U.S. Department of Transportation

Understanding AI Risks in Transportation

AI Assurance for Transportation Whitepaper Series

Huafeng Yu, Daniel Hulse, and Lukman Irshad

September 2024

This whitepaper was developed in collaboration with the NASA System-Wide Safety Project

Table of Contents

Introduction

Our first AI Assurance (AIA) whitepaper, *An Overview of AI Assurance for Transportation*[1], was published in April 2024. One of the identified key areas for AIA research is *AI risk assessment and mitigation*, which has not been explored sufficiently in the past. To promote and accelerate research efforts to address AI risks, this whitepaper—as the second in the series—provides fundamental information about AI risks and their associated techniques in the context of Highly Automated Transportation Systems (HATS).

The U.S. DOT Highly Automated Systems Safety Center of Excellence (HASS COE) has defined HATS as a type of transportation system that "*makes use of automation to achieve its goals—safety, efficiency, speed, or other benefits—in ways that are beyond the understanding, predictability, or possibly even intervention of highly trained operators*[2]." Examples of HATS include automated vehicles such as self-driving cars and unmanned aerial vehicles, as well as enabling infrastructure such as intelligent traffic management systems and unmanned traffic management. In the context of HATS, AI can support automated vehicle functions such as perception, localization, mapping, planning, and control as well as intelligent traffic management functions such as traffic detection, data analysis, and prediction.

AI-Related Terms Used in this Whitepaper:

- **AI function:** A mathematical function that AI provides to achieve a certain task, such as classification or regression.
- **AI component:** A software or hardware component that implements an AI function, which is part of a system.
- **AI-enabled system:** A system with AI components or other software or hardware components, such as sensors or actuators, to achieve complex tasks or missions in operations. Examples include vehicles with automated driving systems and AI-enabled intelligent traffic management systems.

Despite the demonstrated capabilities of AI/ML technology, very limited studies have been performed to identify, assess, and mitigate the unique risk these technologies may pose to transportation systems. This risk, referred to as *AI risk* in this whitepaper, is a category of operational risk introduced because of the integration, deployment, and operations of AI-enabled systems in the transportation eco-system. AI risks may cause a range of potential *hazards* to vehicles, drivers/passengers, pedestrians, and other stakeholders in the transportation ecosystem, which in turn may affect a number of transportation considerations, including safety, performance, security, privacy, reliability, and resilience. For example, safety hazards related to AI may include misdetection of pedestrians and other vehicles, incorrect vehicle localization, and imprecise identification of lanes, which could lead to *harms* such as crashing, staying on an incorrect path, driving in the wrong lanes, or blocking traffic.

[1] H. Yu, T. Lochrane, T. Pham, S. Mandalapu, G. Romanski, and D. Bakar, An Overview of AI Assurance for Transportation, U.S. DOT HASS Whitepaper, April 2024.

[2] https://www.transportation.gov/hasscoe/what-we-offer

To better address these challenges, AI-related hazards should first be identified, assessed, and mitigated systematically before they pose significant risks to streets, highways, or the national airspace. To this end, AI *risk identification, assessment, and mitigation (RIAM)* is an important part of HATS safety assurance. AI RIAM focuses primarily on implementing safety technologies, processes, and practices to address AI risks in the development, deployment, and operations of AI-enabled systems.

Both *qualitative* and *quantitative* approaches may be used to identify and assess AI risks, based on operational experience and lessons learned, as well as real-world, simulation, or external data. Risk mitigation is subsequently performed to reduce the opportunities for these risks to arise and lessen their impact on the transportation system. AI RIAM is critical for the *safety assurance*, either voluntary or mandated *certification*, and safe operation of AI-enabled systems across modes, from aviation to surface transportation.

Diverse perspectives exist concerning the application of AI RIAM—from safety and security, to reliability and resilience, among others—and there are a variety of standards for risk mitigation that apply to different areas of the transportation industry. While it is important for practitioners to understand the details of these standards and techniques, the goal of this whitepaper is to provide an overview of the basics of AI RIAM that may be applied across transportation domains. As such, this whitepaper is considered a starting point when it comes to understanding the basics of AI RIAM, rather than a reference on specific practices to perform in a particular industry or application.

To better understand the risks that AI components pose across the transportation system, it is important to first understand how they may be used in the overall transportation system. The next section describes these uses and how different AI use cases in transportation may translate into different types of AI risks.

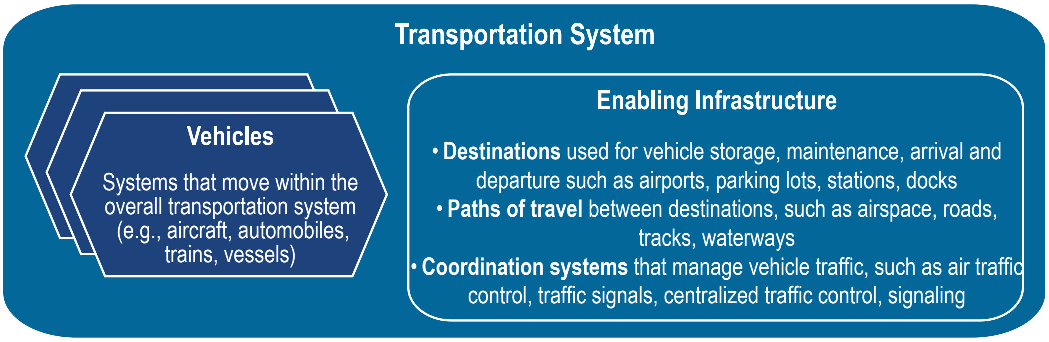
Use of AI in Transportation

Across the transportation sector, there are many potential applications where AI could provide useful functionality, including automated vehicles, traffic management, digital infrastructure, and vehicle and infrastructure maintenance. The specific role that AI plays in the context of an application is a major determinant of the types of risks it may pose to the transportation system. To help inform the understanding of AI risk, this section provides basic information about AI uses in transportation, including contexts, use cases, and operational concepts.

Different Contexts of Potential AI Usage

AI-enabled systems and applications in the transportation system include both vehicles and infrastructure. Figure 1 provides a few examples of these usage contexts and their relevant considerations for how risk should be considered. Specifically, vehicles are manufactured by the private industry and operated by private or public operators, which bear individual and collective responsibility for safety to protect their customers and the overall safety of the public and other stakeholders. Often, such operators help ensure this by following applicable laws and regulations. The infrastructure, on the other hand, is typically designed, operated, and maintained by public or private infrastructure providers in a distributed manner for public use, such as state and local governments. The system-wide responsibility for safety-enabling infrastructure thus rests collectively on policymakers who must balance safety against other considerations, including efficiency and societal costs.

Figure 1. Potential systems with AI applications in the overall transportation context.


