From Functional Insufficiency to Behavior-Policy Gate: A Cross-Standards Framework for ODD Boundary Enforcement in Level-4 Robotaxis under Degraded Environmental Conditions

Jherrod Thomas
Independent Researcher — The Lion of Functional Safety
jherrodthomas.com • May 2026

Abstract—Two recent United States National Highway Traffic Safety Administration actions — the May 2026 Waymo voluntary recall 26V-XXX covering 3,791 robotaxis after an unoccupied vehicle drove into a flash-flooded road in San Antonio, and the March 2026 escalation of Engineering Analysis EA26002 covering approximately 3.2 million Tesla Full Self-Driving vehicles in reduced-visibility crashes — together expose a common gap in current Level-4 automated-driving practice: the absence of a structured method for translating perception-availability evidence into an enforceable behavior-policy gate that prevents the vehicle from entering a region of state space outside its Operational Design Domain. ISO 21448 SOTIF, UL 4600, ISO 34502, and ISO/PAS 8800 each supply pieces of the answer, but no single instrument closes the loop from a measured functional insufficiency to a verifiable run-time monitor that gates the planner. We propose a six-step framework that ingests degraded-environment perception evidence, derives ISO 21448 triggering conditions, allocates SOTIF and SOTIF-adjacent functional-safety requirements, and emits a UL 4600 claim-based gate and an ISO 26262 minimal-risk-condition transition. We demonstrate the framework on a worked example combining the Waymo flood-traversal failure and the Tesla glare-blind crash, and provide a numbered fault tree, an availability-aggregation equation, and a comparative table of ODD-monitor styles. The framework is intended as a reviewer-grade input for applicants pursuing Level-4 deployment under the NHTSA Standing General Order on Crash Reporting and equivalent international regimes.

Index Terms—ISO 21448, SOTIF, UL 4600, Operational Design Domain, Level-4 Robotaxi, Behavior Policy, Triggering Condition, Run-Time Assurance, Functional Insufficiency, Minimal Risk Condition.

I. Introduction

The deployment of SAE Level-4 robotaxis in U.S. metropolitan areas has accelerated meaningfully in the 2024–2026 window, with commercial passenger service in Phoenix, San Francisco, Los Angeles, Austin, San Antonio, and Atlanta now supported by fleets exceeding three thousand vehicles in aggregate [1], [2]. Two regulatory actions in the first half of 2026 highlight a structural problem in how these vehicles handle the boundary of their Operational Design Domain (ODD) under degraded environmental conditions. The first is the voluntary recall filed by Waymo with the National Highway Traffic Safety Administration on 30 April 2026, covering 3,791 fifth- and sixth-generation robotaxis after an unoccupied vehicle on 20 April 2026 detected a flash-flooded San Antonio roadway, slowed, and continued into the flood, ending up swept into Salado Creek [1], [2]. The second is the 18 March 2026 escalation of the NHTSA Engineering Analysis EA26002 covering approximately 3.2 million Tesla vehicles equipped with Full Self-Driving (Supervised), in which the agency finds that the camera-only perception stack lacks an adequate degradation-detection function to warn the human supervisor when cameras are blinded by sun glare, fog, or precipitation [3].

The two events differ substantially in technology stack, automation level, and recovery model. The Waymo robotaxi is SAE Level-4 with a multi-modal radar–LiDAR–camera fusion and no in-vehicle human supervisor; the Tesla Full Self-Driving vehicle is SAE Level-2 with a camera-only vision-only fusion and an explicit human-supervisor fallback [3], [4]. The common engineering observation, however, is that both stacks performed *exactly as intended* — both perceived the degraded condition, and both then proceeded with a behavior policy that had not been gated on the perceived condition. In the Waymo case, the vehicle detected standing water and continued at reduced speed; in the Tesla case, the vehicle continued at the commanded operating speed without escalation to the human driver until a crash was imminent [1], [3].

This pattern is precisely the failure type that ISO 21448 *Safety of the Intended Functionality* (SOTIF) was authored to address: the system behaved within the bounds of its intended functionality, but the intended functionality was insufficient for the encountered scenario [5], [6]. The complementary instruments — UL 4600 *Standard for Safety for the Evaluation of Autonomous Products* [7], ISO 34502 *Test scenarios for automated driving systems* [8], ISO/PAS 8800 *Road vehicles — Safety and artificial intelligence* [9], and the SAE J3016 taxonomy of automation levels and minimal risk conditions [4] — each supply pieces of the answer. None of them, however, closes the loop from a measured functional insufficiency to a verifiable run-time monitor that gates the behavior policy.

The contributions of this paper are as follows. First, we propose a six-step framework that ingests perception-availability evidence collected in the Operational Design Domain and produces a UL 4600 claim-based gate, an

ISO 21448 triggering-condition catalogue with derived SOTIF requirements, and an ISO 26262 minimal-risk-condition transition contract. Second, we demonstrate the framework on a worked example that combines the Waymo flood-traversal failure (untraversable-lane class) and the Tesla glare-blind failure (perception-degradation class) into one cross-failure functional thread for an L4 robotaxi. Third, we provide one numbered equation for stack-level availability aggregation, one fault tree for the joint failure mode, and one comparative table of ODD-monitor styles. The framework is intended as a reviewer-grade input for applicants pursuing Level-4 deployment under the NHTSA Standing General Order on Crash Reporting [10] and equivalent international regimes.

