



Massachusetts
Institute of
Technology

Runtime Safety Engineering for AI-Driven Autonomous Mobile Robots

A STPA-derived Architecture with Safety Supervisor

[Proposed Framework – Development in Progress]

Gianpiero Negri, Ph.D.

Global Head, Robotics Safety Center of Excellence, Amazon Robotics

MIT STAMP/STPA Workshop

Cambridge, March 23rd–26th, 2026



Today's Agenda

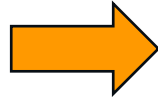
- **Why STPA for AI Robots at Scale**
- **Application Use Case: AI driven AMR**
- **STPA: 4 steps**
- **Runtime Safety Supervisor Architecture**
- **Alignment with AI Regulations and Standards**
- **Current Status and Next Steps**
- **Conclusions and Key Takeaways**

AI Robots at Scale: Why STPA?

Classical safety standards (ANSI B11, ISO 12100, IEC 61508, ISO 13849, etc.) assume:

- ❑ Behavior fully specified at design time
- ❑ Faults enumerable (fault trees, FMEA)
- ❑ Validation is a one-time activity

None of these assumptions hold for AI/ML robotics systems at scale (1M+)



Advantage	Why it's suitable for Robotics AI systems
Control-structure based	Captures AI perception - decision - actuation interactions
Context-sensitive UCAs	Identifies unsafe actions in AI-specific contexts, including nominal operations (failures, system deficiencies/limitations)
Incremental Safety Architecture Definition	Safety requirements emerge from analysis not assumed upfront
Aligned with Robotics AI Regulations and Standards	Operationalizes AI hazard and fault analysis

Application Use Case: AMR in logistics

System: Autonomous Mobile Robot (AMR) in logistic centers

- Transports inventory pods or items between storage and picking stations
- Operates 24/7 in shared human-robot workspace
- Perception Stack: • 360° LiDAR (30 Hz, range 25 m) • RGB-D depth camera (30 fps, forward-facing) • AI/ML fusion: ResNet-50, 180M parameters

Assumption: Baseline architecture has NO mechanism to:

- Detect OOD inputs (e.g., unusual lighting)
- Monitor for data drift or concept drift
- Provide graduated speed control proportional to risk
- Deliver real-time AI performance visibility to the operator
- Guarantee reliable delivery of safety-critical commands

STPA analysis conducted on this baseline to derive the safety requirements from which the Runtime Safety Supervisor will emerge

Operational Design Domain Assumptions:

Parameter	Value
Environment	Indoor warehouse, structured layout
Speed range	0 – 2.0 m/s (nominal 1.5–2.0 m/s)
LiDAR	30 Hz, 360°, nominal range 25 m
RGB-D camera	30 fps, forward-facing
Lighting	1,000 – 10,000 lux (fluorescent + LED)
Human presence	Continuous: pickers, maintenance, supervisors
Fleet density (max)	40 AMRs per zone (ODD-defined limit)
Compute	On-board GPU; no cloud dependency for safety

Application Use Case: Losses, Hazards, System Level Constraints

Loss ID	Description	Severity
L-1	Human worker injured or killed by AMR	Critical
L-2	Damage to warehouse infrastructure or equipment	Major
L-3	Mission failure / operational downtime	Moderate



Hazard ID	Description	Leads to Loss
H-1	AMR in physical contact with, or within close proximity of, a human worker	L-1
H-2	AMR operating at speed in proximity to humans	L-1
H-3	AMR present in a restricted zone or unable to move	L-1, L-2, L-3



SLC ID	Description	Derived from
SLC-1	AMR must maintain distance > braking distance from any human at all times	H-1
SLC-2	AMR speed must be ≤ 1.0 m/s within 2 m of a human; ≤ 0.5 m/s within 1 m	H-2
SLC-3	AMR must not enter or remain in restricted zones under any operating condition and must be always able to move in normal operations	H-3

