14539

Leveson.N.G. (2011). Engineering a safer world: Systems thinking applied to safety. Retrieved February 23rd, 2024, from https://library.oapen.org/handle/20.500.12657/26043

Lidl ,R; Pilz,G. (2004). Applied Abstract Algebra. New Delihi: Springer India.

NIST(2023, January). The Artificial Intelligence Risk Framework((AIRMF) V1.0 NIST AI 100-1. Retrieved February 16th, 2024, from National Institute of Standards and Technology: https://bit.ly/450fzOo

NTSB. (2020, June 26th). Reissued Report-Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian Tempe, Arizona March 18, 2018.Highway Accident Report NTSB/HAR-19/03. Retrieved February 18th, 2024, from NTSB: https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf

Perrow, C. (1984;1999). Normal Accidents (1999 ed.). New Jersey: Princeton University Press.

Rasmussen,J et al. (1994a). Cognitive Systems Engineering. (A.P.Sage, Ed.) New York: John Wiley and Sons, Inc.

——— (1997). Risk Management in a Dynamic Society. Retrieved July 4th, 2020, from http://sunnyday.mit.edu/16.863/rasmussen-safetyscience.pdf

Read, G. J., et al. (2022, September). Learning lessons for automated vehicle design: Using systems thinking to analyse and compare automation-related accidents across transport domains.

Safety science, 153(105822), 14. Retrieved February 27th, 2024, from https://www.sciencedirect.com/science/article/pii/S0925753522001618

Reason J. (2001). Human Factors Aspects of Safety Management Systems. Proceedings of the 15th Annual FAA/TC/CAA Maintenance Human Factors (p. 7). London: CAA. Retrieved May 26th, 2022, from https://bit.ly/49MOyPG

——— (1993). The Identification of Latent Organisational Failures in Complex Systems: Human Factors Issues. (J. H. Wise, Ed.) Retrieved December 11th, 2021, from https://link.springer.com/chapter/10.1007/978-3-662-02933-6_13

Reason,.J et al. (2006). Revisting the « Swiss Cheese » Model of Accidents. Retrieved September 23rd, 2011, from https://bit.ly/3FaPU7T

——— (1990b). Human Error (17th ed.). New York, USA: Cambridge University Press.

Redmill, F. (2000). Installing IEC 61508 and supporting its users–nine necessities. The Australian Workshop on Safety Critical Systems and Software (p. 13). Melbourne, Australia: The Newcastle University. Retrieved February 24th, 2024, from https://bit.ly/4bxlI74

RSSB. (2020, June). Development of a new Safety Risk Model. Retrieved January 21st, 2022, from https://bit.ly/3w3lWAK

Senge. P. (1990b). The fifth discipline: The art and practice of the learning organization. (2006 ed.). New York: Double Day. Retrieved February 15th, 2023, from https://bit.ly/3yARD8e

Shwartz, Reva, et al. (2022, March). *Towards a standard for identifying and managing bias in artificial intelligence.* Retrieved May 31, 2024, from US Department of Commerce, National Institute of Standards and Technology.: https://doi.org/10.6028/NIST.SP.1270

Simon H.A . (1997). Administrative behavior: a study of decision-making processes in administrative organizations (Fourth ed.). New York: Free Press.

Sir Roger Penrose. (2004). The Road to Reality : A complete Guide to the Laws of Universe (2004 ed.). London: Jonathan Cape.

Starbuck W.H; Farjoun M. (2005). Organization at the limit: lessons from the Columbia disaster. Makden, MA, USA: Blackwell.

ISO/IEC 22989:2022-Artificial intelligence AI concepts and terminology. Retrieved May 4th, 2024, from https://www.iso.org/standard/74296.html

The BSI PAS 1881: Assuring the operational safety of automated vehicles – Specification. Retrieved February 25th, 2024, from https://bit.ly/456cRqI

The DSIT. (2024, January 17th). Policy paper: Introducing the AI Safety Institute. Retrieved February 16th, 2024, from https://www.gov.uk/: https://bit.ly/3T86JK8

The NRI Foundation,. (2009, December 20). NRI MORT User's Manual and Fault Tree . Retrieved March 16, 2017, from https://www.nri.eu.com/mort.html

Tuccio W.A PhD. (2011, Spring). Heuristics to Improve Human Factors Performance in Aviation. Retrieved December 14th, 2021, from The Journal of Aviation/Aerospace Education & Research (JAAER): https://bit.ly/3q6R9ih

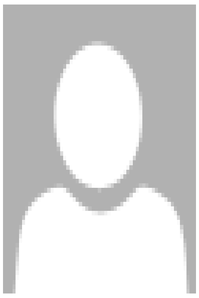
Vaswani, A., et al. (2017,version 1). Attention is all you need v7(2 Aug 2023). The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS), (p. 30). Long Beach. Retrieved February 18th, 2024, from https://arxiv.org/abs/1706.03762

Wiener, N. (1948, Novermber 1st). Cybernetics. Scientific American, 179(5), pp. 14-19. Retrieved February 18th, 2024, from https://www.jstor.org/stable/24945913

——— (1960, May 6th). Some Moral and Technical Consequences of Automation: Science, New Series, 131(3410), 1355-1358. Retrieved February 26th, 2024, from https://www.jstor.org/stable/1705998

Wilson, J.R. (2014). Fundamentals of systems ergonomics/human factors. *Applied Ergonomics, 45*(1), 5-13. Retrieved June 15th, 2017, from https://bit.ly/3cqbA6g

de Winter, J. C. (2019). Pitfalls of automation: a faulty narrative? Commentary on Hancock (2019) Some pitfalls in the promises of automated and autonomous vehicles. *Ergonomics,, 62*(4), 505-508.



Sanjeev Appicharla is an electrical engineering (1983) graduate from formerly known as Karnataka Regional Engineering College, India. An electrical systems designer, systems engineer, a business manager and safety researcher with over 25 years' experience in safety critical industries in India and the UK. Since 2006, he has turned his attention to the theme of System safety and Accident Case Study analysis. He is due to present talks on resident pathogens in systems engineering in the Boeing 737 Max 8 crashes and AV case studies at the INCOSE WSRC 2024 and 2024 HISE INCOSE Conferences. Publications and other details are available at https://orcid.org/0000-0002-8314-7387; https://bit.ly/3AuM2x9.