Safety Engineering Challenges in LLM Era

Ramesh S

General Motors R&D

General Disclaimer

- Based upon my observations
- Not the Opinion of GM
- Tried my best to quote the original source of diagrams wherever possible

Talk Summary

Emergence of AI as important components and tool in engineering next generation systems

- AI based components in ADAS and ADS features
- AI based Tools in the automotive software development life cycle

Safety Engineering Guidelines for `traditional AI' based systems

- Many Standards & guidelines emerging
  - Extensions of ISO 26262 and 21448, TS5083 Appendix
- ISO PAS 8800, USCAR DL-SPICE

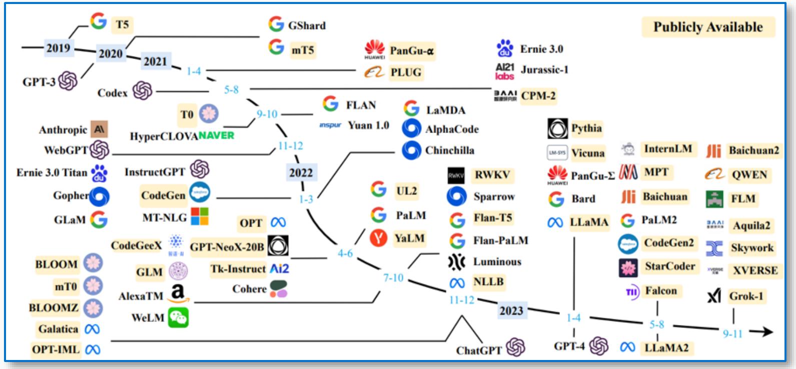
Safety Engineering Challenges for LM based applications

- Do the existing standards/guidelines extend to LLM based Systems?
- Feedback by human/tools
- Agentic Approach - `run time' verification

Next Generation AI based systems

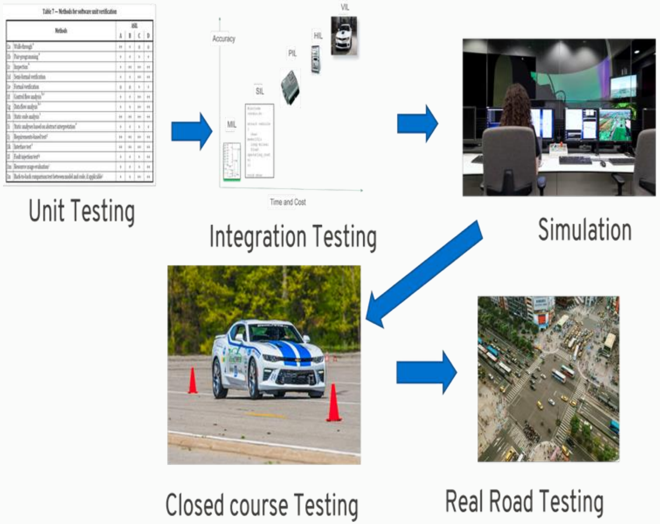
- Leverages GenAI Technologies and advancements in NLP
  - LLM based applications
- Makes use of foundation models
  - Models trained on enormous amount of textual and (in future) multi-modal data
  - Huge and Historical data, Bulkier and General Purpose
- Host of Foundation Models
- AI systems would be applications on top of foundation models
  - Summarization, translations
  - Frameworks for development(LangChain)

AI Applications



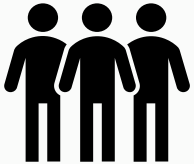
Engineering Automotive Systems - Today

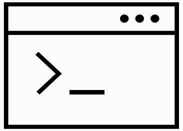
- Engineering Systems involves a variety of artefacts and tools
- Some of these are very formal objects or tools:
  - Models, State Machines, Code, Executable Test Scripts, Verification results
  - Differential equations, solvers, optimization algorithms
- But many are informal objects/tools:
  - Requirements, use cases, scenarios, diagrams, conversations,
  - design constraints, objective functions and criteria and decisions
  - V&V activities and assets

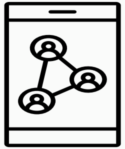


Engineering Automotive Systems - Today

- Humans play a crucial and critical role in translating informal artefacts into formal objects (models) so that they can be processed
  - Gap between informal intention and derived formal extension
  - V&V Validation an important activity to fill the gap

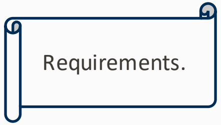


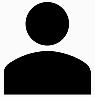


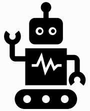


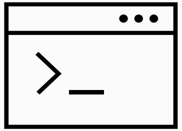
Engineering System – in the emerging AI era

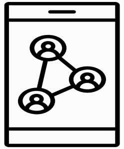
- AI agents and tools to be used extensively
- AI components in end-system components
  - Capability to deal with human based semantic artefacts
  - Inputs and outputs closer to human
  - Human based semantic features consumed
- AI agents in System Development Life Cycle
  - Both processes and Artefacts
  - Requirement Engineering, Design Model Development, Testing, Validation
  - Analysis, Debugging
- Interaction between human and AI agents/tools
  - Human level
  - Manage fuzzy and ambiguous sematic information