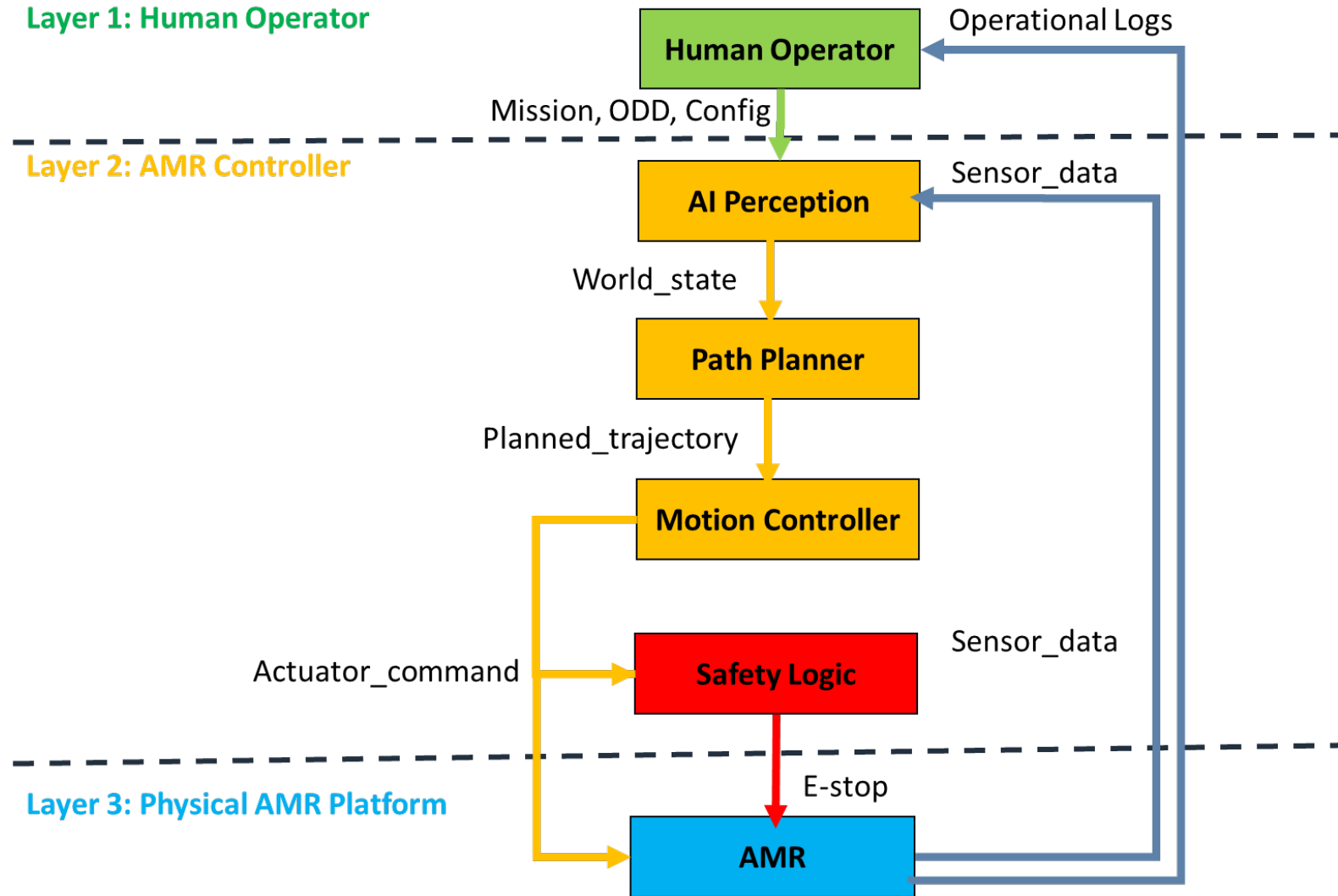
Baseline Control System Architecture

Control Hierarchy (3 levels):

- Level 1: Human Operator
- Level 2: AMR Navigation Controller
 - ❑ AI Perception (SLAM + ResNet-50 fusion)
 - ❑ Path Planner
 - ❑ Motion Controller (PID/MPC)
 - ❑ Local Safety Logic (binary proximity check)
- Level 3: Physical AMR Platform

Key structural observations:

- No intermediate risk assessment layer between AI Perception and Motion Controller
- Human Operator receives only periodic logs (no real-time AI performance visibility)
- Local Safety Logic: binary response only (NOMINAL / E-STOP) with no graduated control
- No OOD detection, no drift monitoring, no acknowledgement-based communication



