Waymo’s Safety Case and Use of Reference Models

Presented to the 42nd Session of the Informal Working Group on Functional Requirements for Automated Vehicles, Mountain View, CA, July 2023

Scott Schnelle
Kristofer Kusano
Johan Engström
Safety Research & Best Practices

WAYMO

Introducing Waymo's Approach to Safety: a Safety Case

A "safety case" is a structured argument, supported by a body of evidence that provides a compelling, comprehensible, and valid case that a system is or will be adequately safe for a given application in a given environment.

[UK Ministry of Defense DS 00-56, adapted subsequently in UL 4600]

The determination of safety is, at its heart, a **risk assessment** process

Safety Case Structure

GOAL
(Overarching statement)

The top-level goal of Absence of Unreasonable Risk. **Safety** is defined in ISO as **Absence of Unreasonable Risk (AUR)**

LOGICAL ARGUMENT
(Decomposing the statement)

Building a Credible Case for Safety: Waymo's Approach for the Determination of Absence of Unreasonable Risk (March 2023)

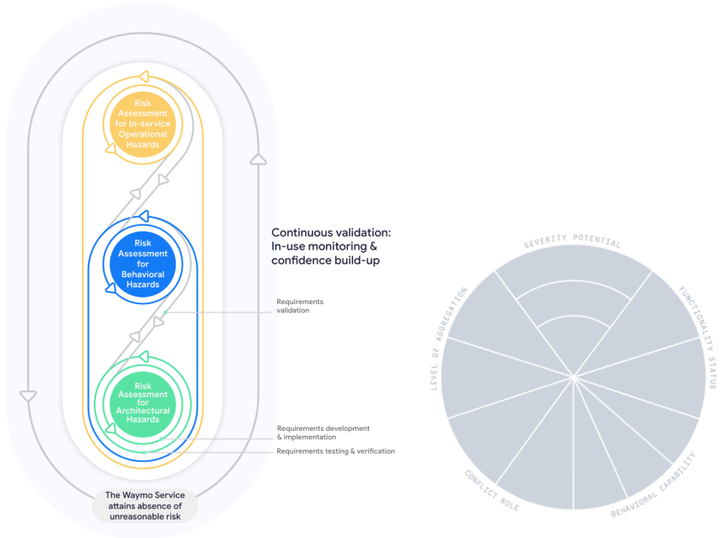
EVIDENCE
(Compelling proof)

Waymo's Safety Methodologies and Safety Readiness Determinations (October 2020)

Introducing Waymo's Approach to a Safety Case

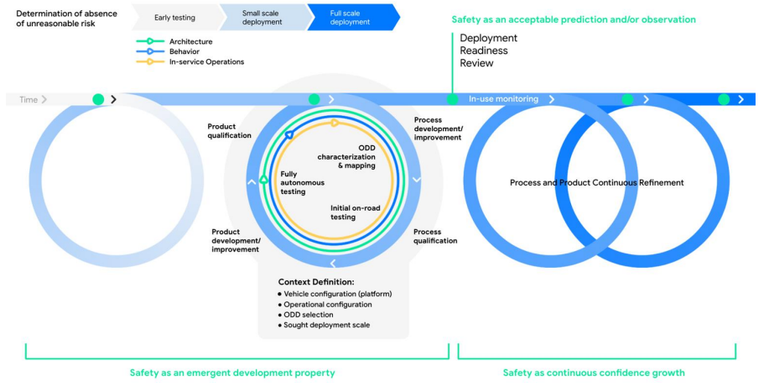
1

A layered approach to safety



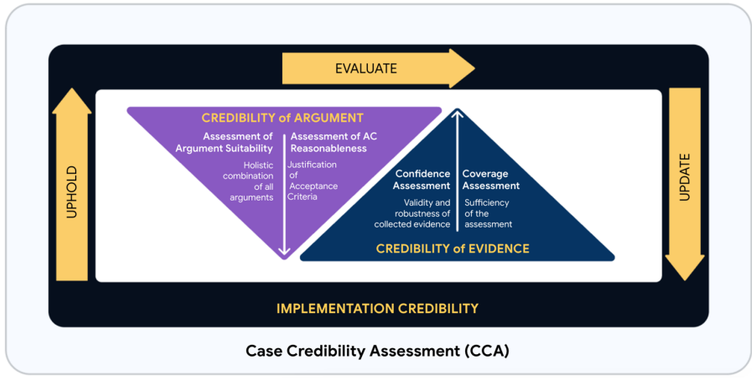
2

A dynamic approach to safety



3

A credible approach to safety



Case Credibility Assessment (CCA)

