From Dilemma to Metric:

A Publishable Crash Policy for Autonomous Vehicle Fleets

Adhyaay Karnwal

September 2026

Working paper

Abstract

The "crash dilemma" for autonomous vehicles (AVs) is usually posed as a trolley problem: when a collision is unavoidable, whom should the vehicle harm? We argue that this formulation is unanswerable by construction and that answering it, in any direction, is commercially and legally self-defeating. We propose a reframing on three levels. First, the vehicle's decision problem is not a single moral choice under certainty but a *policy* executed under perception uncertainty across billions of miles; the ethical object is the policy, and a policy must be stable, predictable, publishable and auditable. Second, the dominant lever on outcomes is not the choice made at the moment of unavoidability but whether that moment is reached at all; we define a *dilemma state* as a planning cycle in which no feasible trajectory has collision probability below a threshold under a published behavior envelope for other agents, and we show that its incidence is computable on every mile driven without a crash. Third, we specify a six-rule lexical policy (the *Crash Doctrine*): avoid the dilemma state; when in it, minimize total expected injury; weight every human equally with no personal attributes; never impose serious-injury risk on agents outside the conflict; prefer the predictable maneuver under model uncertainty; and log, publish and stand behind the result. We derive the *Dilemma Incidence Rate* (DIR) as the fleet-level leading indicator, show why equal weighting of occupants and non-occupants is acceptable to riders once injury asymmetry is accounted for, describe implementation on both modular and end-to-end driving stacks, and propose a governance mechanism (published doctrine, liability safe harbor, no-fault compensation fund) that turns disclosure from a litigation liability into a competitive asset. The contribution is not a resolution of the trolley problem but its replacement by a measurable, improvable engineering quantity.

Keywords: autonomous vehicles, machine ethics, trolley problem, responsibility-sensitive safety, runtime assurance, safety metrics, liability, no-fault compensation.

1 Introduction

Public and regulatory discussion of AV ethics has been dominated for a decade by a single scenario: a vehicle that must choose between harming its occupants and harming pedestrians, or between two groups of pedestrians. The MIT Moral Machine gathered roughly 40 million such judgments from participants in 233 countries [1]. Bonnefon, Shariff and Rahwan showed that respondents approve of vehicles that minimize casualties while declining to ride in one that might sacrifice them [2]. Germany's Ethics Commission on Automated and Connected Driving produced twenty rules, including a prohibition on weighting persons by personal characteristics, but left the occupant-versus-third-party question explicitly open [3]. Industry has responded mostly with silence, punctuated by

episodes such as a 2016 statement by a Mercedes-Benz safety executive that its vehicles would prioritize occupants, withdrawn within days [4].

This silence is locally rational. Any explicit choice is discoverable in litigation and a target for juries. But the aggregate result is that the policy governing the most consequential decisions a fleet makes is written after the fact, one case at a time, by the least appropriate forum. Meanwhile, roughly 40,000 people die on United States roads each year and about 1.19 million globally [5, 6]; Waymo's rider-only fleet, across more than 100 million miles, reports serious-injury crash rates reduced on the order of 90% relative to human benchmarks [7]. If those figures hold at scale, hesitation in deployment is itself the largest trolley problem in the field.

This paper argues that the dilemma dissolves under three moves, each standard in a neighboring discipline but not previously combined into a deployable policy:

1. Treat the vehicle's behavior as a **policy under uncertainty**, not a choice under certainty (Section 3).
2. Treat entry into the unavoidable-collision state as a **measurable event** rather than a thought experiment, and make its frequency the primary published safety metric (Section 5).
3. Treat disclosure as a **governance instrument** exchanged for liability protection, on the model of prior technologies that were safer in aggregate but litigated out of existence in particulars (Section 8).

Section 2 reviews related work. Section 4 states the policy. Section 6 analyzes the occupant-weighting question quantitatively. Section 7 covers implementation on modular and end-to-end stacks. Section 9 lists limitations and Section 10 concludes with concrete recommendations.

2 Related Work

Empirical moral psychology. The Moral Machine [1] and its successors establish that public preferences are broadly utilitarian in the abstract, vary by culture in the margins, and reverse when respondents imagine themselves as occupants [2]. We take these findings as constraints on *publishability*, not as inputs to the policy: a rule that cannot be stated in public in every market in which the fleet operates is not a rule a fleet can run.