Transportation System
Vehicles
Systems that move within the overall transportation system (e.g., aircraft, automobiles, trains, vessels)
Enabling Infrastructure
• **Destinations** used for vehicle storage, maintenance, arrival and departure such as airports, parking lots, stations, docks
• **Paths of travel** between destinations, such as airspace, roads, tracks, waterways
• **Coordination systems** that manage vehicle traffic, such as air traffic control, traffic signals, centralized traffic control, signaling

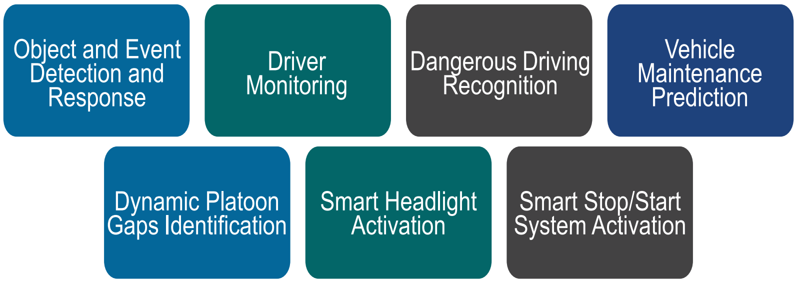
Before the development, deployment, and operations of AI systems, several questions should be considered: (1) who is responsible for the system(s) (e.g., operators, private owners, state/local/federal governments); (2) who uses the system(s) (e.g., operators, private owners, state/local/federal governments) and how they are expected to do so; (3) rules and regulations defining how

the systems can be legally operated; and (4) the cyber-physical form that makes up the operational system (e.g., a moving vehicle versus a static airport supporting many vehicles). These considerations help determine the overall implementation practices that must be followed in the design to mitigate risk, as well as how much leeway the organization may have to mitigate AI risks as they choose.

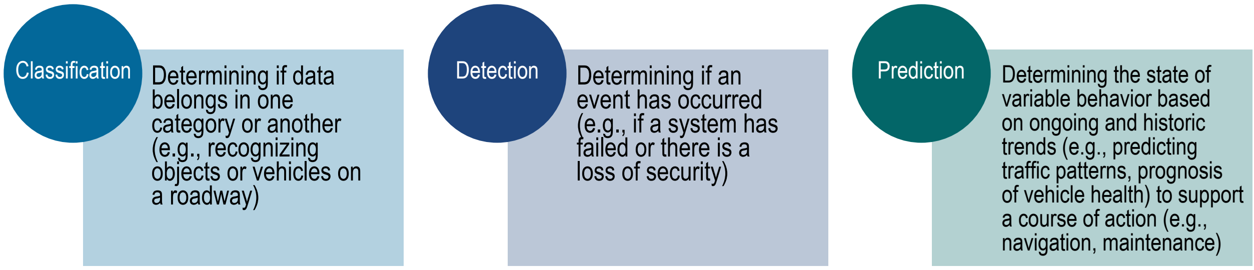
AI Use Cases

An AI use case is a specific function that AI performs within the system concept of operations. Examples of typical AI use cases in transportation systems include predictive maintenance, automated vehicles, vehicle tracking, driver behavior analysis, and traffic management. Figure 2 shows several typical examples of AI applications in automotive advanced driver assistance systems and automated driving systems[3]. These use cases comprise broad functionality that may be performed not just by AI components, but by a larger automated system or subsystem such as an autopilot or maintenance system.

Figure 2. Examples of AI use cases in transportation.


Object and Event Detection and Response
Driver Monitoring
Dangerous Driving Recognition
Vehicle Maintenance Prediction
Dynamic Platoon Gaps Identification
Smart Headlight Activation
Smart Stop/Start System Activation

Within these systems, AI components may perform tasks crucial to the overall function, which may include, but are not limited to:



Classification

Determining if data belongs in one category or another (e.g., recognizing objects or vehicles on a roadway)

Detection

Determining if an event has occurred (e.g., if a system has failed or there is a loss of security)

Prediction

Determining the state of variable behavior based on ongoing and historic trends (e.g., predicting traffic patterns, prognosis of vehicle health) to support a course of action (e.g., navigation, maintenance)

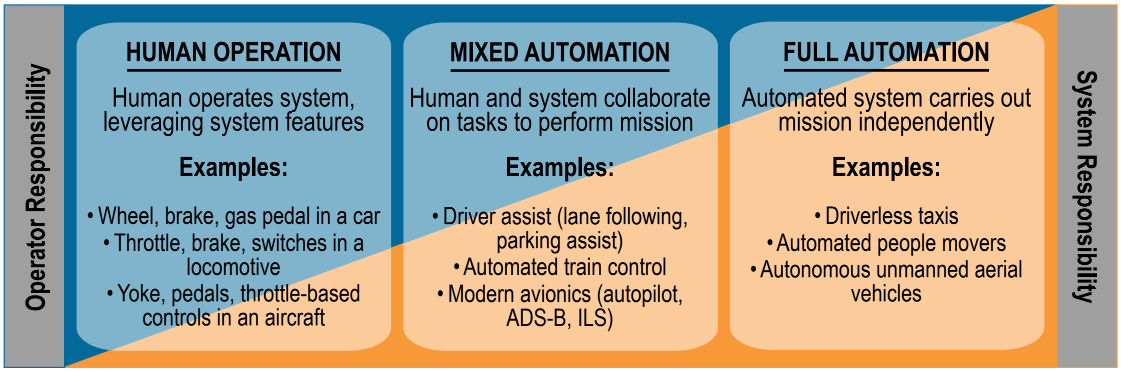
These tasks may be crucial for the operation of a given automated system or function because they enable the automated characterization of system states traditionally performed by an operator to control the system.

[3] M. Vasudevan, et al., Identifying Real-World Transportation Applications Using Artificial Intelligence (AI): Summary of Potential Application of AI in Transportation, Intelligent Transportation Systems Joint Program Office, Report no. FHWA-JPO-20-787, 2020.

AI-Enabled Operational Concepts

AI components can increase the operational automation of transportation systems by automating tasks that otherwise would be performed by human operators (e.g., drivers or pilots). This overall description of how the system is controlled is referred to as an *automation concept*. Because of the central responsibility that human operators traditionally play in mitigating hazards, the degree to which the system is automated can present unique challenges for *risk management*. Typically, in the automated driving domain, the degree of control that is given to the automated system is referred to as the *“level” of driving automation* (codified by SAE J3016[4] and other transportation standards/resources). Between “full human operation” and “full automation,” more discrete levels can be delineated between these two levels. Figure 3 illustrates the basic concept.

Figure 3. Examples of different levels of automation.



HUMAN OPERATION

Human operates system, leveraging system features

Examples:

- Wheel, brake, gas pedal in a car
- Throttle, brake, switches in a locomotive
- Yoke, pedals, throttle-based controls in an aircraft

MIXED AUTOMATION

Human and system collaborate on tasks to perform mission

Examples:

- Driver assist (lane following, parking assist)
- Automated train control
- Modern avionics (autopilot, ADS-B, ILS)

FULL AUTOMATION

Automated system carries out mission independently

Examples:

- Driverless taxis
- Automated people movers
- Autonomous unmanned aerial vehicles

In a human-operated system, the operator directly controls the functions of the system to achieve a mission, while in a fully automated system, the system carries out the mission independently. In a mixed-automation concept, the operator and the system may share or trade control or information over the system to accomplish certain tasks during a mission.

It should be noted that each of these automation concepts pose different types of risk:

- **Human-operated systems** are subject to human-factors related hazards, such as diminished skills[5] (e.g., the ability to perform complex maneuvers as needed), loss of ability (e.g., sight, hearing, or attention), or undesired intent (e.g., hijacking or cybersecurity breach). These hazards can arise from the challenge of the task itself as well as from human variability, which can make them difficult to characterize and fully eliminate. Additionally, because human-operated systems often leverage an in-situ operator, they inherently contribute to some baseline risk because they introduce the possibility of operator injury.