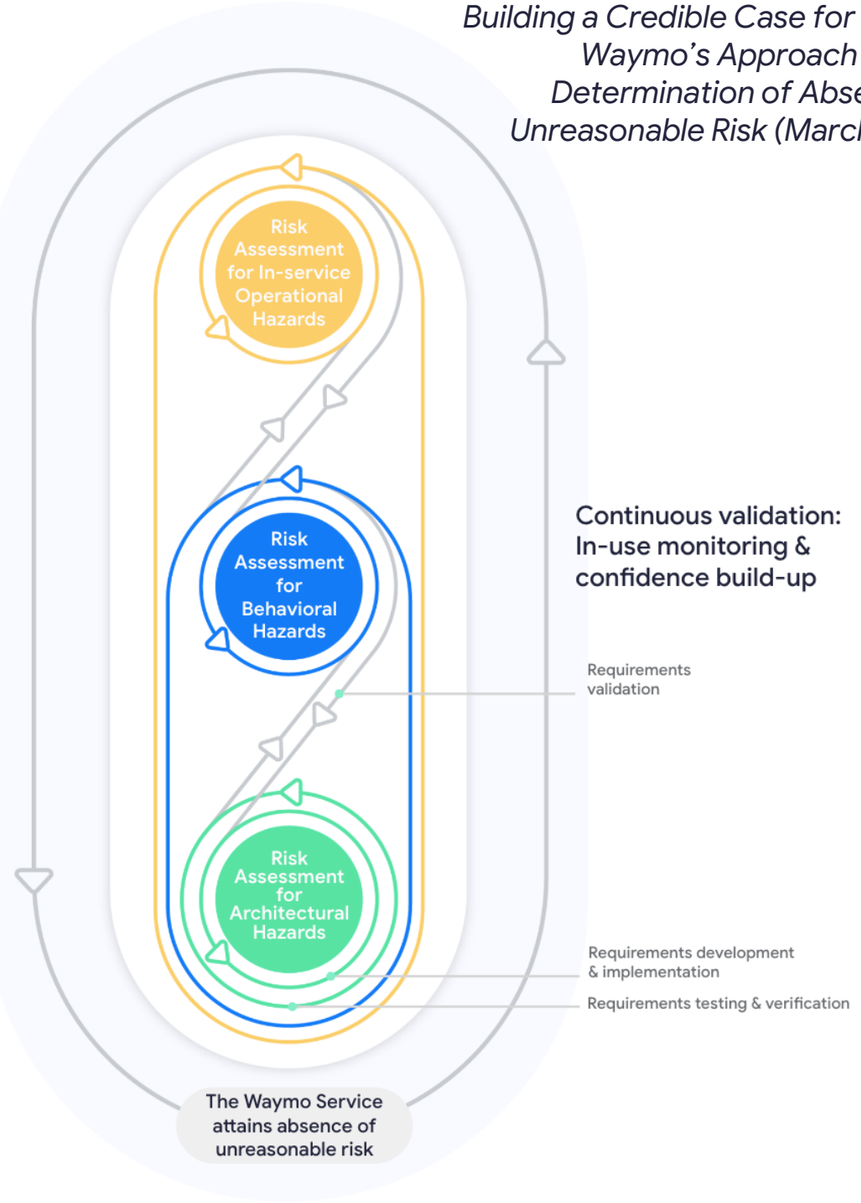
WAYMO

Introducing Waymo's Approach to a Safety Case

A layered approach to safety

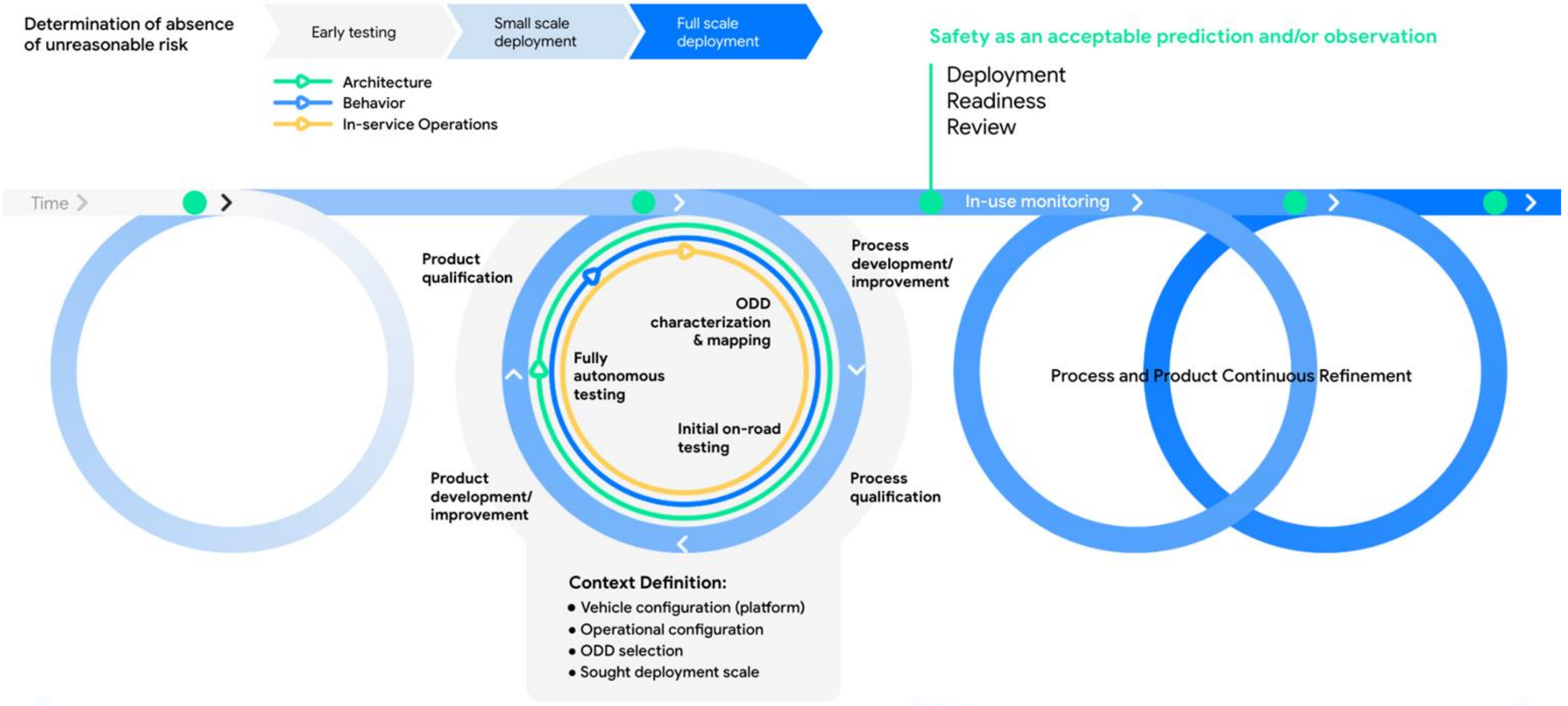
Waymo's Safety Methodologies and Safety Readiness Determinations (October 2020)

Building a Credible Case for Safety: Waymo's Approach for the Determination of Absence of Unreasonable Risk (March 2023)