Normative guidance. The German Ethics Commission [3] supplies the two constraints we adopt directly: no distinction by personal characteristics (its Rule 9) and the principle that parties who generate mobility risk may not sacrifice uninvolved parties (Rule 7). Goodall [8] and Lin [9] argue that AVs cannot avoid encoding some crash behavior and that the encoding should be deliberate. Nyholm and Smids [10] make the observation we build on: AV crash decisions are decisions about risk distribution made in advance by many stakeholders, not split-second individual choices.

Formal safety. Shalev-Shwartz, Shammah and Shashua's Responsibility-Sensitive Safety (RSS) [11] defines longitudinal and lateral safe distances such that, if the AV respects them and other agents behave within stated bounds, the AV cannot be the cause of a collision. Our dilemma-state definition generalizes the RSS "dangerous situation" to arbitrary trajectory sets and probabilistic perception, and DIR measures how often the RSS guarantee is lost in practice.

Injury biomechanics. Pedestrian fatality risk as a function of impact speed is well characterized [12, 13]. Occupant injury risk as a function of delta-V and restraint use is likewise established from crash databases [14]. These curves make severity-weighted harm minimization computable rather

than rhetorical.

Runtime assurance. Simplex-style architectures [15] and shielding for learned controllers [16] provide the pattern by which a rule can be enforced on an end-to-end policy without editing its weights: a verifiable monitor overrides the learned output when a safety property is about to be violated.

Governance precedents. The National Vaccine Injury Compensation Program [17] and Volvo's 2015 unilateral acceptance of liability for autonomous-mode crashes [18] are the precedents on which Section 8 is modeled.

3 Problem Formulation

3.1 Why the trolley framing fails

The trolley formulation assumes (a) certainty about the state of the world, (b) knowledge of the identity and characteristics of those at risk, and (c) a one-off choice. None holds. Perception delivers probability distributions over object class, position and velocity; reliable classification is at the level of "pedestrian" or "cyclist," not age or intent; and any decision rule is authored once and executed across the fleet's entire operational life. Ethical evaluation of a policy differs from that of an act: a policy must additionally be *predictable* to those it affects, *stable* across superficially similar situations, *publishable* without contradiction across jurisdictions, and *auditable* after the fact. These are engineering requirements, and they eliminate most candidate answers to the trolley question before any moral argument is made.

3.2 State, trajectories and the behavior envelope

At planning cycle t, the vehicle holds a belief state b_t over the positions, velocities and classes of surrounding agents $\mathcal{A}_t$. The planner considers a finite candidate set of trajectories $\mathcal{T}_t$ over a horizon H. Let $\mathcal{E}$ be a **behavior envelope**: a published set of bounds on how other agents may behave over H (maximum longitudinal and lateral acceleration by class, maximum reaction delay, maximum drift from lane or path). For each candidate trajectory $\tau \in \mathcal{T}_t$ and agent $i \in \mathcal{A}_t$, define the collision probability

$$p_i(\tau) \;=\; \Pr[\,\text{AV on } \tau \text{ collides with agent } i \mid b_t,\ \text{agent } i \text{ behaves within } \mathcal{E}\,]. \tag{1}$$

Definition 1 (Feasible collision-free set)**.** For a published threshold $p^* \in (0, 1)$,

$$\mathcal{F}_t \;=\; \left\{ \tau \in \mathcal{T}_t \;:\; \max_{i \in \mathcal{A}_t} p_i(\tau) < p^* \ \text{ and } \tau \text{ is dynamically feasible} \right\}. \tag{2}$$

3.3 The dilemma state

Definition 2 (Dilemma state and entry)**.** The vehicle is in a *dilemma state* at cycle t if and only if $\mathcal{F}_t = \varnothing$. A *dilemma-state entry* occurs at t if $\mathcal{F}_{t-1} \neq \varnothing$ and $\mathcal{F}_t = \varnothing$.

Each entry is attributed to one of two causes:

- **self**: the entry would not have occurred had the AV's own trajectory over the preceding k cycles been chosen from a stricter set; equivalently, the RSS-style guarantee was lost by the AV's action;
- **other**: an agent's realized behavior fell outside $\mathcal{E}$ and closed the set.

This attribution is computable offline from logs by counterfactual re-planning, and does not require a collision to have occurred.