[4] SAE J3016, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, 2021.

[5] Diminished skills could come from a number of human conditions, such as fatigue, impairment (e.g., drugs or alcohol), distraction, workload, etc. The skills can be manual or cognitive.

- **Fully automated systems** are subject to hazards related to the interaction between the automated system and the operational environment. While automated systems may perform better than human operators at individual well-defined tasks, (e.g., cruise control), they often are not as effective at performing the correct task in a complex operational domain (e.g., city driving). Additionally, fully automated systems are typically seen to have less resilience to system failures or unexpected circumstances than human operators, since they lack the general intelligence to "just know what to do" in these scenarios. However, fully automated systems remove the need for an in-situ operator who may be injured when the system fails, which may reduce the severity of certain hazardous scenarios.

- **Mixed-automated systems** are subject to both the hazards related to fully automated systems and the hazards related to human-operated systems, while introducing a new class of human-automation interaction hazards. This does not mean that these systems pose more risk than solely human-operated or automated systems, but it does expand the scope and complexity of hazards that may occur. For example, the system may be designed in a way that the human operator is able to compensate for the hazards posed by the automated features and vice versa. Understanding human-automation interactions is thus key to understanding the overall risks posed by mixed-automated systems.

As discussed here, the context, use cases, and operational automation concept of AI are important for understanding what kinds of risks AI may pose to the transportation system. The next section briefly introduces the basic concepts of AI hazards and risks. Many of these concepts are closely related to AI contexts, use cases, and operational automation concepts.

Basic Risk Management Concepts for AI

To address AI risk, it is important to first understand the basic concepts involved in *risk analysis*. Risk is a combination of the *probability* of a hazardous event occurring and the *severity* of the outcome of the event. Sources of risks may be performance degradation (e.g., wear and tear), discrepancies between the system and its requirements, external hazardous events (e.g., disasters), or random chance. When these mechanisms cause an event with a potential for harm, it is called a *hazard*. Hazards may be called faults, defects, or errors in different contexts.

AI risk is a specific category of risk that is related to the integration of AI in the system or the deployment and operations of AI-enabled systems. AI risk-related activities include the identification and analysis of the source, dependency, probability, severity, and impacts of AI risks on system functional and operational safety. In the following subsections, we discuss different perspectives on addressing risk and outline a general three-step process to manage risk.

Perspectives on Addressing Risk

While the overall concept of risk is universal, there are different perspectives on risk analysis and mitigation that are used depending on the specific concern, industry, and system(s) involved. These perspectives include:

- **Safety**, which is concerned with minimizing damage to property and harm to people such as injury or death. Depending on the industry, safety may be stringently regulated and must be proven prior to entering system operations.

- **Security**, which is concerned with preventing intentional or unintentional access or control being granted to external actors. This is a relatively new consideration that is becoming more relevant as systems become increasingly connected and automated.

Safety Terms Used in this White-paper:

- **Harm:** Realized damage or injury to people or property.

- **Hazard:** Condition that could lead to harm.

- **Fault:** Hazardous event characterized by the undesired operation of a system element.

- **Severity:** Measure of harm to stakeholders from a realized hazard.

- **Probability:** Likelihood of an event occurring ranging between 0 and 1, where 0 means the event could never happen, and 1 means that the event will certainly happen.

- **Rate:** Expected number of event occurrences over a given duration of time (e.g., operational window or lifecycle).

- **Risk:** Overall expected harm borne from a hazard or set of hazards. Risk is a combination of probability and severity.

- **Safety:** Determination that a set of risks in the system satisfy desired requirements to the overall allowable probabilities of harms to people and property.

• **Reliability**, which is concerned with lowering the rate of component or system failures below a certain threshold. Reliability engineering is a long-standing discipline with standardized tools and techniques.

• **Resilience**, which is concerned with making sure that the system can mitigate hazards as they arise, ensuring safe outcomes, both by proactively avoiding them and by restoring key functionality soon after they occur.

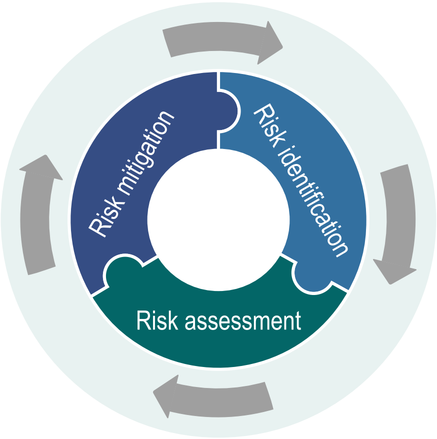
Each of these perspectives can be important for addressing risks, but that does not mean that one is a substitute for another. Particularly, the safety perspective is often important for being able to prove to a regulator that a system is fit to operate, while reliability and resilience perspectives are more important for mission fulfillment and economic considerations.

A major challenge and unifying thread in risk management is the importance of human factors. All AI-enabled transportation systems (even ones with the highest level of automation) will encounter humans in one form or the other throughout their lifecycle. Humans can be both a source of risk and a risk mitigation factor in AI-enabled transportation systems. For example, a human can cause accidents and mishaps or prevent them, depending on the system design and their response to hazardous conditions. As a result, any comprehensive assessment of AI risks will not be complete without the consideration of human elements. This perspective is discussed further in the section, “Human Considerations in AI Risks.”

AI Risk Management

Three steps are usually taken to address potential AI risks in the system and its operations: *risk identification, assessment, and mitigation* (*RIAM*, Figure 4).

Figure 4. Three common steps to manage AI risks.



- **Risk identification** is the first step to identify potential AI hazards in the system or its operations. It often considers AI's context, use, and automation concept as well as the potential events that may arise within its scope of use.

- **Risk assessment** is the determination of the properties of hazards, including their causes and effects, as well as an overall assessment of its probability, severity, and overall risk.

- **Risk mitigation** is the minimization of the risks posed by hazard through dedicated measures such as elimination, prevention, operational avoidance, active monitoring, and contingency management and/or recovery. Risk mitigation is often performed in conjunction with risk identification and assessment to determine the effectiveness of mitigation measures in reducing risk against the new risks they may introduce by failing.

These steps can be performed in both design-time and operation-time of automated systems, as shown in Table 1.

Table 1. Risk-related techniques in design-time and operation-time.

Activities	Design-Time Assurance	Operation-Time Assurance
Risk Identification	Identify occurrence of hazards and risks from the designs of the systems including design requirements, design data, implemented systems, and design environment.	Identify occurrence of hazards and risks in operations, which were not discovered in design time, considering operational requirements and data, human aspects, deployed systems, and operational environment.
Risk Assessment	Assess the properties of AI risks based on design documents, prototype/test data, well-documented knowledge and experience, analytic models, or lessons learned.	Assess the properties of AI risks based on operation-time data and events, which are not predicted during design time.
Risk Mitigation	Reduce or remove potential safety risks by adding safety features to the design, eliminating failure mechanisms, adding architectural features such as redundancy, and/or imposing design requirements (e.g., on reliability or required contingency).	Reduce or remove potential safety risks by entering emergency/safety modes taking the system offline, avoiding hazardous mission profiles, and/or performing maintenance actions.