WAYMO

A Dynamic Approach to Safety

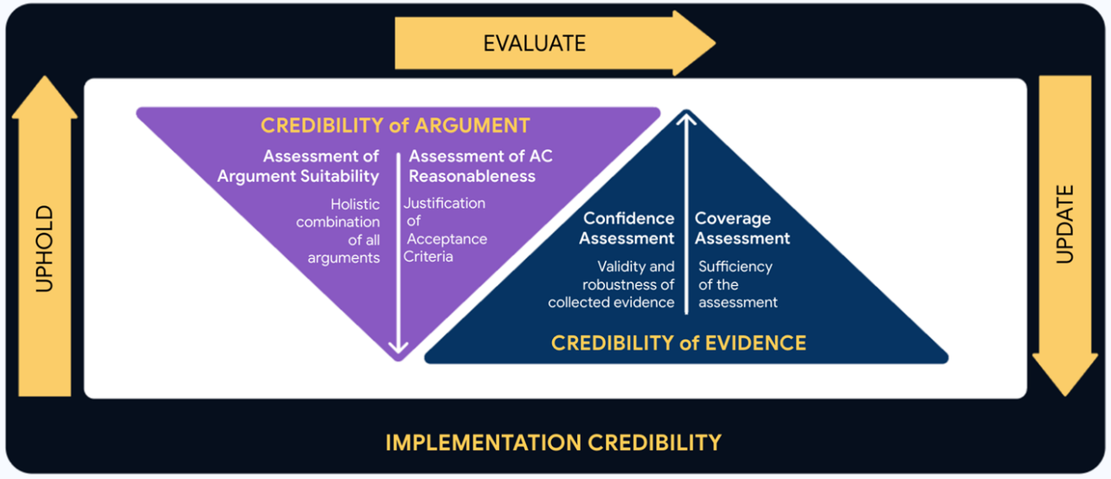


WAYMO

A Credible Approach to Safety

A “safety case” is a structured argument, supported by a body of evidence that provides a compelling, comprehensible, and valid case that a system is or will be adequately safe for a given application in a given environment.

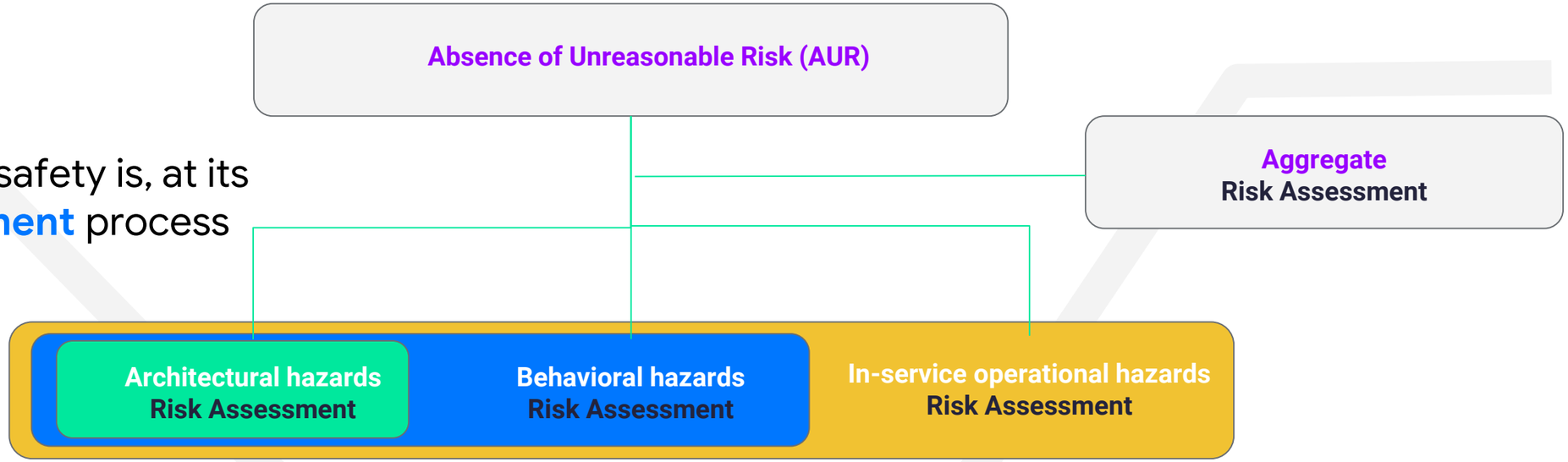
- Goal: Overarching statement
- Logical argument:
  Decomposing the statement
- Evidence:
  Compelling proof



Case Credibility Assessment (CCA)

A Layered Approach to Safety: Decomposing AUR

The determination of safety is, at its heart, a **risk assessment** process



Architectural hazards: those associated with potential sources of harm inherently embedded within the platform because of architectural choices. Example: undesired presence of blind-spots, stemming from architectural choices related to sensors' typology and placement.