The remainder of the paper is organized as follows. Section II reviews prior art in SOTIF, run-time assurance, behavior-policy formal methods, and adverse-weather perception. Section III presents the six-step framework. Section IV applies the framework to the worked flood-and-glare example. Section V discusses limitations, threats to validity, and where the standards lens stops. Section VI concludes.

II. Background and Related Work

0.2.1 A. The SOTIF Lens

ISO 21448:2022 defines a *functional insufficiency* as an insufficiency of specification or performance of the intended functionality, and a *triggering condition* as a scenario condition that activates a functional insufficiency and leads to hazardous behavior at the vehicle level [5]. The standard organizes scenarios into four quadrants — Known Safe, Known Unsafe, Unknown Safe, Unknown Unsafe — and prescribes that the SOTIF activity move scenarios from the right column (unknown) to the left (known) and from the bottom row (unsafe) to the top (safe) through analysis, simulation, and field validation [5], [6]. Khan reviews SOTIF validation methods for autonomous driving and identifies four open problems: triggering-condition completeness, performance-insufficiency injection at scale, the absence of a closed-form coverage metric for the ODD, and the lack of a structured method for migrating residual Unknown Unsafe scenarios into the Known column [11]. Wang and colleagues, in a comprehensive *Engineering* survey of SOTIF for autonomous vehicles, observe that triggering conditions involving rare environmental phenomena — flash flooding, sudden sun glare, snowburst — are the most resistant to scenario-based migration because the underlying physical phenomena are not amenable to bulk simulation [6]. Liu and co-authors describe SOTIF requirement decomposition for perception systems with quantitative performance-insufficiency budgeting [12].

0.2.2 B. UL 4600 and Claim-Based Gates

UL 4600 takes a claim-based approach to safety evaluation of autonomous products, requiring the safety case to enumerate claims, supporting arguments, and supporting evidence — including data integrity, autonomy validation, and metrics for conformance [7]. Edition 3, issued in March 2023, expanded the scope to include autonomous trucking and clarified the treatment of over-the-air software updates [13]. Koopman has argued that the claim-based pattern enables the safety case to remain coherent under the high rate of software change typical of L4 deployments [14]. The standard does not, however, prescribe pass/fail thresholds and does not benchmark the road testing of prototype vehicles [7].

0.2.3 C. ISO 34502 and Scenario-Based Verification

ISO 34502:2022 provides a scenario-based safety evaluation framework for automated driving systems, with a classification scheme that decomposes each ADS function into *recognize, judge, and operate* substages and systematically enumerates the potentially hazardous events for each [8]. The German PEGASUS family of projects supplies the source-corpus methodology that ISO 34502 codifies [15]. Recent work by Hasuo and colleagues formalizes the ISO 34502 critical-scenario language in temporal logic and couples it to the Responsibility-Sensitive Safety (RSS) longitudinal-distance contract [16].

0.2.4 D. ISO/PAS 8800 and the AI Bridge

ISO/PAS 8800:2024 *Road vehicles — Safety and artificial intelligence* fills the gap between ISO 26262 (random hardware faults and systematic design defects) and ISO 21448 (functional insufficiencies of the intended functionality) by addressing AI-specific output insufficiencies, systematic errors, and random hardware errors arising from AI elements [9], [17]. The standard treats data integrity, training-set coverage, and post-deployment behavioral monitoring as first-class safety activities, and explicitly anticipates a bridge to ISO 26262 and ISO 21448 through cross-references in the AI Safety Lifecycle [9].

0.2.5 E. Run-Time Assurance and Behavior-Policy Formal Methods

Mehmed et al. survey runtime monitoring approaches for automated driving systems and classify them into ODD monitors, safety-envelope monitors, and trajectory-bound monitors [18]. Mauritz and colleagues describe a runtime monitoring concept that observes the perception stack and the planning stack jointly and invokes a minimum-risk maneuver when either crosses a predefined bound [19]. Cheng and co-authors propose a

runtime monitoring approach for traffic-light behavior policy with formal proof of bounded recovery time [20]. Mobileye's Responsibility-Sensitive Safety (RSS) provides a mathematical contract for safe longitudinal and lateral distances between road users, grounded in five formal rules [21]. IEEE P2846 takes a parallel path by codifying a minimum set of foreseeable scenarios and reasonable assumptions for ADS behavior models [22].