The following sections describe the RIAM process in more detail as it relates to AI components and AI-enabled systems.

AI Risk Identification

The first step of the risk assessment process is risk identification. Conventionally, this identification is performed by brainstorming potential risks and using operational experience from previous similar systems to create an overall list of hazards. Because AI-enabled systems are relatively novel, there may be a lack of operational experience to base this identification on, making systematic identification processes, outlined here, crucial to addressing risks before the system enters operations.

Preparation for Risk Identification

To prepare for risk identification, the context and use cases of the AI-enabled system should be well-understood, as well as the overall system automation concept. As such, prior to risk identification it may be helpful to gather information related to the tasks(s) being performed and develop an understanding of how the system functionally interacts with its environment (interfacing functions, operators, components, etc.), as exemplified in Figure 5. This usage context will help determine how hazards will affect the operators, other vehicles, and the transportation system as a whole.

Figure 5. Hazard contexts in transportation systems.

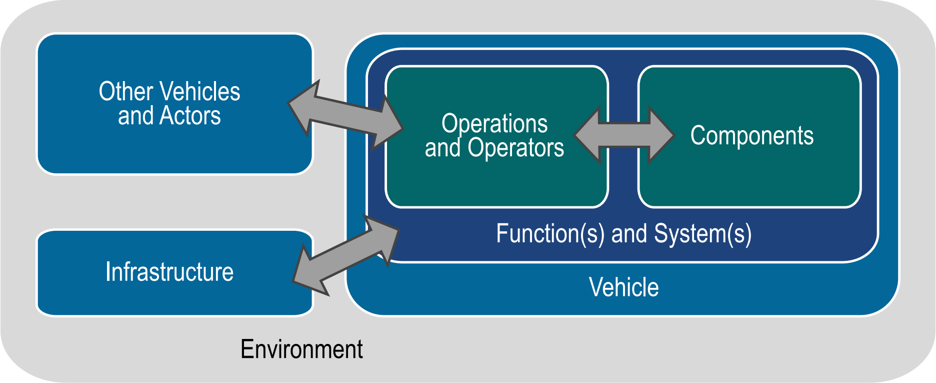
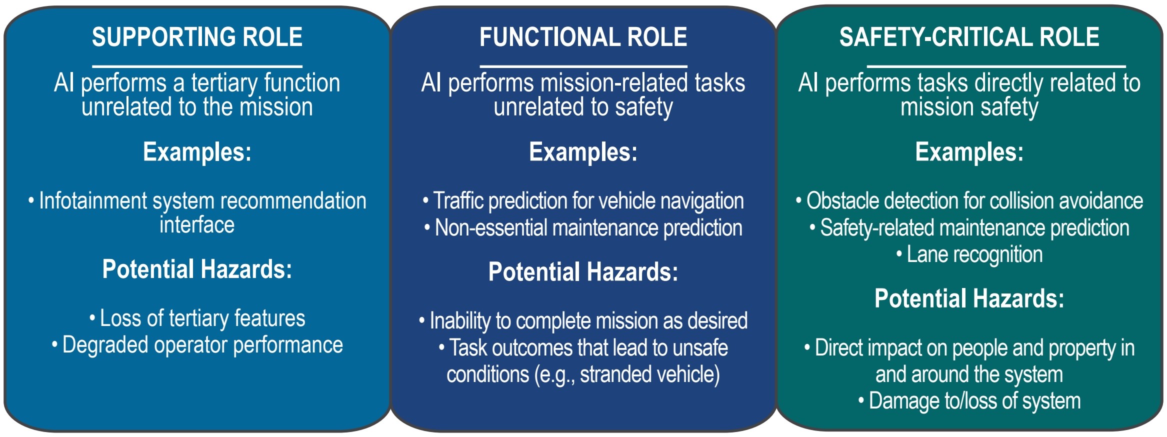


Figure 5 illustrates examples of different usage contexts and their interactions within the transportation system. Within this model of the transportation system, hazards may (1) originate in a particular context of the system, (2) create hazardous interactions between context elements, and (3) cause harm in context elements. Understanding these interactions is crucial for determining how hazardous events may arise in AI functions, as well as the potential effects and harms these events may cause.

Based on the usage of the AI, it should then be possible to determine the type of safety role that the component plays in the overall system. Roles for AI components may be categorized by their criticality to mission safety and operations, as shown in Figure 6. Note that complexity of the distributed system (or system of systems) will tend to increase the difficulty in understanding the roles, interdependencies, and safety risks of each component.

Figure 6. Types of roles given to AI components and prototypical associated hazards for road vehicles.



SUPPORTING ROLE

AI performs a tertiary function unrelated to the mission

Examples:

- Infotainment system recommendation interface

Potential Hazards:

- Loss of tertiary features
- Degraded operator performance

FUNCTIONAL ROLE

AI performs mission-related tasks unrelated to safety

Examples:

- Traffic prediction for vehicle navigation
- Non-essential maintenance prediction

Potential Hazards:

- Inability to complete mission as desired
- Task outcomes that lead to unsafe conditions (e.g., stranded vehicle)

SAFETY-CRITICAL ROLE

AI performs tasks directly related to mission safety

Examples:

- Obstacle detection for collision avoidance
- Safety-related maintenance prediction
- Lane recognition

Potential Hazards:

- Direct impact on people and property in and around the system
- Damage to/loss of system

While classifying the roles of AI components does not itself demonstrate that they are safe (or that a given automation concept is safe), it can help determine the overall risk contribution and risk tolerance for the component, as well as the overall needs of the hazard assessment. In functional and supporting roles, for example, less performance/reliability is necessary from the AI function to achieve safe operations and risk mitigation may be simple. Putting AI functions in safety-critical roles, however, necessitates stringent requirements for risk mitigation.

Hazard Identification

Once the role and context of the AI function has been characterized, its hazards can then be identified. While there are many methodologies that may be used to support the hazard identification process (such as FHA[6], HARA[7], STAMP/STPA[8], and others), the output of these analyses is a table with:

- causes (otherwise known as mechanisms) of hazards,

[6] Defined in SAE ARP-4761A, Guidelines for Conducting the Safety Assessment Process on Civil Aircraft, Systems, and Equipment.

[7] Defined in ISO 26262, Road vehicles - Functional safety.

[8] Nancy Leveson and John Thomas, STPA Handbook, 2018: https://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf

- hazards themselves (e.g., behaviors, interactions, or more general concerns),
- harms that may potentially result from these hazards, and
- other industry or methodology-specific fields based on the standard (e.g., ARP-4761A[9], ISO-26262[10], and ARP-926C[11]) or methodology.

Table 2 provides example outputs of the hazard identification process.

Table 2. Example outputs of the hazard identification process in different contexts for AI components.