Behavioral hazards: those associated with potential sources of harm resulting from the ADS's displayed driving behavior, whether intended or unintended. Example: undesired degree of proximity to surrounding road users.

In-service operational hazards: those associated with potential sources of harm resulting from the fact that the ADS operates in a complex ecosystem, and that do not belong to the other two categories. Example: improper securing of cargo or undesired access to the vehicle from a malicious actor.

AUR Determination and Risk Assessment

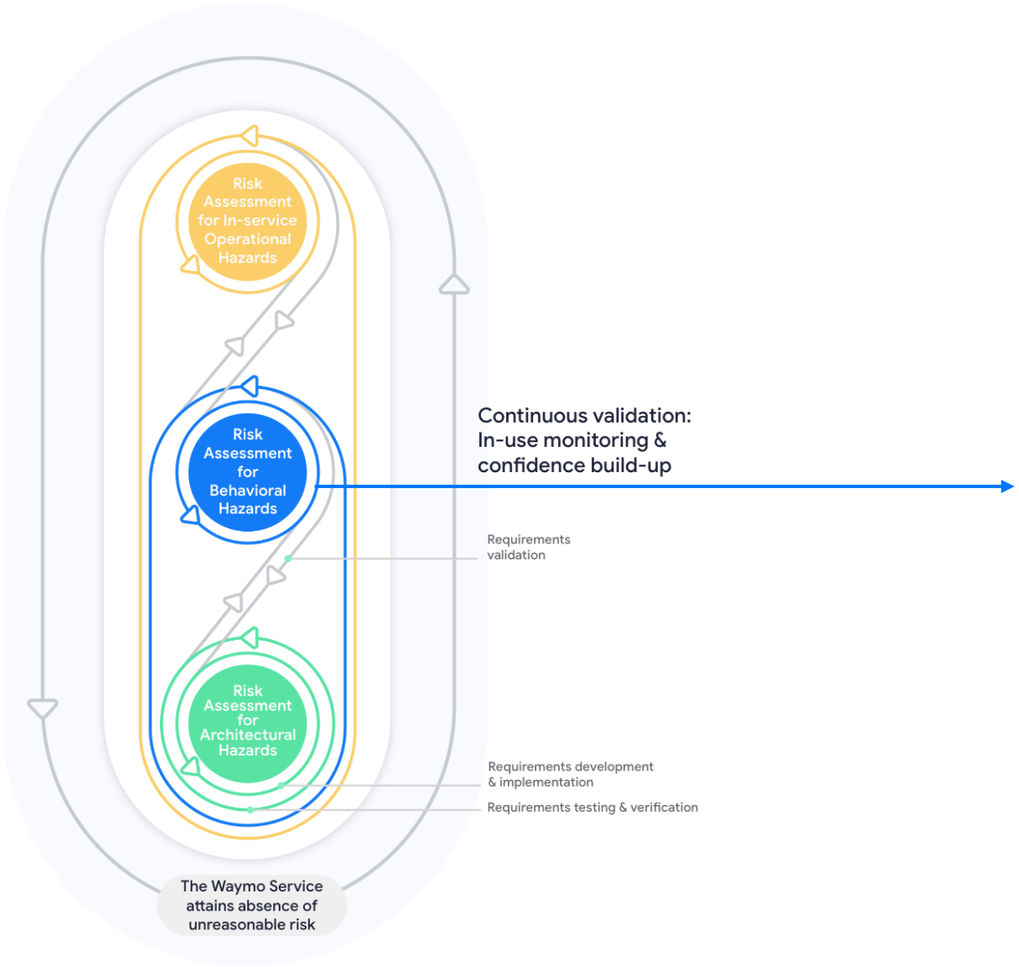
- For all hazard categories, a set of **explicit risk acceptance criteria** should be stated to assess if the residual risk reached an acceptable level or further mitigations are required.
- The crux thus remains of how to determine that a certain collection of acceptance criteria adequately covers a certain category of hazards.
- The process of setting appropriate acceptance criteria relies on the following three assumptions:

A. A sufficiently exhaustive list of hazards can be identified and covered by the categories “architectural”, “behavioral”, and “in-service operational”;

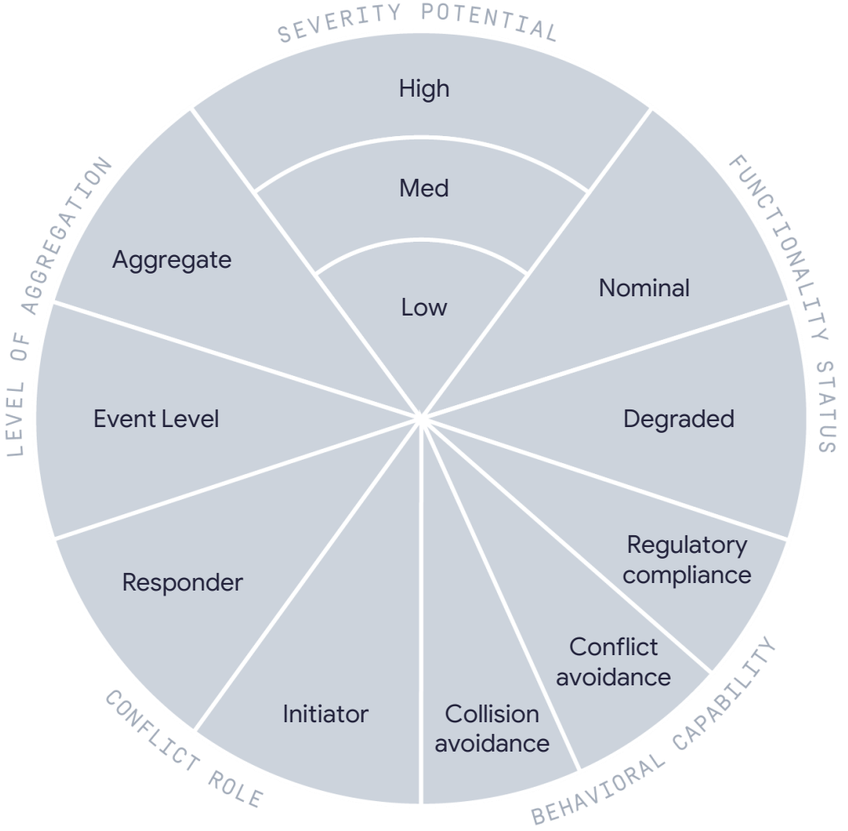
B. We can define indicators of interest mapped to each hazard category to set an explicit acceptance criterion for risk evaluation;

C. We can define the minimum set of dimensions of interest to state completeness of the set of acceptance criteria and establish credibility

B. We can define indicators of interest mapped to each hazard type to set an explicit acceptance criterion for risk evaluation;



C. We can define the minimum set of dimensions of interest to state completeness of the set of acceptance criteria and establish credibility