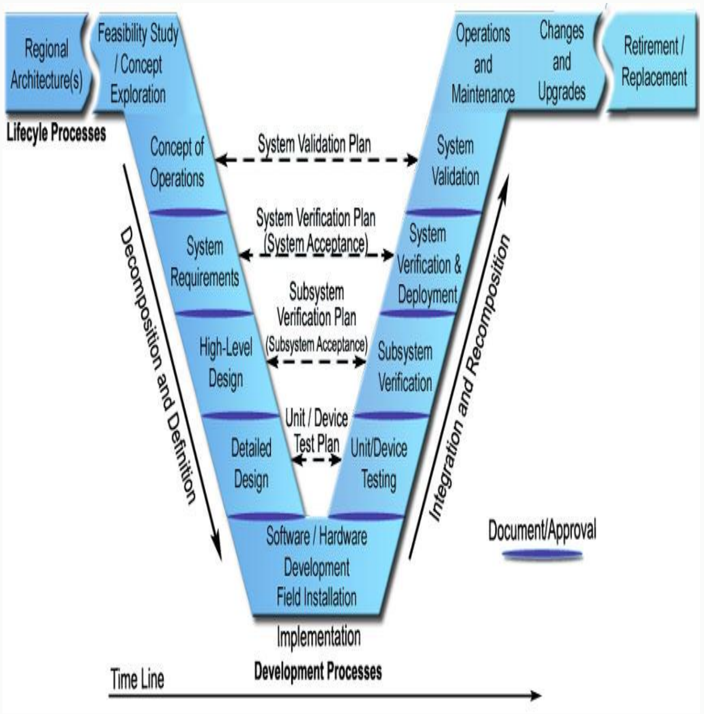
AI Components in End-Systems

- Automated Driving System
  - SAE Level 3 and 4
  - Complex features
  - Perception & Planning
- AI Technology and Machine Leaning (ML) increasingly being used in vehicle applications some of which are safety-related
  - Supercruise uses Camera which uses ML function
  - More L3 and L4 Functions coming
- Deep Neural Networks for perception tasks
- Reinforcement Learning for planning

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
***Human driver* monitors the driving environment**						
0	No Automation	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	Driver Assistance	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	Partial Automation	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/ deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
***Automated driving system* ("system") monitors the driving environment**						
3	Conditional Automation	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	High Automation	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	Full Automation	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

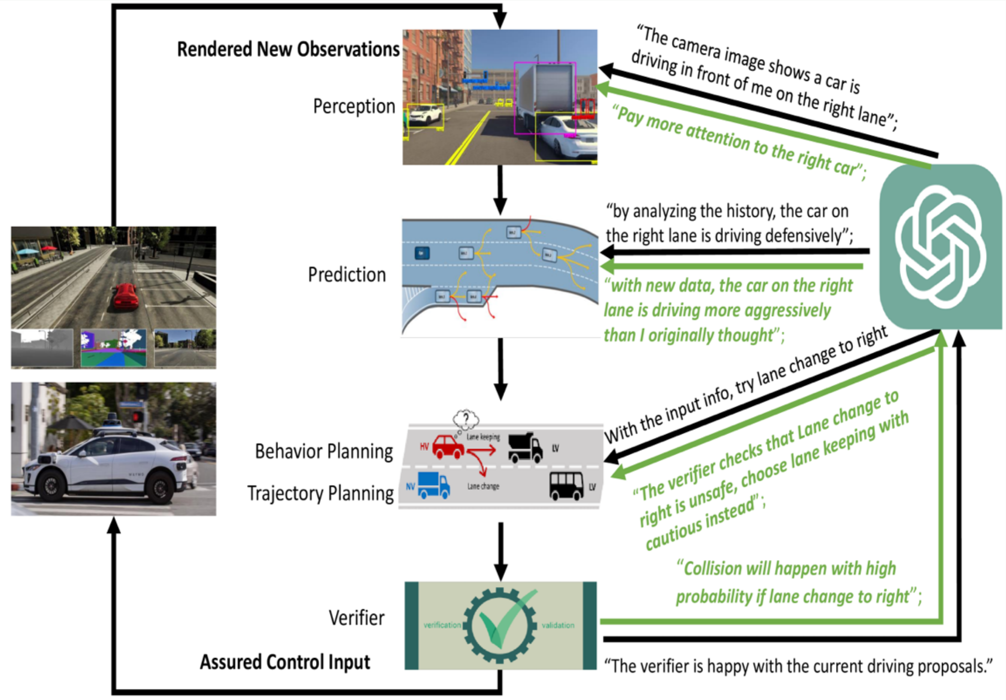
AI in SDLC

- Recent significant advances in LLM
- Extensive experimentation in progress everywhere including GM
- Quickly would be adopted in every aspect of system development life cycle
- GenAI in
  - Code development – code copilot
  - Requirement engineering
  - Safety Engineering
  - testing & verification
  - Scenario generation
  - Simulation
- Neuro-Symbolic Computing and Verification



LLM based Automotive Applications

- Enhancing in-vehicle non-ADS functionality
  - Online diagnosis & prognosis?
  - Engine or Motor control applications?
  - Battery Management
    - SOC, SOH and RUL Prediction
- Enhancing In-vehicle ADS Functionality
  - Navigation, Perception, Planning, Decision Making, On-line verification
- Assisting/Enabling Automotive Life-cycle activities



Source: arXiv:2312.00812v3 [cs.AI] 18 Dec 2023

Data Augmentation using AI

Generated with OpenAI's ChatGPT 4o (all images)

Prompt: *Create a photorealistic image of a traffic scene on a highway with some pieces of gravel on the road.*

Prompt: *Create a photorealistic image of a traffic scene in the city with snow on the ground.*

Prompt: *Create a photorealistic image of a traffic scene on a highway with some shadows on the ground.*

Prompt: *Generate a photorealistic image of a German stop traffic sign with dirt on the sign*

Prompt: *Generate a photorealistic image of a German 50 km/h speed limit traffic sign with a graffiti on the sign*

Prompt: *Generate a photorealistic image of a German construction traffic sign during a snowy winter*

Prompt: *Generate a photorealistic image of a German construction traffic sign during night, with dirt on the sign, and taken with a camera with a steamed over lens*

Prompt: *Generate a photorealistic image of a German stop traffic sign with heavy dirt on the traffic sign, during night, and with steam on the camera lens. Regarding the dirt on the traffic sign, dirt should only be on the image area of the traffic sign, not the background. Regarding night, the traffic sign reflects due to a car's headlights. But the areas with dirt do not reflect as much. The steam on the camera lens should be simulated as if the picture was taken with a camera that has a steamed over lens.*

Generated with OpenAI's ChatGPT 4o (all images)

Prompt: *Generate a photorealistic image of a German danger traffic sign that is damaged* ***and bent***

Prompt: *Generate a photorealistic image of a German pedestrian crossing traffic sign that is strongly* ***twisted***

Prompt: *Generate a photorealistic image of a* ***German yield traffic sign*** *during night, with dirt on the sign, and taken with a camera with a steamed over lens*