Context	AI-Enabled Functions	Hazards	Potential Harms
Vehicle	Localization and mapping	Insufficient situation awareness	Crash, stop of service, adverse navigation
	Vehicle/object detection and avoidance	Avoiding non-existing objects, hitting unseen objects	Crash, stop of service, adverse navigation
	Planning and control	Violation of operational rules	Crash, stop of service, illegal driving behavior, interruption of traffic
Human Interfaces	AI-based human interface	Misunderstanding human intent	Undesired vehicle control input
	Human/automation interface	Undesired control handoff, lack of control handoff when desired	Operator cannot control vehicle or mitigate hazards. Crash, stop of service, etc.
Environment	Traffic control and signaling	Inability to recognize vehicles	Vehicles stopped inappropriately, vehicles directed into occupied space

Once hazard analysis is completed and hazards are identified, additional risk identification analysis may be necessary to ensure all potential hazards were considered in the identification of safety-critical AI components.

[9] SAE Aerospace Recommended Practice ARP4761A, Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment, 2023.

[10] ISO 26262-1:2018. Road vehicles — Functional safety, 2018.

[11] SAE Aerospace Recommended Practice ARP926C, Design Analysis Procedure for Failure Mode, Effects and Criticality Analysis (FMECA), 2018.

AI Risk Assessment

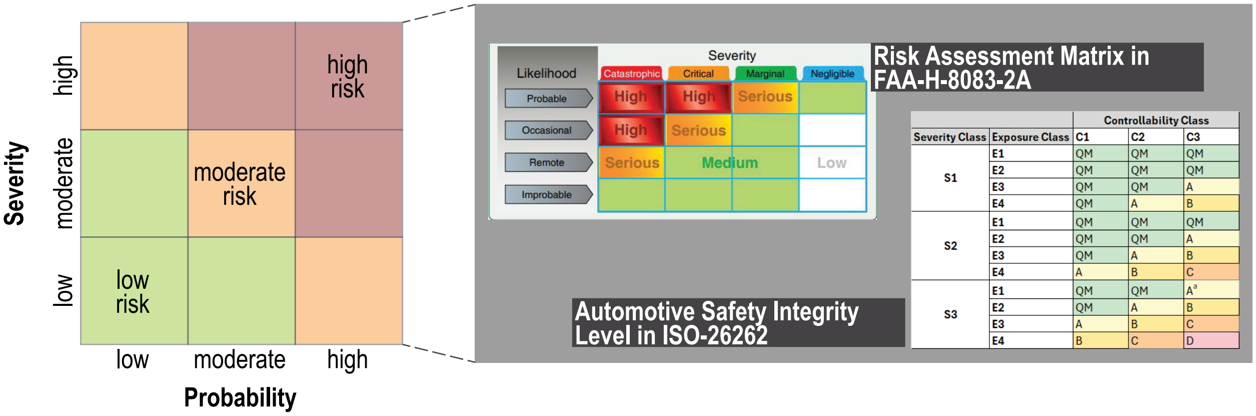
Once hazards have been identified and analyzed for an AI-enabled system, the hazards may be assessed in terms of risk. Risk has two main components: *severity*, which refers to how bad the consequences of a hazard are (e.g., a vehicle loses control and crashes), and *probability*, which refers to how likely it is for the hazard to arise. Severity and probability may be quantified on scales relevant to the industry and subsequently multiplied to determine the risk of the hazard using Equation 1, where R is risk, P is probability, and S is severity.

Equation 1. Assessing risk of a hazard.

$$R = P * S$$

This relationship is illustrated in the risk matrix shown in Figure 7, where high probability/severity hazards pose higher risk and low probability/severity hazards pose less risk. Note that many different standards and frameworks exist for risk consideration that vary by domain, including the levels of severity/probability considered as well as the risk equation itself. For example, in the automotive industry, the *automotive safety integrity level* is calculated as a combination of *severity*, *exposure* and *controllability*, since both exposure and controllability are factors fed into the probability of the severe outcome being realized.

Figure 7. Risk matrix concept and examples in industry.



Risk Assessment Matrix in FAA-H-8083-2A

Likelihood	Catastrophic	Critical	Marginal	Negligible
Probable	High	High	Serious	
Occasional	High	Serious		
Remote	Serious	Medium		Low
Improbable				

Automotive Safety Integrity Level in ISO-26262

		Controllability Class		
Severity Class	Exposure Class	C1	C2	C3
S1	E1	QM	QM	QM
	E2	QM	QM	QM
	E3	QM	QM	A
	E4	QM	A	B
S2	E1	QM	QM	QM
	E2	QM	QM	A
	E3	QM	A	B
	E4	A	B	C
S3	E1	QM	QM	A[a]
	E2	QM	A	B
	E3	A	B	C
	E4	B	C	D

Calculating risk in AI functions is challenging because the probability of failure is difficult to assess. Failure rates in traditional software components are often considered to be “impossible to quantify” because they are considered to arise from design errors, rather than physical mechanisms or variability (RTCA DO-178[12]). However, this concept does not necessarily generalize to AI/ML

because their performance is statistical in nature, which introduces variability into their ability to perform their given function (and therefore, a failure probability). Understanding AI risk probability should thus in part flow out of understanding the statistical performance of these components. However, this may be difficult to calculate because many of the hazards that are expected to arise often come from operating on input data that was not considered in training (i.e., unforeseen situations).

Finally, while the risks of various hazards are often considered individually in hazard identification, it is important to remember that each hazard is only one component of the overall risk posed by the system. This is particularly relevant when the system itself is subject to overall safety requirements on the overall allowable probability of specific harms. Risk assessment should thus be performed at:

- The system level to determine whether overall requirements are met (i.e., that the risks posed by all hazards will meet requirements).

- The individual component hazard level to determine which hazards pose the most risk and thus are the highest priority to mitigate.

Increasing complexity of a system or system of systems, such as AI functions performing Object and Event Detection and Response (OEDR) tasks in transportation systems, makes assessing functional interdependencies of components and subsystems on overall vehicle and system safety outcomes more challenging.

Based on how requirements are met (or not), the risk assessment process motivates the mitigation of hazards, which is explained in the following section.

[12] RTCA DO-178C, Software Considerations in Airborne Systems and Equipment Certification, 2012.

Mitigating AI Risks

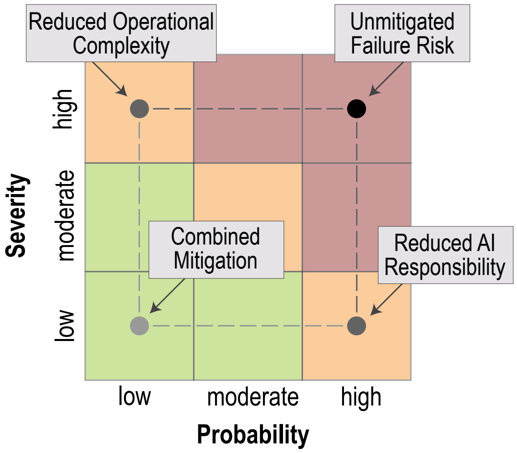
Risk mitigation is the process of reducing the probability and/or severity of hazards in a system so that the overall risks to be borne in operations meets requirements. Reducing risk may occur by reducing the probability and/or the severity of a given hazard. Two general design strategies may be used to reduce the overall risk that AI components pose to system operations—system-level risk mitigation and component risk mitigation.