Unsafe Control Actions

UCA ID	Controller	Control Action	UCA Type	Context	Hazard->Loss
UCA-1	Motion Controller	Stop Command	Not provided	Human within braking distance; AI world state gap or OOD	H-1 → L-1
UCA-2	Motion Controller	Stop Command	Wrong timing (too late)	Human within braking distance; AI latency = 150 ms (fleet overload)	H-1 → L-1
UCA-3	Motion Controller	Stop Command	Applied too long	No human present; object misclassification leading to stop	H-3 → L-3
UCA-4	Motion Controller	Resume Move Command	Started too soon	Human still within safety perimeter (d = 0.8 m); AI confidence recovered	H-2 → H-1 → L-1
UCA-5	Motion Controller	Speed Reduction Command	Not provided	Human within 2 m; no proximity-based speed policy in process model	H-2 → L-1
UCA-6	Motion Controller	Speed Reduction Command	Wrong timing or order	Human within 2 m; speed reduced to 1.5 m/s instead of ≤ 1.0 m/s	H-2 → L-1
UCA-7	Path Planner	Reroute Command	Not provided	Path leads into restricted zone; map stale	H-3 → L-1, L-2, L-3
UCA-8	AI Perception	World State Update	Wrong timing (too late)	SLAM diverged > 0.2 m; world state output based on incorrect position	H-1 → L-1
UCA-9	AI Perception	World State Update	Provided causes hazard	OOD input; high-confidence output with incorrect classifications	H-1 → L-1

Causal Scenarios: Process Model Flaws and Unsafe Interactions

CS ID	Leads to	Category	Description	Causal Mechanism
CS-1.1	UCA-1	Process model flaw/sensor occlusion	Passing AMR-B creates 67 ms LiDAR blind spot; human enters blind spot	Incorrect world model: Stop Command not issued despite human presence
CS-1.2	UCA-1, UCA-9	Training distribution mismatch	Personnel dressing yellow vests	Flawed process model: no component failure
CS-1.3	UCA-1, UCA-9	Gradual sensor degradation	LiDAR SNR: 35 → 18 dB; AI accuracy degrades silently	Process model not updated, no alarm triggered
CS-1.4	UCA-1, UCA-9	Concept drift	Lighting shift 4000K → 5500K; accuracy: 98.20% → 91.70% over 3 months	Process model progressively misaligned: no drift detection
CS-2.1	UCA-2	Emergent interaction - fleet scaling	50 AMRs (designed for 20); latency: 22 → 180 ms	No individual failure: system-level emergent behavior
CS-2.2	UCA-2	Emergent interaction - scene complexity	15 humans + 8 AMRs; total latency = 116 ms; each component within spec	Combined interaction violates SLC-3
CS-3.1	UCA-7	Communication failure	E-STOP lost via EM interference (2.4 GHz); watchdog not tripped	No ACK mechanism: robot unaware of command loss
CS-4.1	UCA-8	SLAM divergence/Perception failure	Moved pallet stack; localization error > 0.2 m; no re-localization trigger	Wrong trajectory from incorrect position estimate
CS-4.2	UCA-8	SLAM sensor fusion conflict	IMU drift 0.8°/s over 4h; inconsistent LiDAR + IMU fusion	No cross-validation: process model staleness undetected

Note: the selected causal scenarios include both degraded-mode and nominal-mode operations. In fact, in several cases the AMR operates exactly as designed, and the hazard arises from unanticipated interactions or incorrect assumptions embedded in the process model.

Controller Constraints/Safety Requirements

Controller constraint/Safety Requirements	Controller	Description	Acceptance Criteria	Addresses
SR-1	Motion Controller	Issue Stop Command within 100 ms (95th pctlile) of human detection within braking distance	95th pctlile \leq 100 ms; 100th pctlile \leq 150 ms	UCA-1, UCA-2
SR-2	AI Perception	Achieve \geq 98.00% Hazard Detection Coverage across all ODD scenarios	HDC \geq 98.00% / 50,000 scenarios	UCA-1, UCA-9
SR-3	Motion Controller, Path Planner	On OOD/drift/degradation: issue Speed Reduction to \leq 0.5 m/s within 2.0 s	100.00% fault injection \rightarrow speed reduction	UCA-5, UCA-6
SR-4	AI Perception	Maintain human detection if one sensor modality fails (single-sensor fallback)	\geq 99.00% single-sensor failure scenarios	UCA-1, UCA-9
SR-5	Motion Controller	Safety interventions \geq 95.00% effective; no new hazards introduced	\geq 95.00% successful; $<$ 2.00% new hazard rate	UCA-3, UCA-4
SR-6	Motion Controller, Local Safety Logic	Safety commands: ACK within 10 ms; retransmit \times 3 if unacknowledged	100.00% delivered within 50 ms	UCA-1, UCA-2
SR-7	AI Perception	SLAM process model synchronized within 0.2 m / 0.1 m/s at all times	99th pctlile position error $<$ 0.2 m	UCA-8
SR-8	Human Operator	Real-time safety KPIs available; KPI degradation $>$ 5.00% \rightarrow investigation within 24h	100.00% of events documented	UCA-1 through UCA-9