Prompt: *Generate a photorealistic image of a* ***German priority road traffic sign*** *with heavy dirt on the traffic sign, during night, and with steam on the camera lens. Regarding the dirt on the traffic sign, dirt should only be on the image area of the traffic sign, not the background. Regarding night, the traffic sign reflects due to a car's headlights. But the areas with dirt do not reflect as much. The steam on the camera lens should be simulated as if the picture was taken with a camera that has a steamed over lens.*

LLMs in Safety Engineering

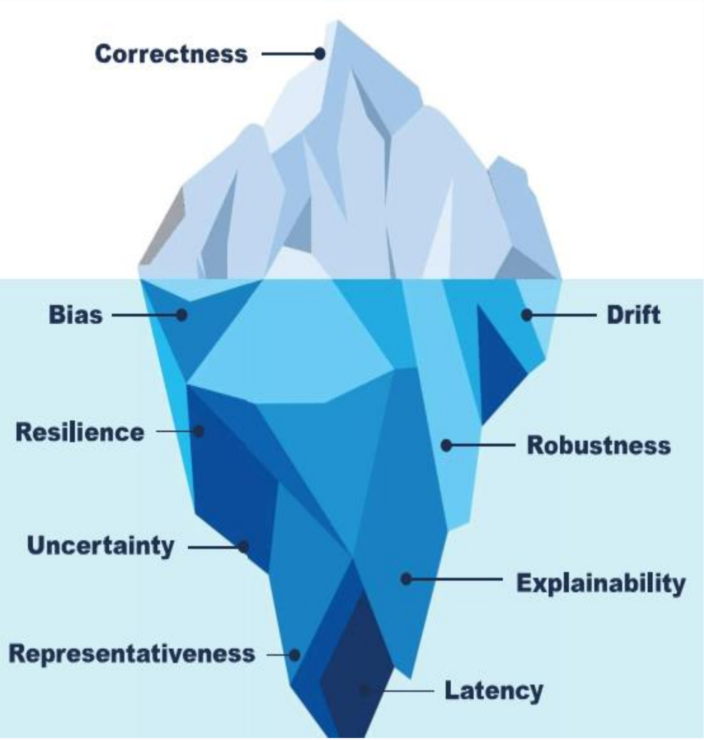
- Some recent papers
  - Huang, Xiaowei, et al. "A survey of safety and trustworthiness of large language models through the lens of verification and validation." Artificial Intelligence Review 57.7 (2024): 175. https://arxiv.org/abs/2305.11391
  - Bullwinkel, Blake, et al. "Lessons From Red Teaming 100 Generative AI Products." arXiv preprint arXiv:2501.07238 (2025). https://arxiv.org/abs/2501.07238
  - Rawat, Ambrish, et al. "Attack Atlas: A Practitioner's Perspective on Challenges and Pitfalls in Red Teaming GenAI." arXiv preprint arXiv:2409.15398 (2024). https://arxiv.org/abs/2409.15398
  - Evil Geniuses: Delving into the Safety of LLM-based Agents, https://arxiv.org/abs/2311.11855
  - AGENT-SAFETYBENCH: Evaluating the Safety of LLM Agents https://arxiv.org/abs/2412.14470
  - AI security risk assessment Best practices and guidance to secure AI Systems, Microsoft Security.
  - Durante, Zane, et al. "Agent ai: Surveying the horizons of multimodal interaction." arXiv preprint arXiv:2401.03568 (2024). https://arxiv.org/abs/2401.03568
- We ourselves exploring the use in safety engineering tasks
  - Feature Interaction
  - Sotif Scenario Generation

LLM and GM

- Great Impact among many groups
- Several Exploration studies in R&D
- Across the domain
  - Code copilot has become common place
  - Manufacturing Robotic Programming
  - Root Cause Analysis
  - Engineering Design
- SDLC
  - Requirement Engineering
  - Requirements to Test Cases
  - Test Cases to Test Scripts
  - Code refactoring

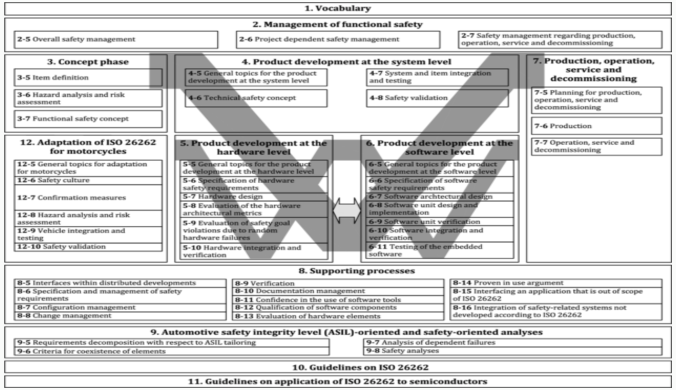
Systems Engineering Challenges in the context of AI

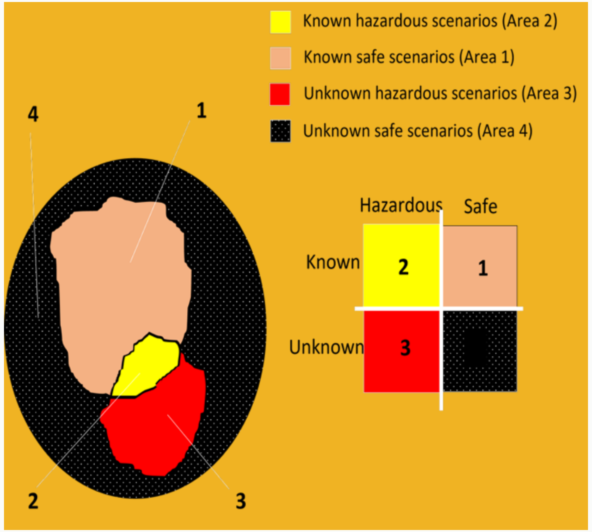
- SE is concerned with building systems (and components)
- Several properties expected of systems
  - Quality, Reliability, Safety, Security, Predictability, Trustworthiness
- The AI Problem
  - Success and Failure are Remarkable
  - Performance (70 – 90% accuracy) suspect
  - Probablistic
  - AI cheats - Learns wrong features
  - AI hallucinates- Explores new features as a learning strategy
  - AI fakes- Bad actors can influence
  - Human based concepts and features
    - inherently ambiguous, inconsistent and incomplete
    - Contextual which could be implicit or explicit
- Review and Revision of Safety Engineering Standards