System-Level Risk Mitigation

System-level risk mitigation for AI is the process of reducing risk that occurs outside of a component but within the interfacing systems and functions that the component interacts with. This mitigation often occurs earlier in the design process, as it involves changing the overall concept and requirements of the system that flow down to the AI design. Reducing risk at this level may take two forms, illustrated in Figure 8:

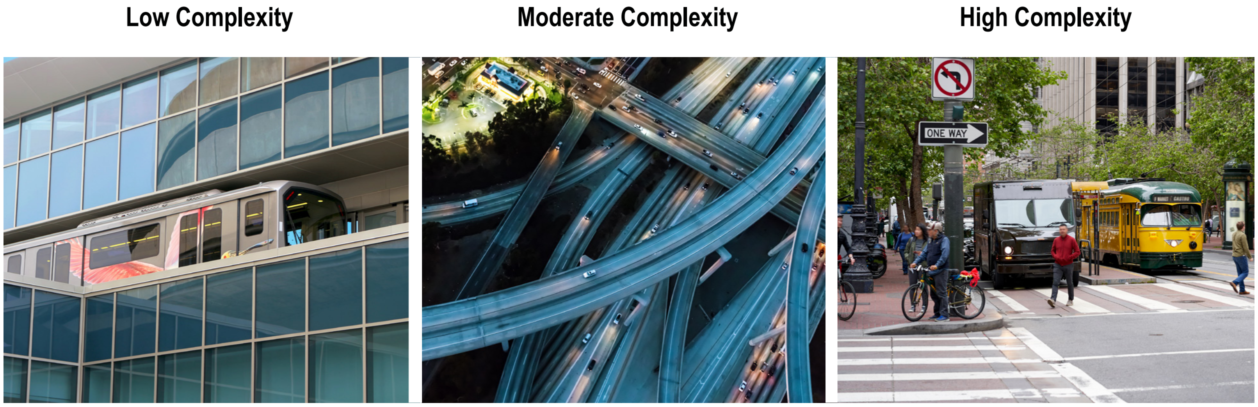
1. Reducing the probability of failure by reducing the operational complexity. This makes the task easier for the AI system to perform, thus decreasing its probability of failure.
2. Reducing the severity of failure by limiting the AI responsibility. This makes it so that AI system failures result in fewer harms, thus decreasing failure severity.

Figure 8. Mitigating AI risks during system function allocation.



Reducing hazard probability by limiting task *complexity* means reducing the *variability* of the operational environment. This is illustrated in Figure 9, which shows low- and high-complexity environments for automation.

Figure 9. Examples of low- and high-complexity automation environments.



Low Complexity

Moderate Complexity

High Complexity

An automated people mover (APM) has low complexity because it operates on a fixed guideway that is protected from potential conflicts.

Driving in a highway context has moderate complexity because the expectations of road use are narrowly-defined, forgiving, and predictable.

Driving in an urban street context has high complexity because of the varying perception and communications tasks the operator must perform to safely interact with the environment.

In practice, examples of reducing environmental variability at the vehicle level may include:

- Operating only on highways instead of busy streets (self-driving cars).
- Operating only in a prescribed right of way (trains).
- Operating only in specific airports/routes where training data has been acquired (aircraft).

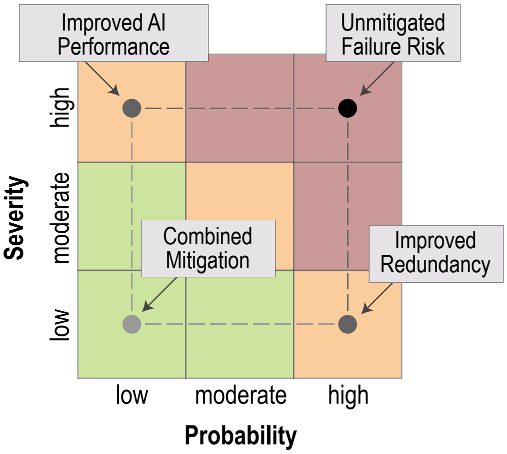
At the infrastructure level, reducing environmental variability is a part of the design of the overall transportation system. Particularly, reducing vehicle conflicts and simplifying the intended uses of infrastructure may make it easier for automated vehicles to perform safely in the overall transportation system.

Reducing hazard severity may be accomplished by reducing the role of AI so that it does not affect the safe operation of the system. This could mean giving the AI-enabled system tasks that have a lower-criticality role (as outlined in Figure 6) assigned to the AI function, as opposed to higher-criticality roles, thereby making failure effects less severe. Often, this is infeasible if the goal or desired innovation of the system is to act on its own. However, this can also be achieved in part by reducing the scope of AI functions to only the tasks where they are necessary (e.g., for perception tasks but not controls or planning tasks) and by relying on external functionality (e.g., operator takeover or redundant safety monitoring systems) to achieve a higher level of safety.

Component Risk Mitigation

Component risk mitigation is the process of reducing risk that occurs within the AI component. Component risk mitigation at this level is thus a matter of mitigating the faults that would cause functional failures in the system, which can be achieved by reducing their probability or severity, as illustrated in Figure 10.

Figure 10. Mitigating AI risks during component and system design or implementation.



In the context of component risk mitigation, reducing the probability of component failure involves improving the performance of the AI components over a wide range of scenarios (including known failure modes, such as optical illusion) such that it is within acceptable bounds for the given task.

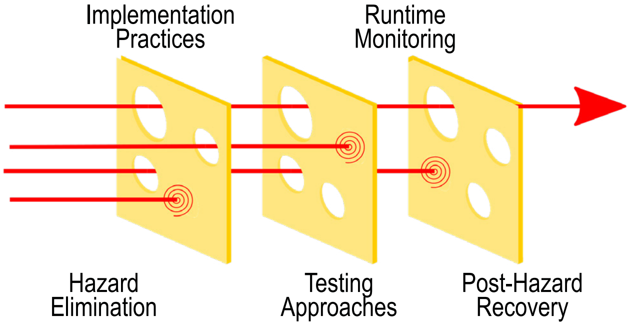
Reducing the severity of component failure, on the other hand, may be achieved in part by reducing the effects of individual AI component failures on the function achieved by the AI-enabled system. This can be accomplished by introducing diverse confirming checks at safety-critical decision points (either alternative AI or traditional functions), redundancy, or fail-safety at the component level (e.g., ensemble models, runtime monitors, self-diagnosis, etc.) to ensure that individual training faults do not result in functional failures for the AI component. To do this, mitigations should ensure that they are not subject to common mode errors—errors in which individual fault modes affect multiple redundant parts (e.g., multiple models in an ensemble failing due to poor training).

Risk Management Strategy

One important concept for system-wide risk mitigation involves defense-in-depth, safety architectures, and the related Swiss cheese model[13], illustrated in Figure 11.

[13] Reason, J. (1997). Managing the Risks of Organizational Accidents (1st ed.). Routledge. https://doi.org/10.4324/9781315543543

Figure 11. Swiss cheese model for AI hazard mitigation approaches (Wikipedia[14]).



In this model, different hazards are mitigated via distinct approaches, meaning that reducing overall risk involves a combination of approaches. This is similar to the concept of a safety architecture, which codifies the causes and effects of various hazards, as well as the mitigating (prevention and recovery) factors that are in place to avoid worst-case outcomes.

Given the diverse sources and effects of hazards, it is recommended to recognize the importance of using multiple types of approaches to maintain and assure safety. Table 3 provides examples of hazard mitigation approaches and their goals.