0.2.6 F. Adverse-Weather Perception and Sensor Degradation

The fusion of camera, radar, and LiDAR data is the dominant architectural pattern in production automotive ADAS [23]. Bijelic and colleagues demonstrate a fog-robust deep fusion network on the SeeingThroughFog corpus and quantify per-modality degradation under dense water particulates [24]. A 2024 IEEE Transactions on Intelligent Transportation Systems survey by Yurtsever et al. tabulates per-modality degradation curves across five weather classes [25]. Recent work on sun-glare detection demonstrates self-supervised convolutional architectures that produce a per-frame degradation score with measurable correlation to downstream object-detection recall [26]. Multi-modal water-hazard detection — including the WaterScenes 4D radar–camera fusion dataset — addresses the specific class of failure mode exhibited in the Waymo San Antonio recall [27]. Sensor-fusion uncertainty quantification via deep ensembles, Monte Carlo dropout, and evidential learning has been shown to estimate both epistemic and aleatoric uncertainty at frame rate, providing a basis for a runtime degradation monitor [28].

III. Approach: A Six-Step Framework

We propose a six-step framework that ingests perception-availability evidence collected in the candidate Operational Design Domain (ODD) and produces a UL 4600 claim-based gate, an ISO 21448 triggering-condition catalogue with derived SOTIF requirements, and an ISO 26262 minimal-risk-condition transition contract. Fig. 1 sketches the stage flow.

0.3.1 A. Step S1 — Perception-Availability Evidence Capture

The first step ingests three classes of evidence. The first is per-modality sensor-availability curves under the candidate ODD, derived from controlled trials and from field deployments. Following Bijelic et al. [24] and the survey of Yurtsever et al. [25], we treat camera availability $A_{\text{cam}}(c)$, radar availability $A_{\text{rad}}(c)$, and LiDAR availability $A_{\text{lid}}(c)$ as functions of the environmental condition vector c that includes visibility, illumination, precipitation rate, and water-on-surface depth. The second is ODD-coverage statements per UL 4600 Section 8 [7] and per ISO 34502 scenario-class enumeration [8]. The third

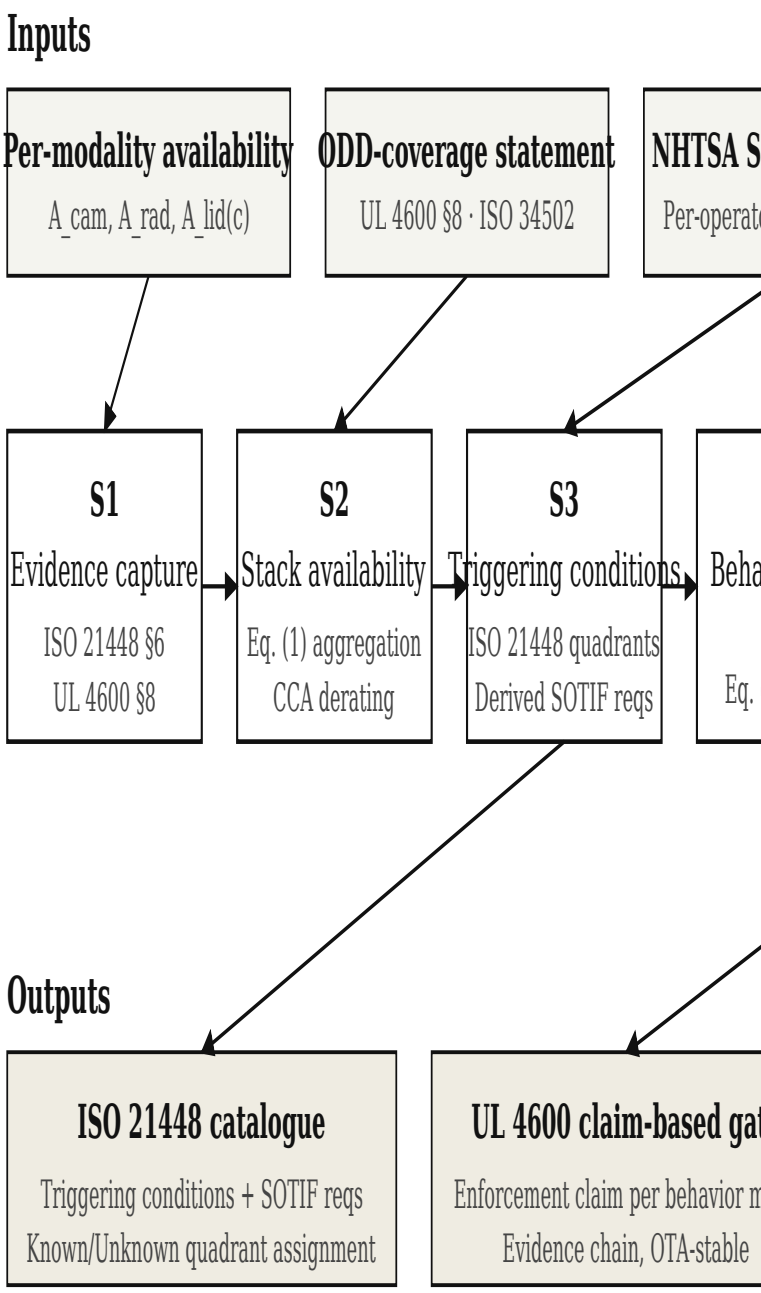

Fig. 1. Six-step cross-standards framework. Inputs (top) → s
Each stage is annotated with the gover

Figure 1. Cross-standards framework. Inputs (top) flow through six numbered stages (S1–S6) into UL 4600, ISO 21448, and ISO 26262-compatible outputs (bottom). Each stage is annotated with the governing clause.

is bounded inference from publicly available NHTSA Standing General Order on Crash Reporting data [10], which supplies per-population incident rates by ADS operator and class.