Safety Engineering and Road Vehicles

- Has a long history
- Two Standards and subsequent revisions
  - ISO 26262: Functional Safety
  - ISO 21448: Safety of the Intended Function (SoTIF)
- Functional Safety
  - Safety under random failures of HW and systematic failures of SW
  - ASIL and elaborate Design, Verification & Validation guidelines
- SoTIF
  - Safety in spite of functional insufficiency or misuse
  - Trigger conditions and Acceptance Criteria
  - Scenario based testing
- Both standards pre-date the recent developments of GenAI
  - Extensions to AI based systems - limited

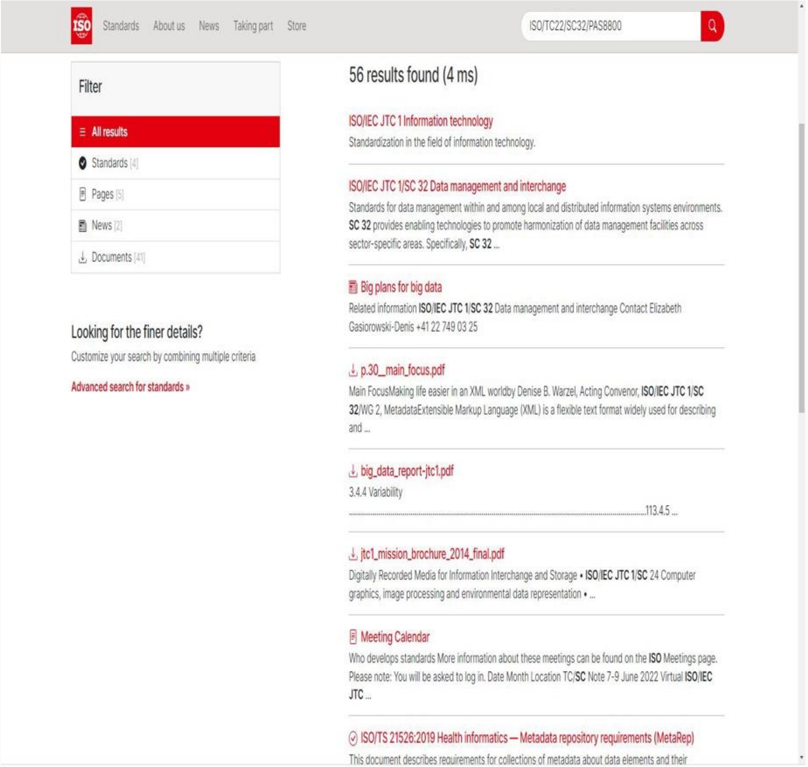




Source: ISO 26262/21448 Documents

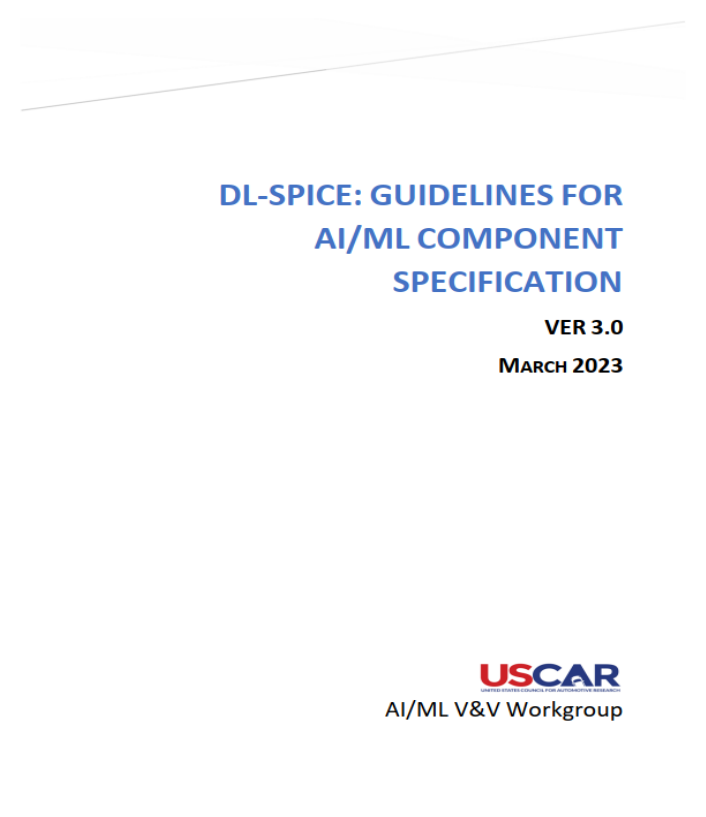
The Standards Landscape for AI based systems

- Quite rich, has been an intense focus for the last few years
- More than 100 guidelines and standards in the general context have come out or under development
  - ISO/TC 22/SC32 – Electrical & Electronic components and general systems aspects
  - ISO/IEC JTC 1/SC42 Artificial Intelligence
- Participating in
  - USCAR DL-SPICE Guidelines Document
  - ISO/AWI TS 5083: Road Vehicles – Safety for ADS – Design, V&V
  - ISP/AWI PAS 8800 Road Vehicles – Safety and AI
- UL 4600 – Safety standard for the evaluation of Autonomous Products



DL-SPICE Guidelines

- Define
  - Life-cycle development, V&V process and the associated artefacts
  - Metrics for the process and artefacts
- The expectation is that the suppliers provide these assets, besides the end artefact
- The guidelines and the metrics would help the OEMs assess and evaluate the robustness and quality of the system
- The guidelines illustrated and validated with a typical example