STPA Traceability Matrix

Loss ID	Hazard ID	System Level Constraint ID	UCA ID	Causal Scenarios ID	Safety Requirements/Controller Constraints
L-1	H-1	SLC-1	UCA-1, UCA-2, UCA-4	CS-1.1, CS-1.2, CS-1.3, CS-1.4, CS-2.1, CS-2.2	SR-1, SR-2, SR-5, SR-6
L-1	H-2	SLC-2	UCA-5, UCA-6	CS-2.1, CS-2.2	SR-3
L-1, L-2, L-3	H-3	SLC-3	UCA-2, UCA-3, UCA-7, UCA-8, UCA-9	CS-1.2, CS-1.3, CS-1.4, CS-2.1, CS-2.2, CS-3.1, CS-4.1, CS-4.2	SR-2, SR-3, SR-4, SR-6, SR-7, SR-8

Runtime Safety Supervisor

The updated architecture introduces a Runtime Safety Supervisor as a new control level between the Human Operator and the AMR Navigation Controller.

The Supervisor comprises three components:

- **Monitor:** detecting Out-of-Distribution inputs, drift, latency, and sensor degradation
- **Decision Engine:** providing 3 graduated responses: Mode, Speed, Emergency Stop
- **State Predictor:** maintaining a virtual model for look-ahead risk assessment

Safety-critical commands are transmitted via an acknowledgement-based protocol with retransmission, replacing the unreliable UDP channel. The Human Operator now receives real-time Key Performance Indicator data: Hazard Detection Coverage, Response Latency, and Fallback Effectiveness, enabling proactive safety management and continuous process model validation

RSS+ Architecture

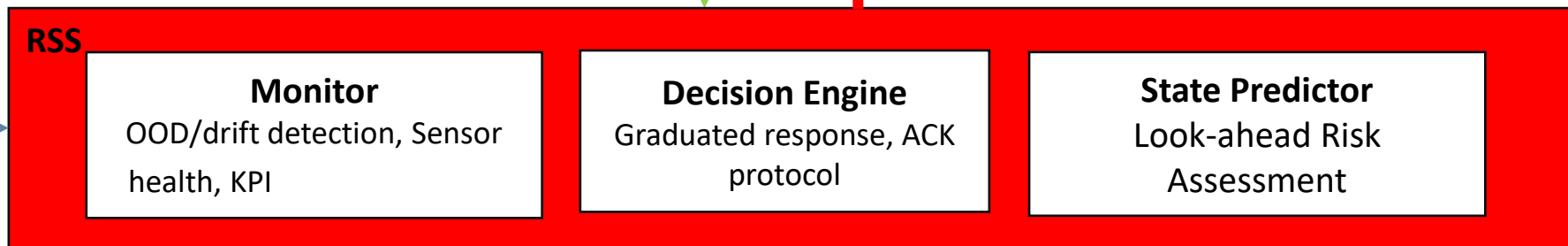
Layer 1



Layer 2

Mission, ODD, Config

Real-time KPIs (HDC, latency, sensor health)