0.3.2 B. Step S2 — Stack-Level Availability Aggregation

The second step aggregates the per-modality availabilities into a perception-stack availability $A_s(c)$ under the assumption of partial independence with explicit common-cause derating. We write

$$A_s(c) = 1 - \prod_{i=1}^{N} [1 - A_i(c) \cdot I_i(c)] - \delta_{\text{cc}}(c) \quad (1)$$

where $I_i(c)$ is an indicator that modality i is within its declared ODD at condition c, and $\delta_{\text{cc}}(c)$ is a common-cause derating that captures correlated degradation across visible-band modalities under fog, sun glare, and standing water. The framework requires $\delta_{\text{cc}} \geq 0.02$ for scenes in which camera and LiDAR are both visible-band and the condition involves dense fog, sun on a wet road, or specular reflection from standing water [24], [25], [27]. We discuss the sensitivity of (1) to δ_{cc} in Section V.

0.3.3 C. Step S3 — Triggering-Condition Catalogue

The third step translates $A_s(c)$ deficits into a catalogue of ISO 21448 triggering conditions. For each region of the condition space c in which $A_s(c)$ drops below an acceptance threshold $A_{\min}$, we enumerate the scenarios that activate the corresponding functional insufficiency and assign them to one of the four SOTIF quadrants [5], [6]. The threshold $A_{\min}$ is calibrated against per-class incident-rate budgets derived from the NHTSA SGO data [10] and is set, for the worked example of Section IV, to $A_{\min} = 1 - 10^{-6}$ per hour for catastrophic-class triggering conditions. Each triggering condition is paired with one or more derived SOTIF requirements, which are then traced into the planning and behavior-policy stack as in Section III-D.

0.3.4 D. Step S4 — Behavior-Policy Gate Derivation

The fourth step is the principal contribution of this framework. We derive, for each triggering condition catalogued in Step S3, a *behavior-policy gate* — a Boolean predicate over the joint state of the vehicle, the perception stack, and the static-and-dynamic environment, whose conjunction is required for the planner to be permitted to enter the corresponding behavioral mode. The gate is the run-time formalization of the triggering-condition catalogue: it ensures that the vehicle cannot enter a region of state space outside its ODD even when the perception stack reports a degraded but non-zero output. For the flood-traversal case, the gate has the form

$$G_{\text{traverse}} \iff [A_s(c) \geq A_{\min}] \wedge \neg F_{\text{standing-water}} \wedge [v_{\text{cmd}} \leq v_{\text{road-class}}]$$

where $F_{\text{standing-water}}$ is a perception-derived flag for standing water at the planned path, v_{cmd} is the commanded speed, and $v_{\text{road-class}}$ is the HD-map road-class speed allowance [29]. Violation of G_{traverse} triggers a transition to an alternative behavior, typically *plan a route around* or *minimal-risk stop*.

0.3.5 E. Step S5 — UL 4600 Claim Closure

The fifth step closes the behavior-policy gates against the UL 4600 claim-based safety case. Each gate becomes an *enforcement claim*: "Claim C-i: The planner shall not enter behavioral mode M-i unless gate G-i evaluates true." Each enforcement claim is supported by an argument that maps to ISO 21448 triggering-condition coverage (the SOTIF lens), ISO/PAS 8800 AI output-insufficiency analysis (the AI lens), and ISO 26262 ASIL allocation for the gate-enforcement function (the functional-safety lens) [5], [7], [9]. The minimum allocation for an enforcement gate whose violation could lead to a Catastrophic outcome (per the ISO 21448 hazardous-behavior classification [5]) is ASIL D under ISO 26262 [30]. The claim-based pattern of UL 4600 allows the gate to be re-verified after each over-the-air software update without re-opening the entire safety case [7], [13].

0.3.6 F. Step S6 — Minimal-Risk-Condition Transition

The sixth step instantiates the SAE J3016 minimal-risk-condition (MRC) transition [4]. For each enforcement gate, we specify a transition contract: the conditions under which the planner shall request a transition to an MRC, the maximum permitted transition latency, and the post-transition supervision regime. For L4 vehicles without an in-vehicle safety driver, the MRC is typically a *stable, stopped* condition at a roadside location safe for evacuation; for L2 vehicles with a human supervisor, the MRC is a takeover request escalated to a controlled hand-back if the supervisor is unresponsive [4]. The transition contract is verified by simulation per ISO 34502 [8] and by closed-course testing per UL 4600 Section 12 [7]. Table I summarizes three representative ODD-monitor styles and their fit to the framework.

IV. Worked Example: Flood-and-Glare on a Level-4 Robotaxi

We apply the framework to a Level-4 robotaxi operating in a metropolitan ODD that includes urban arterials, residential collector streets, and limited-access highways.