DL-SPICE: GUIDELINES FOR AI/ML COMPONENT SPECIFICATION

VER 3.0

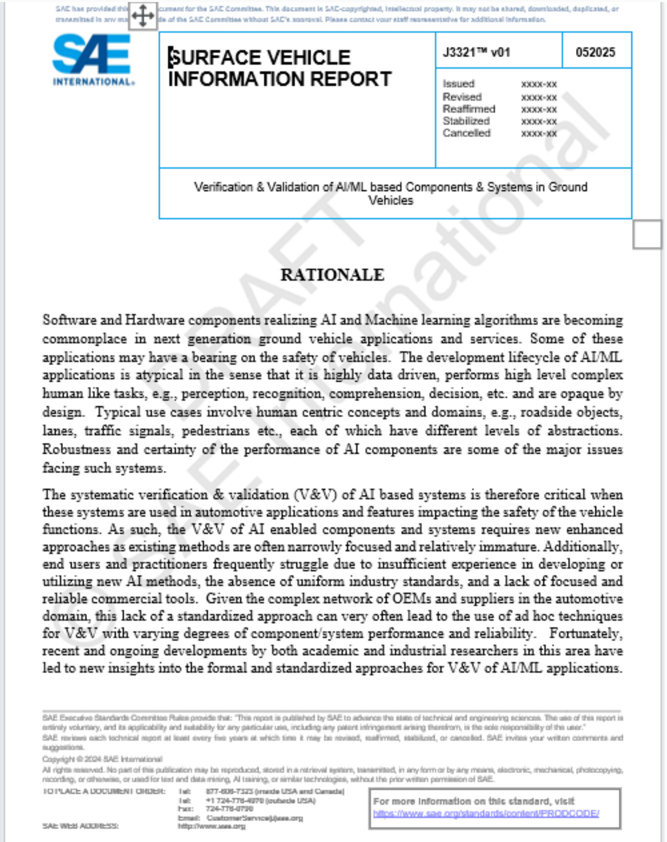
MARCH 2023

USCAR

AI/ML V&V Workgroup

SAE AI Standards Work

- Ground Vehicle AI Committee
  - Use of AI Technology in safe, secure and efficient operation of ground vehicles and transportation infrastructure
- Started a couple of years back
- Several Special Interest Groups and Task Forces ongoing
- A few information reports have come out
  - Terms & Definition, AI Use Cases, AI Data
- Responsible for V&V Task Force
- Soon releasing an information report on V&V of AI/ML
  - SAE J3321



SURFACE VEHICLE INFORMATION REPORT

J3321™ v01	052025
Issued	xxxx-xx
Revised	xxxx-xx
Reaffirmed	xxxx-xx
Stabilized	xxxx-xx
Cancelled	xxxx-xx

Verification & Validation of AI/ML based Components & Systems in Ground Vehicles

RATIONALE

Software and Hardware components realizing AI and Machine learning algorithms are becoming commonplace in next generation ground vehicle applications and services. Some of these applications may have a bearing on the safety of vehicles. The development lifecycle of AI/ML applications is atypical in the sense that it is highly data driven, performs high level complex human like tasks, e.g., perception, recognition, comprehension, decision, etc. and are opaque by design. Typical use cases involve human centric concepts and domains, e.g., roadside objects, lanes, traffic signals, pedestrians etc., each of which have different levels of abstractions. Robustness and certainty of the performance of AI components are some of the major issues facing such systems.

The systematic verification & validation (V&V) of AI based systems is therefore critical when these systems are used in automotive applications and features impacting the safety of the vehicle functions. As such, the V&V of AI enabled components and systems requires new enhanced approaches as existing methods are often narrowly focused and relatively immature. Additionally, end users and practitioners frequently struggle due to insufficient experience in developing or utilizing new AI methods, the absence of uniform industry standards, and a lack of focused and reliable commercial tools. Given the complex network of OEMs and suppliers in the automotive domain, this lack of a standardized approach can very often lead to the use of ad hoc techniques for V&V with varying degrees of component/system performance and reliability. Fortunately, recent and ongoing developments by both academic and industrial researchers in this area have led to new insights into the formal and standardized approaches for V&V of AI/ML applications.

For more information on this standard, visit https://www.sae.org/standards/content/PRODCODE/

ISO/PAS 8800 - Overview

- Industry-specific guidance on the use of AI/ML based systems in safety-related functions of road vehicles
  - Not restricted to specific ML techniques
  - Not restricted to ADS features
    - Annex B of ISO/TS 5083 (under development) adaptation of PAS 8800 for ADS
- Builds on guidance specified in ISO/IEC DTR 5469 (under development)
- Compatible with ISO/IS 26262 and ISO/IS 21448 (SoTIF)
- Harmonizes the concepts in Annex D.2 of ISO/IS 21448


