PROTECTING PEOPLE AND PLACES HSE

# A Risk Assessment Approach for Artificial Intelligence (AI) based Systems in UK Industry: Adapting Traditional Risk Assessment Methods for Application to AI-Driven Autonomous Systems

## PEROSH 2025 – 6th Research Conference Manchester

**Dr Colin Chambers**

HSE Science Division

Disclaimer
This publication, and the work it describes, was funded by the Health and Safety Executive (HSE). Its contents, including any opinions and/or conclusions expressed, are those of the authors alone and do not necessarily reflect HSE policy.

# The Challenge - The Rise of AI in Industry

- The integration of AI systems into industrial operations is a paradigm shift.
- AI offers enhanced efficiency, predictive maintenance, and autonomous decision-making.
- However, AI introduces novel risks that traditional safety assessment methodologies may not adequately address.
- GB industrial sectors operate under robust safety regulations, but AI is challenging established risk assessment methods, potentially making it more difficult to demonstrate regulatory compliance.
- Traditional risk assessment methods (HAZOP, FMECA) were designed for conventional, deterministic systems but fail to capture AI-specific risks such as:
    - Algorithmic Uncertainty; Training Data Dependencies; Adversarial Vulnerabilities and Concept Drift

# Adapting Risk Assessment for AI Systems

We propose a new risk assessment approach, specifically tailored for AI-based systems in industrial applications, by adapting well-established methodologies already used in these sectors, including:

- HAZOP (Hazard and Operability Analysis)
- FMECA (Failure Modes, Effects, and Criticality Analysis)
- FTA (Fault Tree Analysis)
- LOPA ( Layer of Protection Analysis)

This combined approach enables systematic identification of AI-related hazards, failure modes, consequences, and mitigation measures effectiveness, aligned with current industrial safety practice, and could be an initial step toward a practical AI-specific risk assessment framework accessible to traditional engineers working with AI specialists.

# Adapting the Core Methodologies

- Modified HAZOP applied to AI systems introduces new guide words like 'Misinterpretation', 'Overconfidence', and 'Drift' to identify AI-specific hazards.
- Modified FMECA applied to AI systems focuses on new failure modes, such as 'Adversarial Attacks' and 'Training Data Contamination'.
- Modified FTA applied to AI systems accounts for factors such as algorithmic failure branches and correlated failures.
- Modified LOPA applied to AI systems evaluates AI related protection layers, including 'Algorithmic Safeguards' and 'Diverse AI Systems', and verifies their independence.

# Case study system, Autonomous Guided Vehicle (AGV)

The AGV System

- Small, self-driving flatbed vehicle operating inside a manufacturing cell
- Mission:
    - autonomously pick up raw materials, deliver them to assigned processing machines, then collect finished parts and move them to the next station
    - navigates a constantly changing shop-floor using AI vision (cameras) and LiDAR; must detect and avoid people, other AGVs, pallets, and moving equipment
- All steering and speed decisions are made by an AI “black-box” that continuously re-interprets sensor data
- No fixed guide wires or magnetic tracks are used