3.4 Harm

For each agent i (including each occupant of the AV) and trajectory τ, let $S_i(\tau)$ be the expected injury severity conditional on collision, derived from published biomechanical curves as a function of predicted closing speed, impact geometry, agent class and, for occupants, restraint state. Define the expected harm of τ as

$$H(\tau) \;=\; \sum_{i \in \mathcal{A}_t \,\cup\, \text{occupants}} p_i(\tau)\, S_i(\tau). \tag{3}$$

Severity is measured on a common published scale, such as the probability of MAIS 3+ injury or the probability of fatality.

4 The Crash Doctrine

The policy is a lexical ordering of six rules. A lower rule is consulted only within the set of actions permitted by all higher rules; the numbering encodes precedence.

Rule 1 (Avoidance)**.** Choose actions that keep $\mathcal{F}_t$ non-empty. Speed, following gap and lateral position are constrained so that a collision-free trajectory under $\mathcal{E}$ exists at every cycle; under occlusion, the planner treats plausible hidden agents as present and slows accordingly. Every dilemma-state entry is logged with attribution.

Rule 2 (Harm minimization)**.** If $\mathcal{F}_t = \varnothing$, choose

$$\tau^* \;=\; \arg\min_{\tau \in \mathcal{T}_t} H(\tau) \quad \text{subject to Rules 3–5.} \tag{4}$$

Harm is what matters, not a count of bodies: a rule that scores "one versus five" without severities is worse than a rule that scores injury.

Rule 3 (Equal persons)**.** p_i and S_i may depend only on agent class, kinematics, and physical size insofar as it enters injury mechanics. They may not depend on age, sex, race, apparent socioeconomic status, number of co-passengers, customer status, or any re-identification. Occupants of the AV receive the same weight as any other human.

Remark. Rule 3 is the most easily audited: it is a constraint on the *interface* between perception and the harm model, verifiable by inspection independent of the perception model's internals.

Rule 4 (No new victims)**.** Let the *conflict set* $\mathcal{C}_t = \{\, i : p_i(\tau) \geq p^* \text{ for some } \tau \in \mathcal{T}_t \,\}$ at the cycle of entry. Agents $j \notin \mathcal{C}_t$ are protected by a hard constraint: no chosen τ may raise their probability of serious injury above a small ε,

$$p_j(\tau)\ \Pr[\text{serious injury} \mid \text{collision}, j, \tau] \;\leq\; \varepsilon \qquad \forall\, j \notin \mathcal{C}_t. \tag{5}$$

The AV may not redirect harm onto a person on a sidewalk, a vehicle stopped at a signal, or a cyclist in a protected lane to spare parties in the conflict. If the constraint leaves no trajectory, it relaxes to Rule 2 over all agents.

Remark. Rule 4 encodes the doing/allowing distinction recognized in both ethics and tort law, and it most sharply separates a defensible policy from a naive body count.

Rule 5 (Predictability)**.** Let τ_{brake} denote maximum in-lane braking and let σ_H be the harm model's own uncertainty. If $H(\tau_{\text{brake}}) - H(\tau^*) < \sigma_H$, choose τ_{brake}. Straight-line braking has one degree of freedom, the most certain outcome, and is the maneuver other road users anticipate. Predictability is a safety property in its own right.

Rule 6 (Record, publish, stand behind)**.** For every collision and every dilemma-state entry, retain the preceding 10 seconds of belief state, the full candidate set $\mathcal{T}_t$, each candidate's p_i, S_i and H, the chosen trajectory, and the rule that determined the choice. The doctrine text, $\mathcal{E}$, p^*, the severity scale and the metric of Section 5 are public.

5 The Metric: Dilemma Incidence Rate

Definition 3 (Dilemma Incidence Rate)**.** Over a reporting period with M miles driven and N, N_{self}, N_{other} dilemma-state entries in total and by attribution,

$$\text{DIR} = \frac{N}{M} \times 10^6, \qquad \text{DIR}_{\text{self}} = \frac{N_{\text{self}}}{M} \times 10^6, \qquad \text{DIR}_{\text{other}} = \frac{N_{\text{other}}}{M} \times 10^6. \tag{6}$$

Each is reported together with p^*, the envelope $\mathcal{E}$, the severity scale, and the fraction of entries that resulted in any physical contact.

