Assurance arguments for AI systems [12]

Structuring Assurance Arguments for AI Systems [4]

Exercise: Identifying and Analyzing Context and Assumptions in Assurance Arguments

Scenario: AI-Powered Pedestrian Detection System
You are part of a team developing an assurance case for a machine-learning-based pedestrian detection system used in an autonomous vehicle (SAE Level 4). The system uses a deep neural network trained on labeled camera images to identify pedestrians and trigger appropriate driving responses.
Your safety goal is: Ensure that the vehicle identifies pedestrians reliably and initiates braking or avoidance actions in a timely manner. The system has been tested in urban daylight scenarios using synthetic and real-world datasets from Chinese cities.

Task 1: Define the Context

List at least **four key context elements** that should be explicitly stated in the assurance argument to make the claim about safety **valid and bounded**.
Hint: Think about the technical integration, environment, data, operating conditions, and system-level interactions.

Context Element	Why is it Important?
(Example) System is deployed in urban environments	Limits the validity of safety argument to known and tested conditions

Assurance arguments for AI systems [13]

Structuring Assurance Arguments for AI Systems [5]

Task 2: Identify Assumptions

List at least **four assumptions** relevant to the assurance argument. These assumptions might need to be justified, monitored, or discharged by separate argument threads.
Hint: Think about usage conditions, reliability of inputs, performance limits of the AI, or external systems.

Assumption	Risk if Invalid	How to Address or Monitor It
(Example) Camera input is not occluded or blurred	Missed detection of pedestrians	Camera self-check and fallback to radar

Solution to Task 1: Define the Context

Context Element	Why is it Important?
1. The AI system is integrated into an SAE Level 4 autonomous driving stack	This sets the system-level safety expectations and defines interactions with other subsystems (e.g., braking, trajectory planning).
2. Operating domain is urban European cities in daylight and clear weather conditions	Limits the scope of validity of the safety claim to the tested environments, excluding rural, night, and adverse weather.
3. AI model uses RGB camera input mounted at a fixed height on the vehicle front	Defines sensor modality and position, which influences detection performance and field of view.
4. Training and validation data include a mix of synthetic and real-world data with domain-specific augmentations	Establishes the basis for generalization and highlights dataset-related risks and evidence dependencies.

Assurance arguments for AI systems [15]

Structuring Assurance Arguments for AI Systems [7]

Solution Task 2: Identify Assumptions

Assumption	Risk if Invalid	How to Address or Monitor It
1. Camera inputs are unobstructed, clean, and functioning correctly	If dirty or occluded, pedestrians may not be detected → hazardous behavior	Implement self-diagnostics (camera health check), sensor fusion with radar or LiDAR
2. Operational environment matches the defined training and testing domains (e.g., urban, daylight)	Out-of-distribution inputs (e.g., fog, night, snowy conditions) can degrade performance	Monitor operating conditions and trigger fallback or degrade gracefully
3. The pedestrian detection model maintains performance across hardware platforms (e.g., GPU → embedded chip)	Quantization or hardware changes could lead to different behaviors, reducing reliability	Perform cross-platform validation and calibration before deployment
4. AI model continues to generalize well to unseen pedestrian appearances (e.g., new clothing styles, ages, body shapes)	Model may fail to detect non-standard pedestrians, increasing risk of undetected hazards	Monitor real-world detection logs, perform periodic retraining with updated data