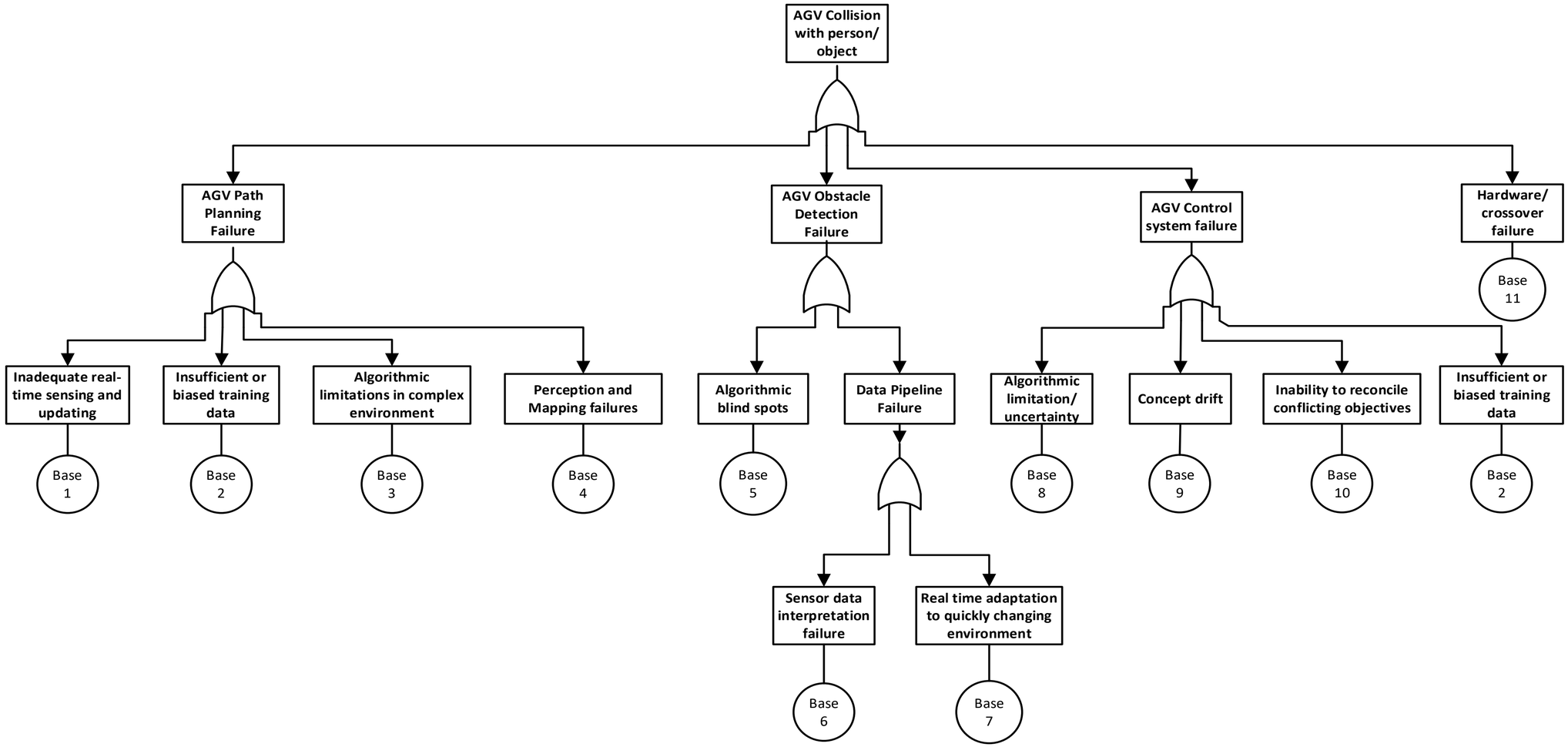
# Conceptual AI-Enhanced HAZOP Worksheet for AGV Operations – an excerpt

| Node /system | Parameter | Guide Word | Deviation | Potential Causes | Consequences | Existing safeguards | Recommendations / action |
|---|---|---|---|---|---|---|---|
| AGV Navigation System | Sensor Input | Misinterpretation | AGV misinterprets reflective floor as open space | Environmental glare; Sensor calibration drift | Collision with fixed structures or personnel; Production line stoppage | Human remote override; Emergency stop buttons; … | Implement real-time sensor data validation; Enhance environmental robustness testing (how well it works in all likely environments) |
| AGV Object Detection | Training Data Quality | Bias Amplification | AGV exhibits biased detection of certain object types (e.g., dark colours) | Insufficient diversity in training data; Unbalanced datasets | Failure to detect critical obstacles or personnel; Near-miss requiring human intervention | Visual alarms; Reduced speed zones | Augment training data with edge cases (limits of likely operation) and diverse scenarios; Implement tracking /detection metrics |
| AGV Path Planning | Model Confidence | Overconfidence | AGV overconfident in obstacle clearance, attempts risky manoeuvres | Inadequate validation of model confidence scores; Lack of uncertainty quantification | Collision with personnel/equipment; Damage to goods; System lockout | Human supervisory control; Geofencing (operation zones) | Implement explainable AI (XAI) for confidence scores; Integrate uncertainty estimation into path planning |
| AGV Control System | Algorithm State | Drift | AGV performance degrades over time, leading to erratic movements | Changes in operational environment; Concept drift; Equipment aging | Unpredictable AGV behaviour; Increased energy consumption; Reduced throughput; Potential minor collisions | Periodic software updates; Performance monitoring | Develop real-time drift detection algorithms; Implement adaptive control strategies |
| AGV Decision-Making Unit | Feature Importance | Generalisation Failure | AGV fails to respond appropriately to previously unseen obstacle configuration | Limited exposure to novel scenarios in training; Overfitting to known data | Inability to navigate around new obstacles; Deadlocks in complex environments; Need for manual intervention | Operator intervention points; Manual override | Implement robust simulation environments; Employ transfer learning for novel scenarios |