FINAL DRAFT
Publicly Available Specification

ISO/DPAS 8800

Road vehicles — Safety and artificial intelligence

Véhicules routiers — Sécurité et intelligence artificielle

ISO/TC 22/SC 32

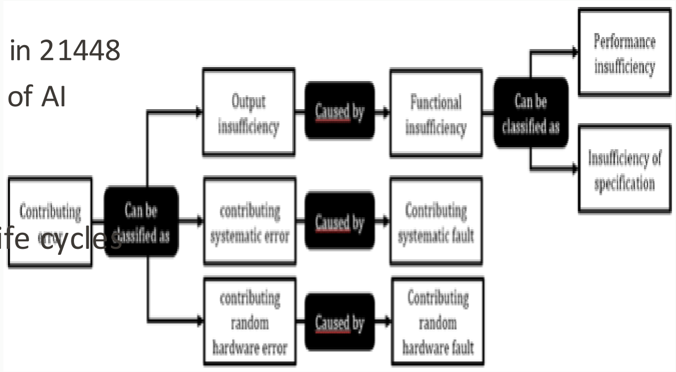
Secretariat: JISC

Voting begins on: 2024-08-21

Voting terminates on: 2024-10-16

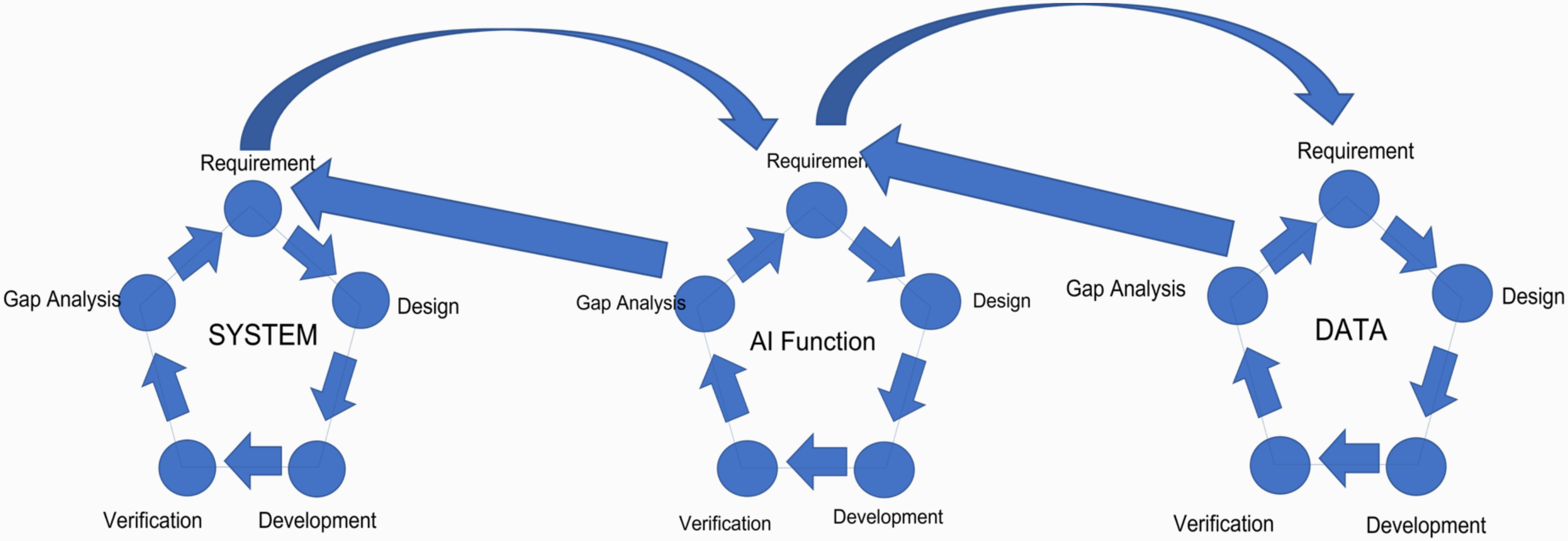
Salient Features of ISO PAS 8800

- Enhancement of ISO 26262 and ISO 21448 process and activities to AI comp.
  - Functional safety-related risks addressed as per 26262
  - Performance Limitation risks by extending the concepts and guidance given in 21448
    - Safety requirements are derived by analyzing performance limitations of AI
- Identification of Two Development Safety Life Cycles
  - AI Component and Data Set safety life cycles
- Monitoring and Identification of Field Issues important component of the life cycles
- Integration of these life cycles with the overall system safety life cycle
- Safety Analysis extended to AI component and Data Set Development
- A number of safety-related properties at the AI component and Data Set level
  - New Notion of Data Set Insufficiency
- Safety requirements include these safety-related properties
- Emphasizes Assurance arguments, besides safety artefacts
  - Safety argumentation includes results of safety analysis of AI system and Dataset Development