Table 1. Representative ODD-monitor styles, their primary failure-mode coverage, and their fit to the proposed framework. Adapted from [18], [19], [20].

Monitor style	Primary coverage	Latency	Fit
Geometric envelope (RSS)	Lon/lat distance violations	Single frame	S4 gate, following
Perception-availability	Sensor degradation	Multi-frame	S2 input, cam-only gate
HD-map class & precondition	ODD-class violations	Sub-second	S4 gate, road-class & speed

The hypothetical functional thread is *traverse a planned route segment under degraded environmental conditions*. We treat two failure modes jointly: an *untraversable-lane class* (flash-flood standing water) and a *perception-degradation class* (sun glare at low-angle insolation on a wet pavement).

0.4.1 A. S1 — Evidence Capture

Following [24], [25], we treat camera availability under heavy rain and sun glare with visibility 200–500 m as $A_{\text{cam}} = 0.88$; LiDAR availability under the same conditions (with specular reflection from standing water adding return-clutter noise) as $A_{\text{lid}} = 0.90$; and radar availability as $A_{\text{rad}} = 0.98$. The bounded inference from NHTSA EA26002 [3] is that camera-only stacks lacking a degradation monitor exhibit an unannounced perception-loss rate that is non-negligible at the population scale; we therefore require an explicit degradation monitor as a precondition for accepting A_{cam} as above. The bounded inference from the Waymo recall [1], [2] is that even multi-modal stacks with degradation monitoring can produce a planner output that proceeds into the degraded region of state space if the behavior policy does not gate on the perceived condition.

0.4.2 B. S2 — Stack Availability

Camera and LiDAR are jointly degraded by sun-on-wet-pavement specular reflection, so we apply (1) with $\delta_{\text{cc}} = 0.04$ (a conservative 4 % common-cause derating to reflect the joint degradation regime [24], [25]). The resulting stack availability is $A_s(c) = 1-(1-0.88)(1-0.98)(1-0.90) - 0.04 = 1 - 0.00024 - 0.04 = 0.9598$, or approximately $1 - 4 \times 10^{-2}$ per second. When integrated over a 1-hour metropolitan operating period and conditioned on the joint-degradation regime obtaining only intermittently, this corresponds to a per-flight-hour perception-loss exposure on the order of 10^{-5}, which by Step S3 is below the catastrophic threshold $A_{\text{min}} = 1 - 10^{-6}$ and therefore activates the triggering-condition catalogue.

0.4.3 C. S3 — Triggering Conditions

The principal triggering conditions for the joint regime are:

- TC-01 *Untraversable lane / standing water at planned path* — perception detects standing water above 6 inches at the planned trajectory; functional insufficiency is the absence of a behavior-policy gate that vetoes path traversal.
- TC-02 *Sun-glare blinding of forward camera* — perception availability drops below A_{min} for a multi-frame window exceeding 200 ms with the forward camera as primary cue; functional insufficiency is the absence of a degradation monitor that escalates to an alternative cue or to a minimal-risk stop.
- TC-03 *HD-map road-class mismatch* — perception or localization reports a road class (e.g., flooded low-water crossing, off-pavement) that does not match the planner's commanded behavior class.

The three conditions correspond, respectively, to the public-record evidence in [1], [2], [3]. Each is assigned to the Known Unsafe quadrant of the SOTIF matrix [5], with derived SOTIF requirements summarized in Table II.

0.4.4 D. S4 — Behavior-Policy Gates

The gate G_{traverse} is instantiated as in (2). The gate G_{vision} is instantiated as the conjunction $[A_{\text{cam}}(c) \geq A_{\text{min}}] \vee$ [LiDAR or radar primary available] $\vee$ [request MRC]. The gate $G_{\text{road-class}}$ is instantiated as a Boolean predicate over the HD-map road-class attribute [29] and the planner's commanded behavior class; violation triggers a planned route replan or MRC.

0.4.5 E. S5 / S6 — Fault Tree and Closure

Fig. 2 sketches the fault tree for the top event *Vehicle enters untraversable region of state space, undetected*. The OR gate below the top event expands into the union G_{traverse} false-negative $\vee$ G_{vision} false-negative $\vee$ $G_{\text{road-class}}$ false-negative. Each branch is supported by an enforcement claim per UL 4600 Section 8 [7]. The MRC transition contract for the worked example specifies a maximum latency of 1.5 s from gate violation to commanded-decel onset, and a minimum decel of 4 m/s^2 to a stable stop on the right shoulder of the active lane [4], [21]. The transition contract is verified by simulation per ISO 34502 [8] using a critical-scenario library that includes the Waymo flood scenario [1] and the EA26002 glare scenarios [3] as ground-truth scenes.

Table 2. Derived SOTIF requirements per triggering condition, with allocated ASIL and behavior-policy gate name. Allocation per [5], [9], [30].