Properties.

1. *Computable today.* Production AV planners already evaluate candidate trajectories against predicted agent motion each cycle. The dilemma-state test is a predicate over data they already compute. At current fleet scale, statistically useful counts accrue in days.
2. *Leading, not lagging.* Existing AV safety metrics (crashes, injury crashes, disengagements per million miles) require harm to occur. DIR counts the near-misses that harm metrics sample only rarely, giving orders of magnitude more statistical power and a direct engineering target.
3. *Fault-honest.* A pedestrian sprinting from behind a bus into a 30 mph lane creates a dilemma no policy can prevent; it falls in $\text{DIR}_{\text{other}}$. A vehicle that took a blind corner too fast falls in DIR_{self}. DIR_{self} is the number under the operator's control and the one to drive toward zero.
4. *Comparable to humans.* With the same $\mathcal{E}$ and p^*, the dilemma-state predicate can be evaluated over naturalistic human driving datasets [19], yielding a human DIR as the benchmark.
5. *Cross-checkable.* Two fleets running similar envelopes in the same cities should observe comparable $\text{DIR}_{\text{other}}$. Divergence signals envelope inconsistency or gaming. Contact rates reported under NHTSA's Standing General Order [20] independently check the DIR-to-contact ratio.

Interpretation. DIR turns the public argument from "whom would your car kill?" into "your DIR_{self} fell from 4.1 to 2.7 this quarter." The former has no winning answer. The latter is an argument a company can win, in public, on evidence.

6 Analysis: Why Equal Weighting Is Acceptable to Riders

The strongest objection to Rule 3 is empirical [2]: riders will not accept a vehicle that might sacrifice them. We show that equal *weight* on persons produces highly unequal *risk* to persons, in the occupant's favor, so the feared scenario is rare under the policy as stated.

Pedestrian risk. Table 1 reproduces the AAA Foundation analysis of U.S. pedestrian crash data [13]. Severe-injury (AIS 4+) probabilities are higher than fatality probabilities at every speed.

Table 1: Probability of pedestrian death as a function of vehicle impact speed [13].

Impact speed (mph)	Pr[death]
16	0.10
23	0.25
31	0.50
39	0.75
46	0.90

Occupant risk. For a belted occupant of a modern passenger vehicle in a frontal impact with a fixed rigid object at 30 mph, fatality probability is in the low single-digit percent range; against another passenger vehicle of similar mass, lower still [14]. Side-impact and rollover regimes are worse but are not the regimes produced by in-lane braking or a controlled swerve into a barrier.

Proposition 1. At typical urban closing speeds, $S_{\text{pedestrian}} \gg S_{\text{occupant}}$ for the same ΔV; consequently the unweighted minimizer of (3) selects the pedestrian-protecting trajectory in nearly every realistic configuration, while the occupant's absolute risk under that trajectory remains low.

Sketch. From Table 1 and the occupant figures above, the ratio $S_{\text{pedestrian}}/S_{\text{occupant}}$ at 30 mph is on the order of 10 or more. For the harm minimizer to prefer striking the pedestrian, the collision probability with the pedestrian along the alternative trajectory would have to be an order of magnitude lower than the collision probability with the barrier or vehicle along the pedestrian-avoiding trajectory, which contradicts the premise that the pedestrian is in the conflict set at $p_i \geq p^*$. The occupant's absolute risk is bounded by S_{occupant} at the impact speed, which Rule 1 keeps low by construction. □

Two tons of crumple zone, airbags and restraints supply the weighting that a "protect occupants first" rule would have added redundantly. The scenario that drives the survey result, a car choosing certain occupant death over pedestrian harm, requires a configuration (high speed, rigid unyielding obstacle, no braking room) that Rule 1 is designed to make rare and Rule 5 further discourages.

What the policy asks of a rider is what we already ask of a human driver: that they will not run down a child to protect a bumper. What the alternative asks of everyone else is to share the road with a vehicle that will kill a pedestrian to spare its occupant a minor injury, and no city will license that vehicle. Equal weighting is the only rule that survives publication in every market simultaneously. Uncertainty, the legitimate core of the "protect the occupant you can be sure of saving" argument, is already represented in p_i and in Rule 5; it should not be smuggled in as an additional coefficient on whoever paid for the ride.