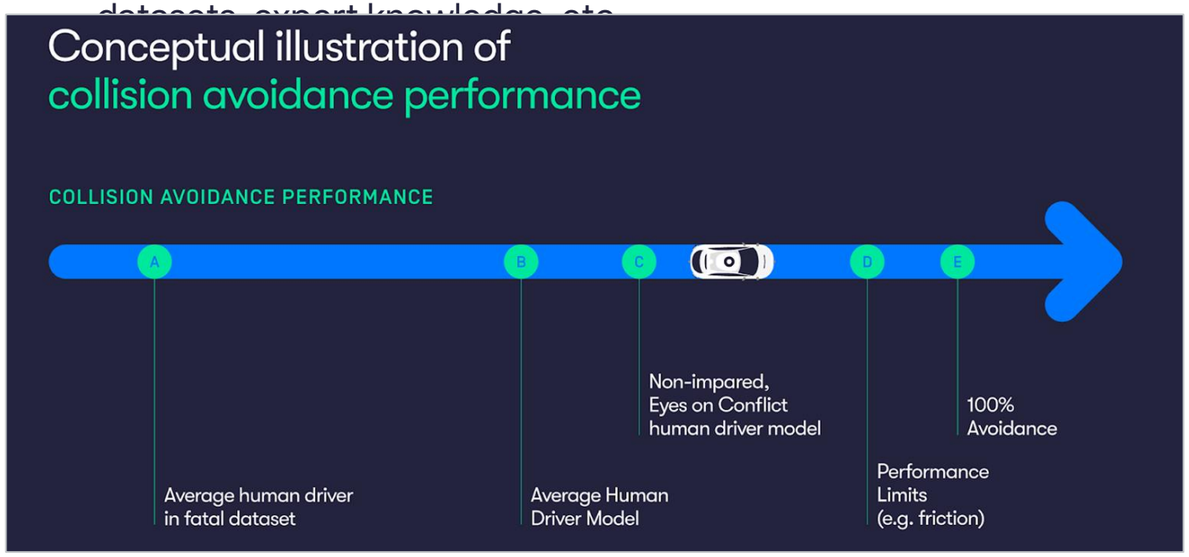
Acceptance Criteria Framework for AUR Behavioral Evaluation

WAYMO

Example of Targeted Scenario-Based Behavior Evaluation: Collision Avoidance Testing (CAT)

Kusano, K., Beatty, K., Schnelle, S., Favaro, F., Crary, C., and Victor, T. 2022. Collision Avoidance Testing of the Waymo Automated Driving System. arXiv:2212.08148

- Virtual, scenario-based testing methodology that evaluates the safety of the ADS's intended function
- Compared to the Non-Impaired driver with Eyes ON conflict (NIEON) model, i.e. high performing human driver
- ADS's ability to avoid situations initiated by others that require urgent evasive maneuvers
- Representative of a given ODD, informed by our driving experiences, public crash datasets, expert knowledge, etc.



WAYMO

Claim Structure: Application of the CCA through Argument construct

combination of all

Claim: AC [insert methodology specific AC] provides an explicit criterion to evaluate predicted RO performance appropriately mapped to dimensions [insert methodology specific AC framework dimensions] for the given context.

Subclaim (SC) #1: The stated acceptance criterion is reasonable

Subclaim (SC) #2: Methodology [insert methodology name] provides credible evidence that the stated acceptance criterion is met

WAYMO

Example: Collision Avoidance Testing

CAT white paper

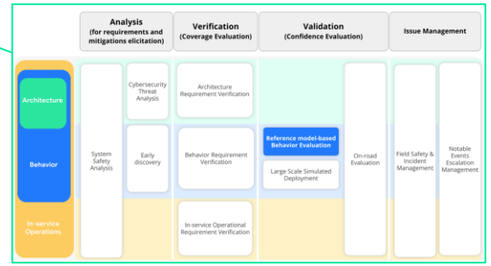
Acceptance Criterion #1 (AC1): The predicted RO collision avoidance capability attained by the Waymo Driver in a number of conflict scenarios initiated by the actions of other road users is assessed through a comparison with a non-impaired, eyes on conflict behavioral reference model made progressively more stringent by decreasing its emergency maneuver response time. Scenario groups are graded at an aggregate level, with individual scenarios within a group contributing to a neutral/positive/negative gap for the ADV when the Waymo Driver shows even/better/worse performance than the artificial driving model in terms of collision outcomes and injury-causing collisions. Minimum passing scores vary between scenario-specific groups, each including either vehicle to vehicle or vehicle to vulnerable road users interactions.