# Conceptual AI-Enhanced FMECA for AGV Subsystems

| System/ Component | Function | Failure Mode | Cause | Effect on System | Severity (S) | Occurrence (O) | Detection (D) | Risk Priority Number (S x O x D) | Current Controls | Recommended Actions |
|---|---|---|---|---|---|---|---|---|---|---|
| AGV Vision System | Object Recognition | Contextual Misinterpretation (failure to correctly interpret its environment) | Novel lighting conditions; Unusual object poses | Incorrect path deviation; Failure to detect human in path | 9 | 5 | 7 | 315 | Human monitoring; Basic visual alarms | Implement real-time anomaly detection for sensor data; Enhance environmental robustness testing |
| AGV Navigation Control Unit | Path Planning | Algorithm Convergence Failures (failure to decide action to take) | Local optima in learning; Insufficient exploration in training (failure to find all possible routs) | Suboptimal routes; Erratic movements; Increased travel time | 7 | 4 | 6 | 168 | Periodic model retraining; Route optimisation algorithms | Develop robust reinforcement learning environments; Implement performance monitoring metrics |
| AGV Decision-Making AI | Collision Avoidance | Adversarial Attacks on Sensor Input | Malicious input injection; Exploitable sensor vulnerabilities | Sudden stop/start; Failure to detect real obstacle; False positive obstacle detection | 10 | 2 | 8 | 160 | Basic cybersecurity protocols; Input filtering | Develop robust adversarial training; Implement cryptographic integrity checks on sensor data |
| AGV Data Pipeline | Data Preprocessing | Training Data Contamination | Corrupted data sources; Manual data entry errors; Biased data collection | Degraded AGV performance in specific scenarios; Unpredictable behaviour | 8 | 3 | 9 | 216 | Data validation checks; Version control for datasets | Implement automated data integrity checks; Diversify data collection sources |

# Conceptual AI-Enhanced Fault Tree for AGV Collision



# Conceptual AI-Enhanced LOPA Worksheet for AGV Operations

## LOPA Scenario – AGV object collision detection error leading to collision with object or people

| Initiating event | Frequency of occurrence | IPL 1 | Independence assessment | IPL 2 | Independence assessment | IPL 3 | Independence assessment | IPL 4 | Independence assessment | Mitigated event likelihood |
|---|---|---|---|---|---|---|---|---|---|---|
| AGV navigation error due to concept drift leads to collision with object<br><br>Target frequency 1E-5/yr | 1E-01/yr | AGV internal anomaly detection puts system into a safe state. (Algorithmic Safeguard)<br><br>PFD 1E-01 | Potential common-mode failure if primary AI and safeguard trained on similar data. | Human operator remote override (Human Oversight Layer)<br><br>PFD 1E-01 | Human factor dependency: operator fatigue, latency in response. | Redundant LiDAR-based collision avoidance system (Diverse AI System)<br><br>PFD 1E-02 | Different sensor modality and algorithm architecture reduces common-mode failure. | Hard-wired emergency stop button requiring operator intervention (Traditional Control System Backup)<br><br>PFD 1-01 | Hardware-based, independent of AI software. | 1E-06/yr |
| Human enters AGV path unexpectedly resulting to collision<br><br>Target frequency 1E-6/yr | 1E-01/yr | AGV's internal anomaly detection (Algorithmic Safeguard)<br><br>PFD 1E-01 | Susceptible to adversarial attacks or contextual misinterpretation. | Area scanning safety laser scanners detect error and force stop<br><br>PFD 1E-02 | Traditional deterministic safety system. | Redundant LiDAR-based collision avoidance system (Diverse AI System)<br><br>PFD 1E-02 | Different sensor modality and algorithm architecture reduces common-mode failure. | Physical barriers/fencing in zoned area (Traditional Passive Protection)<br><br>PFD 1E-01 | Passive, physical separation. | 1E-07/yr |

# Case Study Outcome: Key Risk Assessment Outcomes

The risk assessment performed concentrated on AI related hazards. Hardware and crossover faults were not considered.

- Root causes
    - Most AI related hazards originated from data pipeline & algorithmic issues.
- Risk is dynamic & evasive
    - Subtle, often undetected until a ‘potentially critical incident’ occurs.
    - Data quality and concept drift continuously reshape the risk profile.

Traditional safety assumptions appear to break down

- Independence of protection layers cannot be assumed; AI failures can propagate across software boundaries.

- Mitigations
    - Continuous lifecycle management: drift detection, retraining triggers, bias audits.
    - Layered, diverse safeguards: combine algorithmic, human and physical controls.
- Shift from one-time assessments to 'trigger' based and lifecycle-based AI safety management

# AI Risk Assessment Approach - Key Findings

- Traditional safety methods (HAZOP, FMECA, FTA, LOPA) can be adapted for application to AI based autonomous systems
- This approach could support alignment with regulatory expectations for working with simple AI based systems
- Case study provides practical evidence supporting the adapted methodology.

- Adapted assessment methods represent only an initial step in developing a safety management approach suitable for AI based autonomous systems
- AI’s changing risk profile requires ongoing oversight and monitoring methods to be developed

- Traditional risk reassessment methods, periodic or trigger-based, can fail to capture some time-varying risks unique to AI-controlled systems

PROTECTING PEOPLE
AND PLACES

HSE

# ANY QUESTIONS