TC	Functional insufficiency	Derived SOTIF requirement	ASIL	Gate
TC-01	No water-traverse gate	Planner shall not command traversal of detected water	D	G_{traverse}
TC-02	No cam-degradation escal.	Escalate to MRC if cam-only avail. < A_min for 200 ms	D	G_{vision}
TC-03	No HD-map road-class pre.	Planner shall not enter road class outside ODD	D	$G_{\text{road-class}}$

V. Discussion

The framework is intended as a structured input to a Level-4 deployment safety case, not as a substitute for one. Four limitations and threats to validity merit explicit statement.

0.5.1 A. Sensitivity of (1) to the Common-Cause Term δ_{cc}

The common-cause derating δ_{cc} in (1) is the dominant uncertainty in the stack-availability estimate under joint visible-band degradation. In the worked example of Section IV, halving δ_{cc} from 0.04 to 0.02 moves $A_s(c)$ from 0.96 to 0.98, which moves the per-hour catastrophic exposure from approximately 10^{-5} to 10^{-6} — across the $A_{\min}$ boundary and into a different gate-allocation regime. The literature on multi-modal sensor-fusion uncertainty quantification [28] supplies an upper bound on δ_{cc} via deep-ensemble disagreement, but no closed-form lower bound is yet available. Future work should establish a structured method for δ_{cc} calibration from controlled-trial data, in alignment with ISO/PAS 8800 data-integrity expectations [9].

0.5.2 B. Behavior-Policy Gates Are Not a Substitute for ODD Coverage

The framework presumes that the triggering-condition catalogue of Step S3 is adequately complete. In practice, the SOTIF Known/Unknown migration activity is open-ended; new triggering conditions emerge from field deployment. The framework therefore requires periodic re-execution of Steps S1–S3 informed by post-deployment data — the Standing General Order data [10] in the United States, and the equivalent regulator filings in other jurisdictions — and re-derivation of behavior-policy gates as new triggering conditions are catalogued. The UL 4600 claim-based pattern accommodates this re-execution but does not eliminate the underlying combinatorial completeness problem [6], [11].

0.5.3 C. The Standards Lens Stops at the Behavior-Policy Boundary

The framework codifies how a triggering condition is translated into a behavior-policy gate, but it does not prescribe the behavior policy itself. The choice between *minimal-risk stop in lane*, *minimal-risk stop on shoulder*, *plan a route around*, and *request remote assistance* is an operations-engineering decision that depends on the ODD geometry, the surrounding traffic, the emergency-services availability, and the regulatory regime. UL 4600 Section 13 [7] supplies the claim-based pattern for capturing the rationale, but the analytical methods for choosing among these options are an open research problem [14], [22]. We note that the Waymo recall remedy [1], [2] explicitly combines two of these options — refined extreme-weather operations and area-restricted operations during flash-flood watches — which suggests that the choice is not binary.

0.5.4 D. The L2-vs-L4 Asymmetry

The Tesla Full Self-Driving (Supervised) case [3] is SAE Level-2 with an explicit human-supervisor fallback; the Waymo case [1], [2] is SAE Level-4 with no in-vehicle supervisor. The framework is presented in a Level-4 form, but is applicable at Level-2 with one modification: at Level-2 the MRC transition of Step S6 is replaced by an escalated takeover request with bounded latency, and the gate-enforcement claim of Step S5 includes an explicit driver-attention precondition per SAE J3016 [4] and the IEEE P2846 reasonable-assumption set [22]. The Cruise pedestrian-dragging incident of October 2023 [31], although outside the scope of the worked example, illustrates a third class of behavior-policy failure — where the post-incident behavior policy (pull-over-out-of-traffic) was selected without an enforcement gate that conditioned the policy on the post-incident object-pose state — and is amenable to the same framework with a different triggering-condition catalogue. The framework is therefore claimed to be applicable to L2 and to post-incident L4 behavior, but the demonstrations in this paper are restricted to the in-trip L4 case.

VI. Conclusion and Future Work

We have proposed a six-step cross-standards framework for translating perception-availability evidence into ISO 21448 triggering conditions, UL 4600 claim-based gates, and ISO 26262 minimal-risk-condition transition contracts for SAE Level-4 robotaxis operating under degraded environmental conditions. The framework treats the May 2026 Waymo San Antonio flood-traversal recall and the March 2026 NHTSA EA26002 escalation