Context:

- Use-case: Ride-hailing urban/sub-urban
- Scale of deployment: e.g., fleet size, expected mileage
- Scope (ODD features and ADV behaviors)
  [LINK] ODD feature in-depth description or similar
- Platform: i-Pace, Pacifica
- Release: x.x.x

Claim 1: AC1 provides an explicit acceptance criterion to evaluate predicted RO *aggregate* ADV performance related to *responder-role collision avoidance capability in nominal (i.e., non degraded)* conditions for the given context.

1. **Subclaim #1:** AC1 is a reasonable criterion.
    1. The acceptance criterion is specified at the appropriate level of aggregation.
    2. The scoring assigned to the [VRU/V2V] type is adequate for determining the ADV's performance relative to the reference model.
    3. The NIEON artificial driving model is an appropriate and sufficient benchmark for evaluating responder role collision avoidance.
    4. The AC supports data-driven release qualification and identification of onboard engineering work to continuously improve the Waymo Driver.
    5. The AC is predicated upon appropriate performance indicators.
2. **Subclaim #2:** The *CAT methodology* provides credible evidence that AC1 is met.
    1. Coverage Assessment: The CAT methodology leverages a set of scenario groupings that represent adequate coverage of hazardous situations to develop a safety set that can be assessed for responder role collision avoidance capability for the Waymo ADV in nominal (i.e., non degraded) conditions for the given context.
    2. Confidence Assessment: The CAT methodology attains the appropriate confidence in the collision avoidance performance for the Waymo Driver in responder role predicted for RO operations, and its comparison relative to the chosen behavioral reference model for the given context.
        1. Scoring confidence [...]
        2. Conservativeness [...]
        3. Fidelity [...]
        4. Robustness [...]
        5. Appropriate use of qualified tools [...]
        6. Technical validity of benchmark [...]

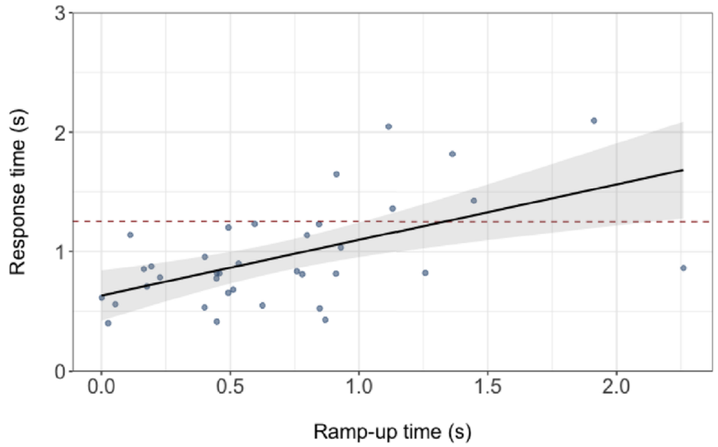


WAYMO

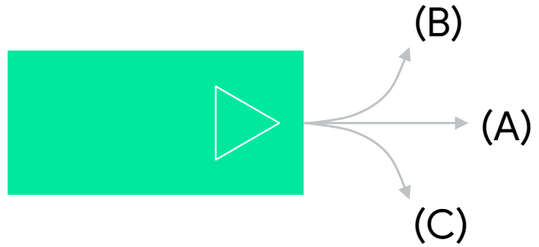
The Non-impaired Eyes ON (NIEON) Conflict Driver

Collision Avoidance Model Only that responds after a conflict has been entered.

1) Attentive with eyes always on the conflict



2) Model fit response response time using eyes-on-road, non-impaired naturalistic driving data*



3) Three chances given (best outcome selected):

(A)Brake only
(B)Brake + steer left
(C)Brake + steer right

Engström, J., Liu S-Y, Dinparastdjadid, A. and Simoiu, C. 2022. Modeling Road User Response Timing in Naturalistic Traffic Conflicts: A surprise-based framework. arxiv.org/abs/2208.08651

WAYMO

Research Needs for Driving Reference Models

Generic vs. Scenario-specific Models

Collision vs Conflict Avoidance Models

Complexity of Models

Validation Methods and Criteria

Closing /
Thank you

Scott Schnelle
Kristofer Kusano
Johan Engström
Safety Research & Best Practices

July 2023

WAYMO