Table 3. Examples of hazard mitigation approaches.

Approach Type	Practices	Goal of Approach
Implementation Practices	Model validation and explainability	Catch design and implementation errors before they are put in operation
Monitoring Features	Run-time assurance, health management, self-diagnosis	Detect poor in-time performance to activate safe modes and other mitigations
Architectural Features	Multi-model redundancies, algorithmic robustness	Improve AI performance and safety failure tolerance
Hazard/Safety Approaches	Safety and hazard assessment	Systematically identify, track, and address all known safety issues

This section described different methods to identify, assess, and mitigate AI-related risks in the context of an overall risk strategy. The next section will present a necessary consideration for effectively managing AI risks: human factors.

[14] https://en.wikipedia.org/wiki/Swiss_cheese_model#/media/File:Swiss_cheese_model_textless.svg

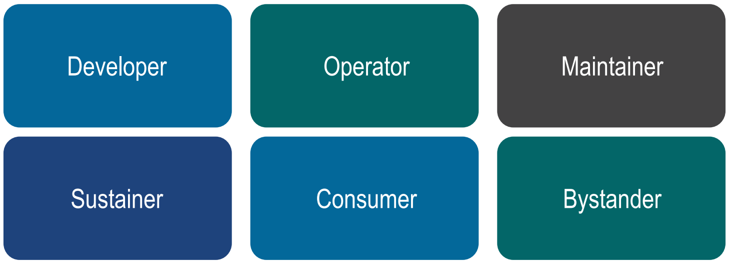
Human Considerations in AI Risks

Because of the key roles that humans play in designing, implementing, operating, and interfacing with AI-enabled systems over their lifecycle, human elements are worth additional consideration in the management of AI risks. To enable the mitigation of human risks in AI-enabled systems, it is important to first understand what roles humans will play in the overall system, the scope of human interactions with AI, and what hazards may arise due to these interactions. These risks may then be mitigated through the processes discussed previously as well as via improved organizational barriers (e.g., safety policies and protocols, company culture, etc.) or human factors requirements (e.g., personnel requirements, operating procedures design, training, interface design, etc.).

Human Actors in AI-Enabled Systems

Human actors may take different roles when interacting with AI-enabled transportation systems. High-level actors include developers, operators, maintainers, sustainers, consumers, and bystanders (Figure 12).

Figure 12. Human actors in the AI lifecycle.



The roles that human actors may assume are not exclusive, meaning that the same people may serve multiple roles. For example, developers may also be maintainers, consumers may be operators, and so on.

- **Developer:** Design, develop, and deploy systems.

- **Operator:** Contribute to the operation of the system locally (e.g., car drivers) or remotely (e.g., ground operations of unmanned aircraft systems).

- **Maintainer:** Maintain and manage changes in the system after the deployment until end-of-lifecycle to ensure optimal performance and safety.

- **Sustainer:** Ensure that AI-enabled systems are safe, ethical, responsible, and usable for intended use cases (for instance, regulators, standard committees, infrastructure providers, and maintainers).

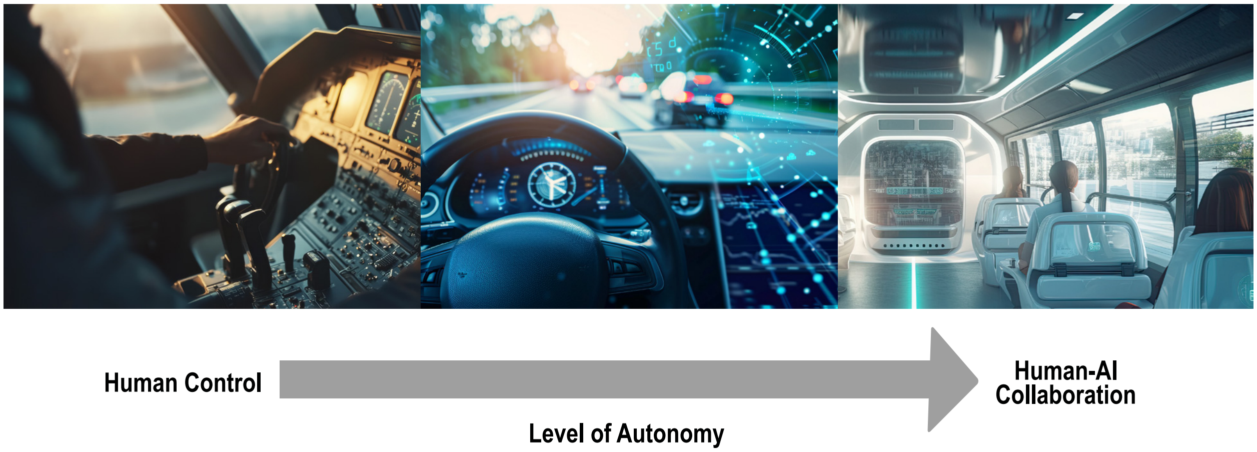
- **Consumer:** End users of AI-enabled systems (e.g., vehicle owners, robotaxi and airtaxi passengers).

- **Bystander:** Any other actors who are not actively involved with the system but who can affect system behavior (e.g., pedestrians for road vehicles, manned aircraft pilots for unmanned aircraft systems).

Scope of Human-AI Interactions

With increasing levels of automation, the paradigm of human-system interactions shifts from human control to human-system collaboration, where the human and AI-enabled systems collaborate to achieve the intended system functions (see Figure 13).

Figure 13. The scope of human-system interactions in automation.

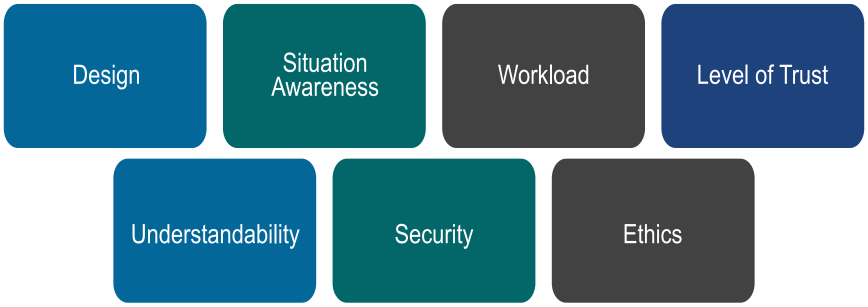


As a result, the scope of interactions for each of the human actors in the system varies with the level of automation. For example, a maintainer of a system with low levels of automation might have to analyze the system state and perform maintenance while relying on preexisting maintenance schedules. On the other hand, a system with high levels of automation will intelligently assess its health state and collaborate with the maintainers to complete the needed maintenance (which may be done by the system itself, with the maintainer providing guidance, or by the maintainer, with the system providing guidance).

Risk Considerations

While the risk factors for human-AI interactions may vary based on the role of the human and the scope of interactions, common high-level risk considerations include (see Figure 14):

Figure 14. Risk considerations for human-in-the-loop.



- **Design:** Design of AI-enabled transportation systems is ultimately a human-driven process that is prone to human error. Human errors that go undetected during the design process (e.g., errors during data labeling or risk assessments) may lead to systems that are poorly designed, which may cause higher risk of failure during operations. Moreover, a poor human factor design could lead to human performance issues (e.g., a confusing user interface).