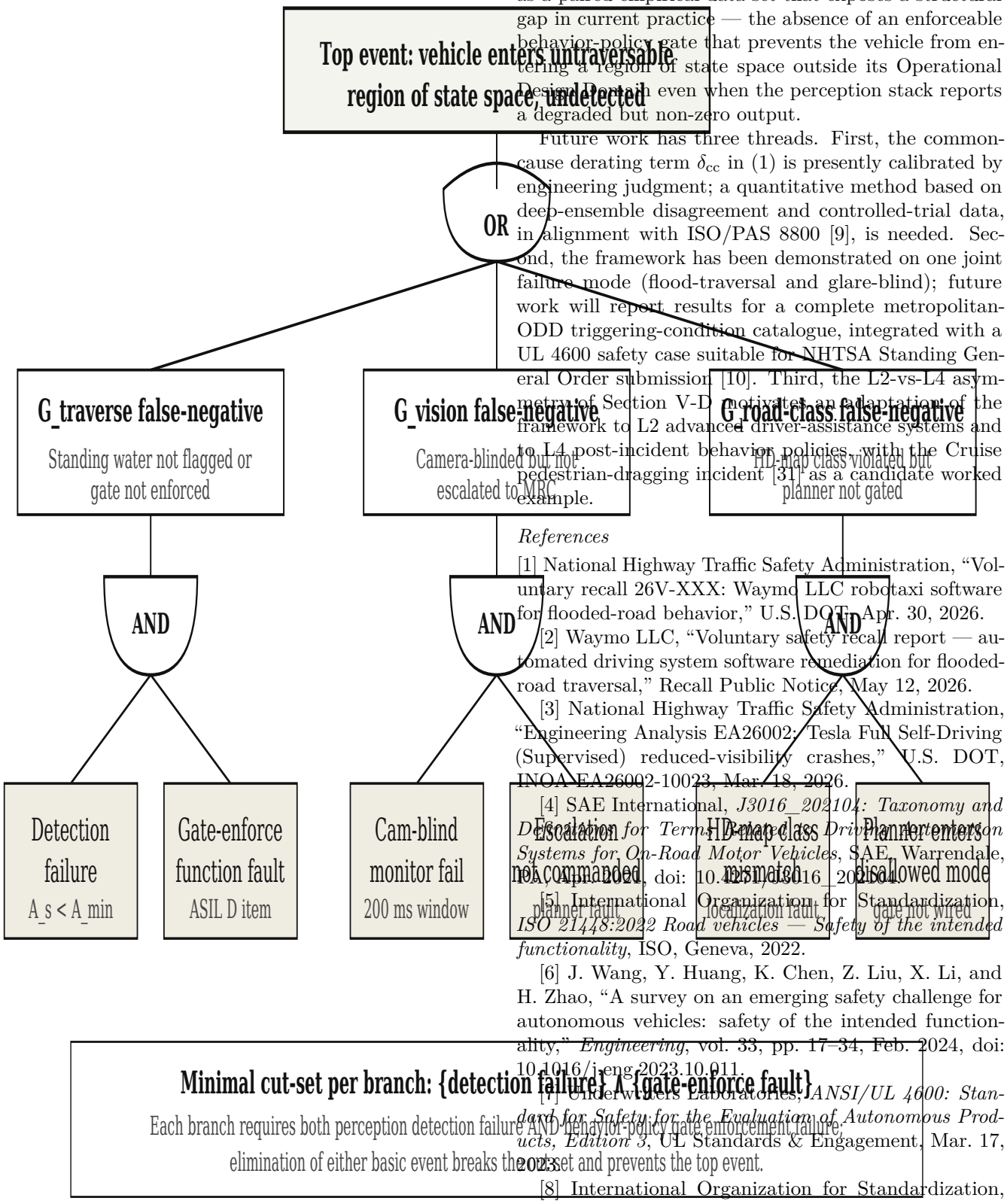

Fig. 2. Fault tree for "vehicle enters untraversable region of state space, undetected" (Section IV-E).

Figure 2. Fault tree for the top event "Vehicle enters untraversable region of state space, undetected." OR gates below the top event correspond to the three behavior-policy gates of Section III-D; AND-decompositions below each gate require both detection failure and gate-enforcement failure.

of Tesla Full Self-Driving in reduced-visibility crashes as a paired empirical data set that exposes a structural gap in current practice — the absence of an enforceable behavior-policy gate that prevents the vehicle from entering a region of state space outside its Operational Design Domain even when the perception stack reports a degraded but non-zero output.

Future work has three threads. First, the common-cause derating term δ_{cc} in (1) is presently calibrated by engineering judgment; a quantitative method based on deep-ensemble disagreement and controlled-trial data, in alignment with ISO/PAS 8800 [9], is needed. Second, the framework has been demonstrated on one joint failure mode (flood-traversal and glare-blind); future work will report results for a complete metropolitan-ODD triggering-condition catalogue, integrated with a UL 4600 safety case suitable for NHTSA Standing General Order submission [10]. Third, the L2-vs-L4 asymmetry of Section V-D motivates an adaptation of the framework to L2 advanced driver-assistance systems and to L4 post-incident behavior policies, with the Cruise pedestrian-dragging incident [31] as a candidate worked example.

References

[1] National Highway Traffic Safety Administration, "Voluntary recall 26V-XXX: Waymo LLC robotaxi software for flooded-road behavior," U.S. DOT, Apr. 30, 2026.

[2] Waymo LLC, "Voluntary safety recall report — automated driving system software remediation for flooded-road traversal," Recall Public Notice, May 12, 2026.

[3] National Highway Traffic Safety Administration, "Engineering Analysis EA26002: Tesla Full Self-Driving (Supervised) reduced-visibility crashes," U.S. DOT, INOA-EA26002-10023, Mar. 18, 2026.

[4] SAE International, *J3016_202104: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE, Warrendale, PA, Apr. 2021, doi: 10.4271/j3016_202104.

[5] International Organization for Standardization, *ISO 21448:2022 Road vehicles — Safety of the intended functionality*, ISO, Geneva, 2022.