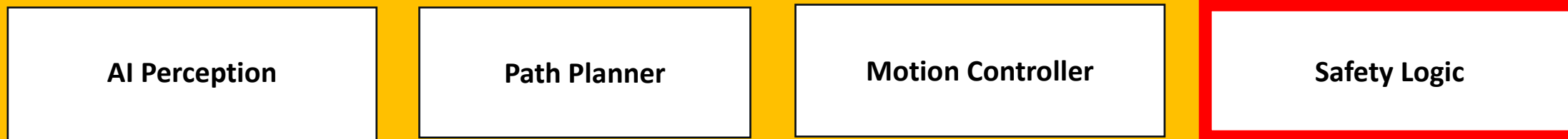


Layer 3

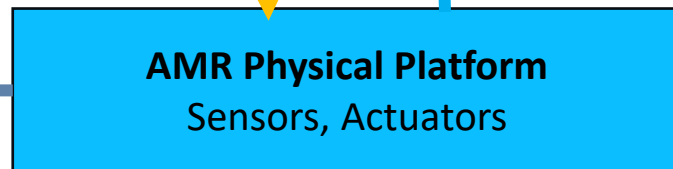
Graduated Control Actions, E-stop

AI state, Manual override

AMR Navigation+Local Safety



Layer 4



Sensor Metrics

Sensor Data

Safety Supervisor vs Baseline: Comparison

SR/CC ID	Description	Baseline	RSS+	How addressed
SR-1	Issue Stop Command within 100 ms from human detection within braking distance	✓	✓	Baseline Motion Controller has dedicated real-time hardware and deterministic execution meeting 100 ms latency without additional components.
SR-2	Achieve $\geq 98.00\%$ Hazard Detection Coverage across all ODD scenarios	✓	✓	Baseline AI Perception with multi-modal sensor fusion achieves 98.00% HDC through extensive training
SR-3	On OOD/drift/degradation: issue Speed Reduction to ≤ 0.5 m/s within 2.0 s	✗	✓	Monitor detects OOD/drift/degradation; Decision Engine issues 4-level graduated response (vs binary baseline) within 2.0 s.
SR-4	Maintain human detection if one sensor modality fails (single-sensor fallback)	✗	✓	Monitor tracks individual sensor health; Decision Engine activates automatic fallback mode maintaining $\geq 99\%$ HDC with remaining sensors.
SR-5	Safety interventions $\geq 95.00\%$ effective; no new hazards introduced	✗	✓	State Predictor performs look-ahead risk assessment; Monitor tracks post-intervention outcomes ensuring $\geq 95\%$ effectiveness and $< 2\%$ secondary hazards
SR-6	Safety commands: ACK within 10 ms; retransmit $\times 3$ if unacknowledged	✗	✓	Decision Engine implements robust ACK protocol with 10 ms timeout and $3\times$ automatic retransmission ensuring 100% delivery within 50 ms
SR-7	SLAM process model synchronized within 0.2 m / 0.1 m/s at all times	✗	✓	Monitor continuously tracks SLAM synchronization error; Decision Engine triggers corrective actions (reset, speed reduction) when 99th percentile exceeds 0.2 m.
SR-8	Real-time safety KPIs available; KPI degradation $> 5.00\%$ \rightarrow investigation within 24h	✗	✓	Monitor computes real-time safety KPIs (HDC, latency, sensor health, SLAM error); automated alerting on $> 5\%$ degradation triggers 24h investigation with 100% logging.

Safety Supervisor vs Baseline: Proposed Validation Approach

Summary: Proposed RSS architecture addresses each of the 8 SRs

How to validate this in practice?

- Preliminary step: Build a simulation-based validation (10k+ scenarios)
- Pilot: 30-day controlled testing (10 AMRs)
- All 8 SR acceptance criteria addressed in preliminary testing
- Next phase: Extended field validation (120 AMRs, 6 months)

Safety Supervisor vs Baseline: Proposed Validation Approach

SR/CC ID	Description	Baseline	RSS+	Validation Method
SR-1	Issue Stop Command within 100 ms from human detection within braking distance	✓	✓	Hardware-in-Loop Testing
SR-2	Achieve ≥98.00% Hazard Detection Coverage across all ODD scenarios	✓	✓	Closed-Loop Simulation
SR-3	On OOD/drift/degradation: issue Speed Reduction to ≤0.5 m/s within 2.0 s	✗	✓	Fault Injection Testing
SR-4	Maintain human detection if one sensor modality fails (single-sensor fallback)	✗	✓	Sensor Failure Simulation
SR-5	Safety interventions ≥95.00% effective; no new hazards introduced	✗	✓	Post-Intervention Analysis
SR-6	Safety commands: ACK within 10 ms; retransmit ×3 if unacknowledged	✗	✓	Communication Stress Testing
SR-7	SLAM process model synchronized within 0.2 m / 0.1 m/s at all times	✗	✓	Ground Truth Comparison
SR-8	Real-time safety KPIs available; KPI degradation > 5.00% → investigation within 24h	✗	✓	Operational Logging Audit