- **Situation Awareness:** Situation awareness may contribute to safety positively (when appropriate levels are present) and negatively (when deteriorated). In a human-AI collaborative environment, both human and AI components must maintain appropriate levels of situational awareness of each other’s state, intent, and environment. For example, in a self-driving car, the car may hand off control to the driver when the driver is not paying attention if the car is not aware of the state of the driver. Similarly, a driver who is not aware of the car’s state may not be prepared to handle emergencies. However, if both the car and driver have appropriate levels of situational awareness, they may collaboratively prevent failures by complementing each other’s performance.

- **Workload:** High or low workload will negatively affect human performance. In a human-AI collaborative environment, when a human is expected to interact with the system and perform tasks, the workload of the human must be maintained at appropriate levels. This is especially crucial in cases where humans are expected to interact with the system only for contingency management because they go from a state of low workload (before a safety issue) to a state of high stress and high workload (once they are prompted to manage safety). In a poorly designed system, this sudden transition may lead to poor human performance.

- **Level of trust:** In a human-AI environment, too much trust may lead to complacency, while low trust may result in underutilization. If humans place too much trust in AI-enabled systems, they may use the systems in ways that were not intended by design. An example of this is a human driver sleeping in a self-driving car when the car is designed with an expectation of the driver taking control during an emergency. In cases where there is low trust, humans may not use AI-enabled systems in scenarios where they should. For instance, due to low trust, an air traffic controller may not use an AI-based traffic management system that aims to reduce the employee's workload.

- **Understandability:** The design of AI-enabled systems must be explainable to allow humans to understand system behaviors. Lack of understanding in AI system behaviors may lead to mis-calibrated trust and situation awareness, increasing the risk of failure. Similarly, AI-enabled systems should possess the ability to understand human actions during human-AI interactions. Inability to understand human actions may diminish the situational awareness of AI-enabled systems.

- **Security:** AI-enabled systems may encounter misuse and abuse throughout their lifecycle due to factors such as over- or under-utilization, operating outside of operational envelopes, and malice. Humans may either be a source of these vulnerabilities or help prevent them depending on their role in the system.

- **Ethics:** An AI-enabled system that is not ethical can pose risks. For example, a self-driving car that was developed without ethical considerations may break the law, posing safety risks for its passengers. Other ethics-related considerations include accountability, bias, privacy, and security. It is widely accepted that even in a fully automated system, some level of judgement and accountability (generally imposed by the developer as requirements on the system) is necessary to ensure that it comprises an ethical system.

Not all risk considerations have relevance to all human actors. However, when they are relevant, humans may contribute positively (that is, mitigate) or negatively (that is, be a source) to the risk considerations. For example, operators, depending on their skills and experience, may mitigate or exacerbate workload-related risks. The risk considerations themselves may lead to other risks. The level of present risk due to each of these risk considerations may vary during the lifecycle of the system. For example, level of trust may be improved or deteriorate over time depending on how the system performs in operation. Additionally, the level of risk from each risk consideration may vary depending on the individual who is interacting with the system. For example, security-related risks may increase when a novice operator interacts with the system compared to an experienced operator.

Perspective

This whitepaper presents fundamental aspects of AI risks and serves as a starting point for more advanced topics related to AI risks and their management. Because of the high-level nature of this whitepaper, it is also expected that subsequent technical papers in the U.S. DOT AI Assurance Whitepaper Series will cover more advanced topics. Additionally, due to the many unknowns and uncertainties present in the development of AI-enabled HATS, there are still challenges in the main steps of risk identification, risk assessment, risk mitigation, and human-AI considerations:

- **Risk identification**, including the coverage of identified risks. In particular, there is not a way presently to have confidence that hazard identification has uncovered all potential internal and external hazards.

- **Risk assessment**, including modeling impact and propagation of AI risks in complex systems. Because AI-enabled systems play an important role in the control flow of the system and have a potentially large input/output space, as well as stochastic performance, it may be difficult to understand the full risk of automating a given system with AI. This is further complicated by roles in which the AI-enabled system must perform hazard-mitigation actions in rare hazardous circumstances.

- **Risk mitigation**, including the efficiency of mitigation mechanisms. The challenge is validating that active hazard mitigation by AI systems can be effective given interactions between AI-enabled systems, operators, vehicles, and infrastructure. Other challenges include identifying appropriate mechanisms to mitigate unpredicted risks and determining combinations of different risk mitigation mechanisms.

- **Human-AI considerations**, including human interactions and process considerations. The automation of human-performed functions minimizes the ability of humans to mitigate hazards. The challenge is to assess this lost human contribution and understand if the AI-enabled system adequately compensates for it. Rather than being considered a system-level construct, the human elements should be accounted for in component-level design tasks, where risks relating to understandability, trustworthiness, ethics, and security can be better addressed.

Further collaboration between government agencies, industry, and academia to explore, develop, and test AI risk-related technologies may help support overall efforts to address the safety of AI-enabled highly automated systems in transportation. These technologies should cover the following topics: risk identification, risk assessment, risk mitigation, human-AI considerations, data management, social impact, policies, and regulation.

AI Risk-Related Research at HASS COE and NASA SWS

AI risk assessment and mitigation is an important research area of U.S. DOT's AI Assurance Program[15], coordinated by HASS COE. In this program, HASS COE and its partners explore AI risks from several different perspectives, including autonomous aircraft, highly automated systems (surface), and highway traffic management systems, as well as generative AI. Technical topics of this exploration are not limited to those already presented in this whitepaper, but also include support to safety standards, safety impact analysis, and support to potential certification. HASS COE works closely with U.S. DOT Operating Administrations (including the Federal Aviation Administration [FAA] and the Federal Highway Administration) and other federal partners (including the National Aeronautics and Space Administration [NASA], National Institute of Standards and Technology [NIST], and Department of Defense) to advance this research.

The NASA System-Wide Safety (SWS) project[16] conducts research and development on technologies to maintain the safety of the national airspace as it is transformed by new technologies such as AI/ML, advanced air mobility, electric vertical takeoff and landing, unmanned aircraft systems (UAS), and UAS Traffic Management. To this end, the SWS project pursues a number of technical challenges related to design safety, verification and validation, and operational assurance of these concepts in the near- and long-term. The SWS project works in partnership with industry, government (including FAA and NIST), and academic partners to incubate safety technologies from the cutting-edge of research into standard industry practice.

[15] U.S. DOT AI Assurance Program: https://www.transportation.gov/hasscoe/highlights/AI-assurance

[16] https://www.nasa.gov/directorates/armd/aosp/sws/sws-project-leadership

Contact:

Author
Huafeng Yu, Ph.D.
Senior Scientist, HASS COE
huafeng.yu@dot.gov

hass-coe@dot.gov
transportation.gov/hasscoe

Co-authors
Daniel Hulse, Ph.D.
Software Systems AST, NASA
daniel.e.hulse@nasa.gov

Lukman Irshad, Ph.D.
Research Engineer, KBR, Inc./NASA
lukman.irshad@nasa.gov

Communications
Denise Bakar
Manager, HASS COE
denise.bakar@dot.gov

U.S. Department of Transportation

HASS
Highly Automated Systems Safety
Center of Excellence