[6] J. Wang, Y. Huang, K. Chen, Z. Liu, X. Li, and H. Zhao, "A survey on an emerging safety challenge for autonomous vehicles: safety of the intended functionality," *Engineering*, vol. 33, pp. 17–34, Feb. 2024, doi: 10.1016/j.eng.2023.10.011.

[7] Underwriters Laboratories, *ANSI/UL 4600: Standard for Safety for the Evaluation of Autonomous Products, Edition 3*, UL Standards & Engagement, Mar. 17, 2023.

[8] International Organization for Standardization, *ISO 34502:2022 Road vehicles — Test scenarios for automated driving systems — Scenario based safety evaluation framework*, ISO, Geneva, 2022.

[9] International Organization for Standardization,

ISO/PAS 8800:2024 Road vehicles — Safety and artificial intelligence, ISO, Geneva, Dec. 2024.

[10] National Highway Traffic Safety Administration, *Standing General Order 2021-01 (third amended), Incident Reporting for Automated Driving Systems and SAE Level 2 ADAS*, U.S. DOT, eff. Jun. 16, 2025.

[11] M. J. Khan, "Validation of safety of the intended functionality for autonomous and ADAS systems," in *Proc. 27th Int. Tech. Conf. Enhanced Safety of Vehicles (ESV)*, Yokohama, Japan, 2023, paper 27ESV-000009.

[12] Z. Liu, H. Cao, and X. Chen, "Decomposition and quantification of SOTIF requirements for perception systems of autonomous vehicles," arXiv:2501.10097, Jan. 2025.

[13] UL Solutions, "UL 4600 Edition 3 updates incorporate autonomous trucking," UL Solutions Tech. Brief, Mar. 17, 2023.

[14] P. Koopman, "An overview of draft UL 4600: Standard for safety for the evaluation of autonomous products," Edge Case Research Tech. Rep., 2020.

[15] PEGASUS Project Consortium, "PEGASUS method — an overview," BMWi Final Report, Berlin, May 2019.

[16] T. Kaibara, S. Onogi, and I. Hasuo, "Temporal logic formalisation of ISO 34502 critical scenarios: modular construction with the RSS safety distance," arXiv:2403.18764, Mar. 2024.

[17] Critical Systems Labs, "Adapting ISO/PAS 8800 to AI and ML system safety assurance within other industries, v1.0," Critical Systems Labs Tech. Rep., Sep. 2025.

[18] A. Mehmed, "Runtime monitoring for safe automated driving systems," Doctoral dissertation, Mälardalen University, 2020.

[19] M. Mauritz, F. Howar, and A. Rausch, "From SOTIF to runtime safety: an early concept evaluation of a runtime monitoring approach for safe automated driving," in *Proc. IEEE Int. Conf. Intell. Transp. Syst. (ITSC)*, 2022, pp. 1473–1480, doi: 10.1109/ITSC55140.2022.9922192.

[20] R. Cheng, J. Park, and S. Kim, "Runtime monitoring approach to safeguard behavior of autonomous vehicles at traffic lights," *Electronics*, vol. 14, no. 12, art. 2366, Jun. 2025, doi: 10.3390/electronics14122366.

[21] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," arXiv:1708.06374, Aug. 2017.

[22] IEEE, *IEEE P2846: A Formal Model for Safety Considerations in Automated Vehicle Decision Making*, IEEE Standards Association, 2022.

[23] J. Wang, S. Park, and L. Chen, "A review of multisensor fusion in autonomous driving," *Sensors*, vol. 25, no. 19, art. 6033, Sep. 2025, doi: 10.3390/s25196033.

[24] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: deep multimodal sensor fusion in unseen adverse weather," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11679–11689, doi: 10.1109/CVPR42600.2020.01170.

[25] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, Mar. 2020, doi: 10.1109/ACCESS.2020.2983149.

[26] J. Andrade, R. Pereira, and A. Almeida, "Self-supervised sun glare detection CNN for self-aware autonomous driving," in *Proc. NeurIPS ML4AD Workshop*, Dec. 2021.

[27] S. Yao, Z. Pi, S. Liu, R. Yang, R. Zhou, and J. Yang, "WaterScenes: a multi-task 4D radar-camera fusion dataset and benchmarks for autonomous driving on water surfaces," arXiv:2307.06505, 2023.

[28] K. Wang, Y. Jiang, and L. Zhang, "Uncertainty-aware prediction and application in planning for autonomous driving: definitions, methods, and comparison," arXiv:2403.02297, Mar. 2024.

[29] K. Park, J. Lee, and Y. Cho, "High-definition map representation techniques for automated vehicles," *Electronics*, vol. 11, no. 20, art. 3374, Oct. 2022, doi: 10.3390/electronics11203374.

[30] International Organization for Standardization, *ISO 26262:2018 Road vehicles — Functional safety*, ISO, Geneva, 2018.

[31] National Highway Traffic Safety Administration, "Consent Order — Cruise LLC, crash reporting deficiencies, October 2023 pedestrian incident," NHTSA Tech. Memo., 2024.