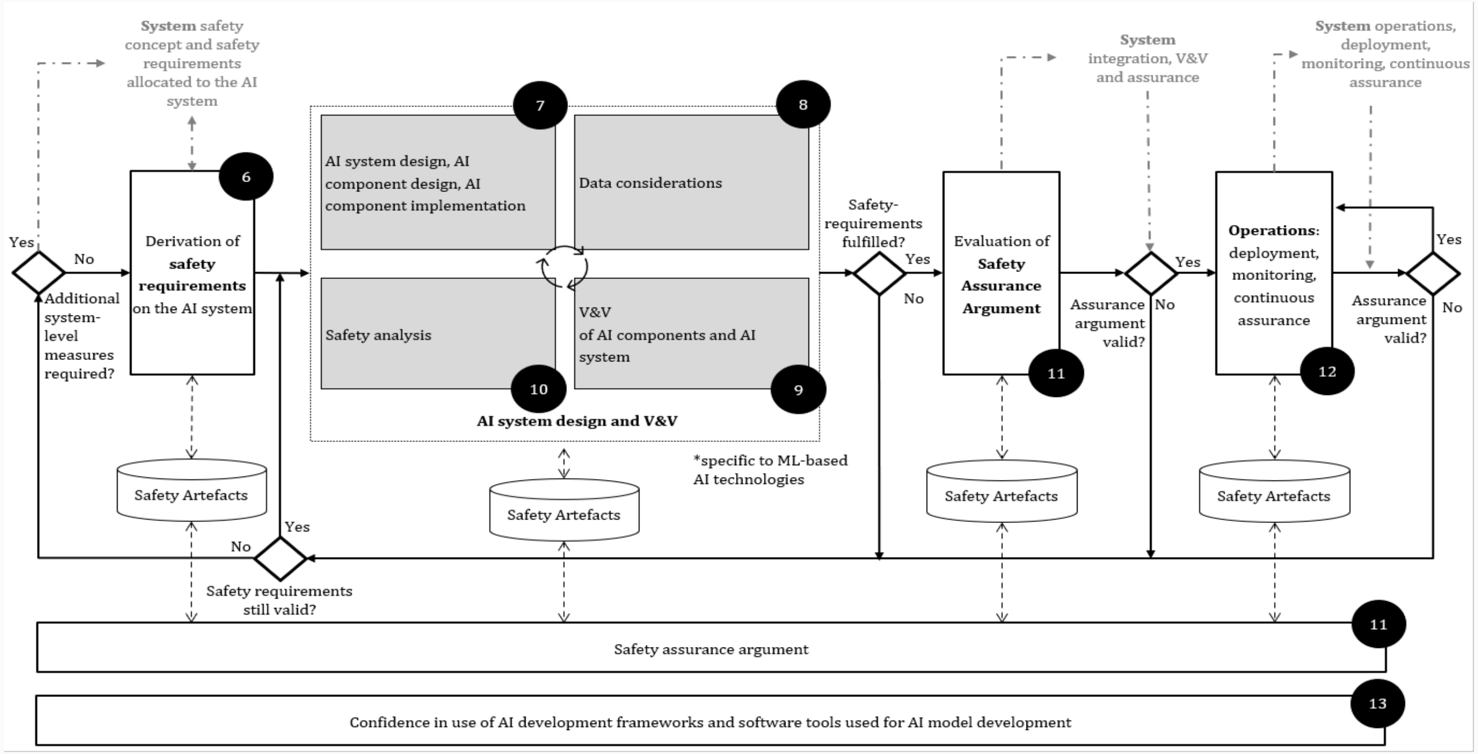
Comprehension of multiple lifecycles

- Traceability from system level to AI and Data life cycle



AI Lifecycle

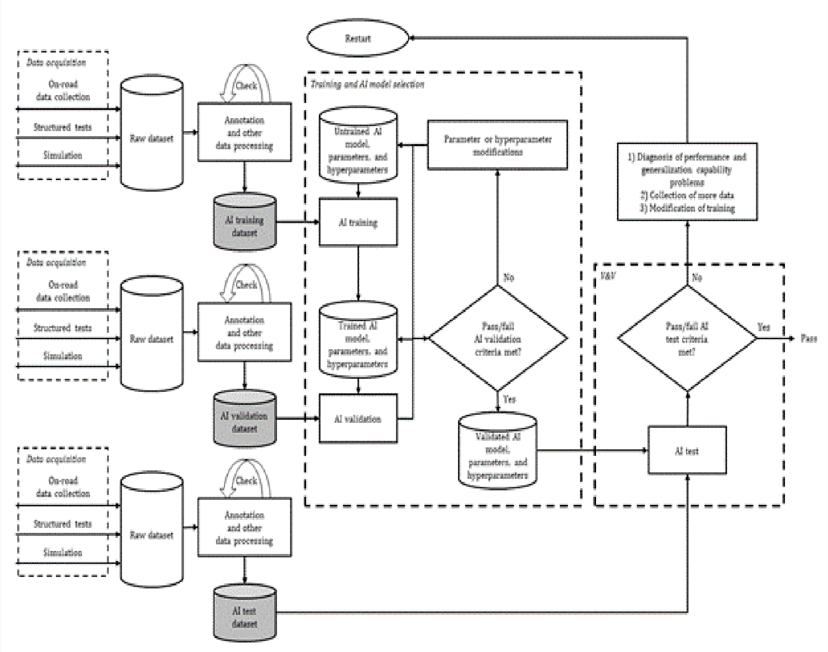
- Exemplary Lifecycle giving rise to Safety Assets forms the basis for the entire document

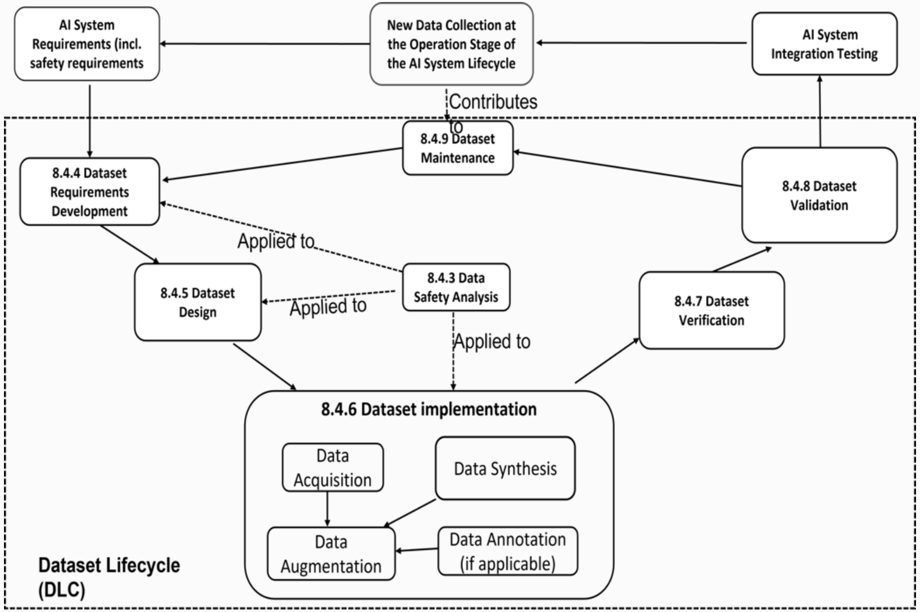


Source: Draft PAS8800 Document

Data Lifecycle

- Data plays a fundamental role in AI system development
  - A dataset lifecycle shall be defined for datasets used in the development of the AI system.
  - The dataset lifecycle shall cover a dataset's requirements development, design, implementation, verification and validation, safety analysis and maintenance.
  - Date-related safety Properties: Integrity, Consistency, Completeness,