Alignment with AI Regulations and Standards

Standard/Framework	Description	Hi-Level Compliance Assessment
ISO/IEC TS 22440 (expected 2026)	AI Functional Safety	✓ Continuous monitoring, graduated degradation, runtime validation
ISO TR 5469	AI Functional Safety	✓ Monitoring (SR-8), redundancy (SR-4), degraded mode (SR-3)
ISO 3691-4	Driverless industrial trucks	✓ Stopping (SR-1), speed reduction (SR-3), redundancy (SR-4)
NIST AI Risk Management Framework	Govern / Map / Measure / Manage	✓ Continuous monitoring, graduated degradation, runtime validation

Alignment with AI Regulations and Standards

EU AI Act Article	Requirement	Addressed by RSS+
Art. 9	Risk management system	✓ STPA-derived SRs embedded; continuous risk assessment via State Predictor
Art. 11	Technical documentation	✓ Complete traceability: Losses → Hazards → UCAs → Loss Scenarios → CC/SRs
Art. 13	Transparency	✓ Real-time KPI dashboard (SR-8) provides AI state, OOD, drift metrics
Art. 14	Human oversight	✓ Human Operator maintains supervisory control; manual override capability
Art. 15	Accuracy, robustness	✓ Accuracy: HDC ≥98%; Robustness: OOD detection, fallback; Security: ACK protocol
Art. 72	Post-market monitoring	✓ Continuous KPI tracking; automated alerting; 100% event documentation

Current Status and Next Steps

Framework Status

Development in progress:

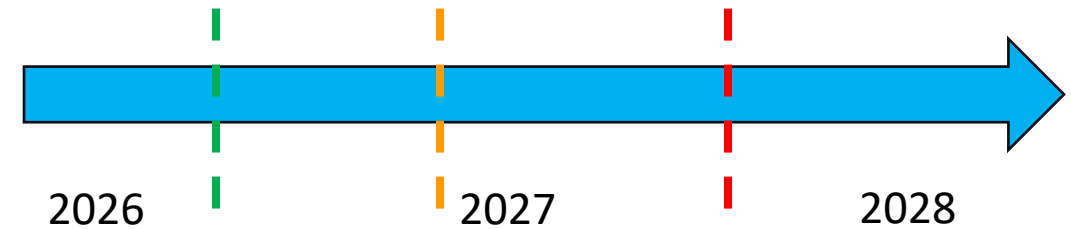
- On-going: STPA analysis, architecture design, simulation
- To be scheduled:
 - 30-day pilot deployment (10 AMRs)
 - Extended field validation (120 AMRs, 6 months)

Next Steps (2026-2027):

- Regulatory readiness and certification
- Integration with fleet management systems
- Production deployment planning

Community Input Welcome:

- Validation methodology for AI-specific SRs
- Scalability considerations for large fleets
- Integration with existing safety standards



Conclusions and Key Takeaways

STPA provides the analytical foundation:

- Identifies unsafe scenarios from both component failures AND nominal operations with unanticipated interactions
- 9 UCAs identified, including 2 specific to AI/SLAM (UCA-8, UCA-9) hardly capturable with traditional FMEA
- 9 Causal Scenarios (selected): majority involve process model flaws, not hardware failures

The Runtime Safety Supervisor addresses operational safety:

- 8 out of 8 Safety Requirements satisfied by the STPA-derived control architecture (RSS+)
- Graduated response (4 levels) vs binary response (baseline): derived from SR-3, SR-5
- Real-time KPI monitoring: derived from SR-8

KPIs verification provides the evidence, preliminary validation demonstrates feasibility:

- Framework Targets: HDC $\geq 98.00\%$ Response Latency ≤ 100 ms, Fallback Effectiveness $\geq 95.00\%$
- Preliminary testing: All targets shall be met in simulation + 30-day pilot (10 AMRs)
- Extended field validation: 120 AMRs, 6-12 months

Thanks for your kind attention!

Any Questions?



Backup Slides



Baseline Control System Architecture

Layer 1: Human Operator



Operational Logs

Mission, ODD, Config

Layer 2: AMR Controller



Sensor_data

World_state



Planned_trajectory



Actuator_command



Sensor_data

Layer 3: Physical AMR Platform



E-stop