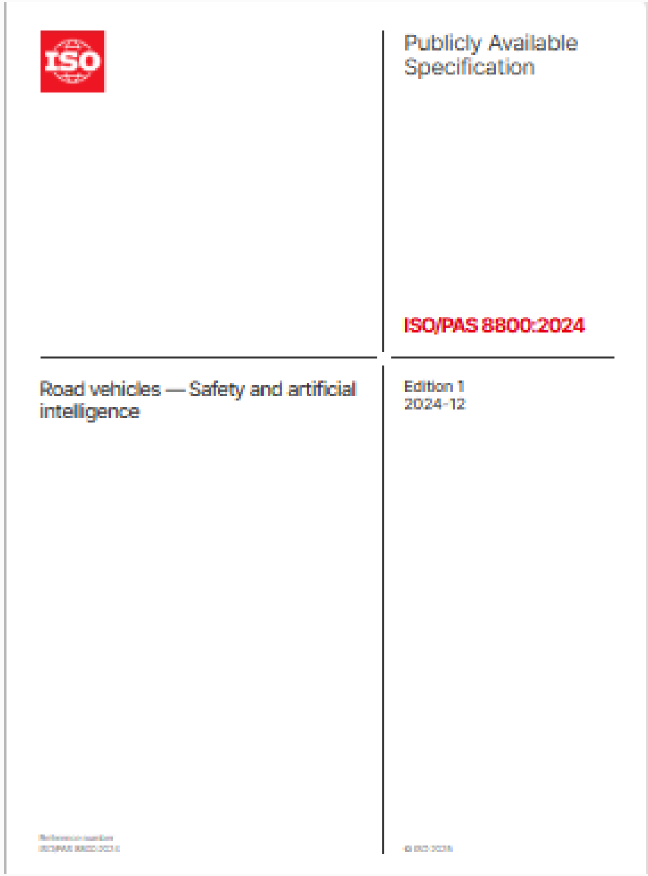




Source: Draft ISO PAS8800 Document

ISO 8800 Status

- Released 12/2024
- Available for use and feedback
- Planning in progress for the next version
- Several Technical Reports and/or extensions planned
  - Use cases,
  - Metrics,
  - AI based Tools
- Guidelines for LLM based applications and LLM use in DLC


ISO

Publicly Available Specification

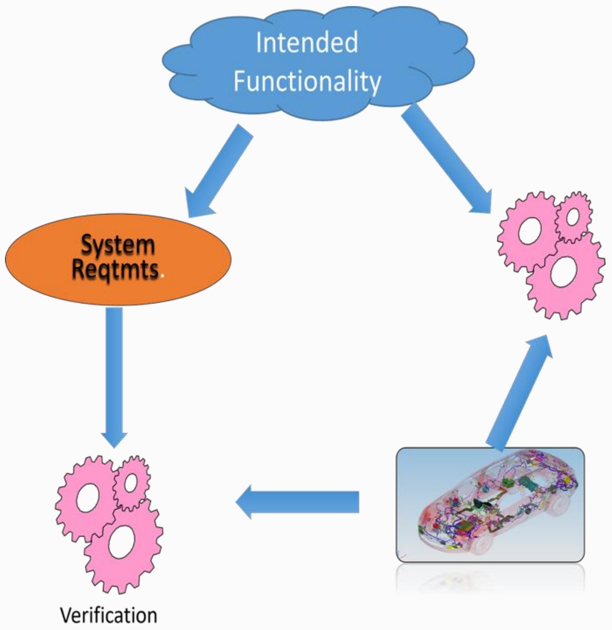
ISO/PAS 8800:2024

Road vehicles — Safety and artificial intelligence

Edition 1
2024-12

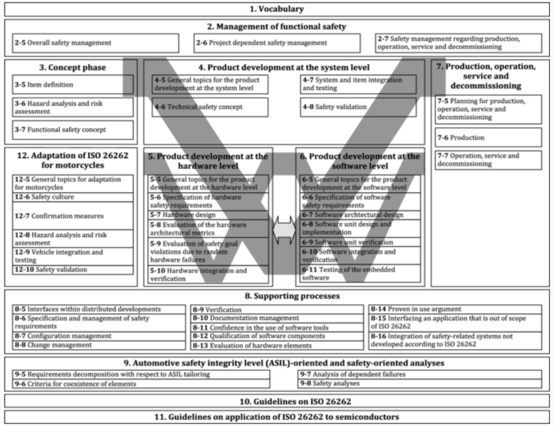
Guidelines for the LLM era

- How do we ensure development of safe systems based upon LLM?
- Rigorous Verification &Validation of Systems
  - Validation is to reduce or eliminate the gap between intension and extension
  - Verification is to check the extension solves the given task
- Verification & Validation problems probably merge
- It is even more crucial
  - Human have less visibility in the system development
  - Humans have a different role to play
- Several Measures in progress
  - Neuro-Symbolic Methods (HSCC2025)
  - ML approach: Guard Rails, Feedback, Reinforcement Learning, providing reasoning

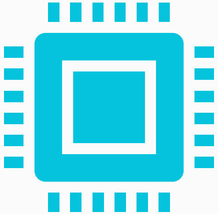


Safety Engineering Guidelines for LLM

- Risk Identification, Analysis & Control/mitigation
- How much of 26262 and 21448 would be applicable/appropriate?
  - Use tool validation guidelines - LLM as a tool
  - Incompleteness issues - Analyze `edge cases'
- Next version of ISO 8800 considering possible inclusion of such systems
- Two possible approaches, as I see it
  - Hide behind the human – make them fully responsible
  - `Run time' verification – verification of outputs
- Overall aim would be to identify
  - Potential risks and possible consequences
  - additional activities, assets and metrics



Verification Challenges



Complex and Large

Billions of parameters

Many tricks for convergence, stability and optimization

Several steps used to build specific tasks

Embedding,

Fine-tuning with and without instructions, single and multitask, parameter efficient (PEFT)

Human Alignment

Tailored to local and dynamic data - RAGs

Verification of LLM based tools

Verification of Prompts and Rigorous prompt engineering	Correctness of prompts Consistency of prompts Completeness of prompts
Verification of Fine Tuning	Similar to traditional AI system verification
Verification of Embeddings	
Verification of use of RAGs	
Most of these are new verification problems	Methods, metrics & tools required

Agentic Approach

- Run-time or on-the-fly verification
  - Verification simpler than synthesis
  - Rigorous Verification possible (Neuro Symbolic Approach)
- Independent Neuro-Symbolic Verifier agent in loop

Talk Summary

Emergence of AI as important components and tool in engineering next generation systems

- AI based components in ADAS and ADS features
- AI based Tools in the automotive software development life cycle

Safety Engineering Guidelines for `traditional AI' based systems

- Many Standards & guidelines emerging
  - Extensions of ISO 26262 and 21448, TS5083 Appendix
- ISO PAS 8800, USCAR DL-SPICE

Safety Engineering Challenges for LM based applications

- Do the existing standards/guidelines extend to LLM based Systems?
- Feedback by human/tools
- Agentic Approach - `run time' verification

Questions