7 Implementation

7.1 Modular stacks

In a perception–prediction–planning architecture, Rule 1 is a constraint on the nominal planner plus an occlusion-aware speed heuristic; Rules 2–5 constitute the objective and constraints of the emergency planner that activates when the nominal planner has no feasible solution; Rule 6 is logging. The novel engineering is the DIR instrumentation, the counterfactual attribution pipeline, and the selection and publication of $\mathcal{E}$. Operators that already publish crash comparisons to human benchmarks [7] are positioned to add DIR with modest effort.

7.2 End-to-end stacks

A single network from sensors to controls cannot have rules edited into it. It can be governed by two standard mechanisms.

Training objective. Rules 1–5 are expressible as a cost over predicted outcomes and can be incorporated into simulation-based training, particularly for the emergency regime, where human demonstration data is sparse and, more importantly, unsuitable: human drivers in the final second before a collision are precisely the behavior a learned policy should not imitate.

Runtime shield. A compact, verifiable monitor between the network and the actuators computes $\mathcal{F}_t$, $\mathcal{C}_t$ and H from the same perception inputs and overrides the network's output when it would violate Rule 1 or Rule 4 by margin, substituting the nearest compliant action [15, 16]. This is the flight-envelope-protection pattern from commercial aviation: the pilot flies, the envelope holds. The shield is also where DIR is computed, so the metric does not depend on trusting the learned policy.

Auditing Rule 3 on a learned policy. Demographic information is latent in raw pixels, so the interface restriction cannot be applied at the network boundary. The defensible claim is instead: (a) the shield and the harm model provably do not consume such features, and (b) the network is trained and tested with counterfactual augmentation such that varying the apparent demographic attributes of a rendered agent, holding kinematics and size fixed, changes control output by no more than measurement noise. Claim (b) is a falsifiable test that an operator with large-scale simulation infrastructure can run and publish.

7.3 Logging

Rule 6 requires retention of roughly 10 seconds of belief state, candidate trajectories and per-candidate harm estimates on each collision or dilemma-state entry. This is a modest extension of existing event-data-recorder practice and should be standardized across operators so that logs are comparable in litigation and regulatory review.

8 Governance: From Liability to Safe Harbor

Disclosure is currently a pure cost to operators, which is why no crash policy has been published. We propose exchanging it for protection, in three parts.

1. **Publication.** The doctrine, $\mathcal{E}$, p^*, the severity scale and quarterly DIR (self/other) are filed through the existing Standing General Order channel [20] and published.
2. **Safe harbor.** Where logs demonstrate that the vehicle operated within its published, regulator-accepted doctrine, the *choice* the vehicle made in the dilemma state is not itself actionable as negligence and is not a basis for punitive damages. Fault for the crash falls where the logs show it falls, including on the operator when the entry is attributed *self* and traceable to a defect in Rule 1 compliance.
3. **No-fault compensation.** Victims are compensated promptly from a fund financed by a per-mile levy on autonomous fleets, modeled on the National Vaccine Injury Compensation Program [17], which addressed a structurally identical problem: a technology safer than its alternative in aggregate, causing rare identifiable harms, and being driven from the market by case-by-case litigation.

The strategic point for operators is first-mover advantage in standard-setting. Volvo's 2015 liability pledge [18] set the terms against which every other manufacturer was subsequently measured. The first operator to publish a crash doctrine and a DIR series will define the envelope, the threshold and the severity scale that regulators adopt. The second will adopt a doctrine written by its competitor.

9 Limitations

The envelope is a political parameter. $\mathcal{E}$ determines how much misbehavior by others the fleet must absorb, and therefore how cautious and how slow it is. This is a speed-limit-type decision that belongs with regulators and the public. The doctrine makes the parameter explicit; it does not choose its value.

Injury models are approximate. Severity curves are population averages with wide confidence intervals, and vehicle-to-vulnerable-road-user models at oblique geometries are less mature than frontal models. The standard proposed is documented, uniform and improvable, not exact.

Learned policies are bounded, not interpreted. A shield constrains an end-to-end network's behavior in the dilemma state; it does not explain the network's behavior elsewhere. Rule 3 on a learned policy is a statistical claim under counterfactual testing, not a structural guarantee.

Attribution is counterfactual. The self/other split depends on re-planning under a stricter set and on the envelope. Different operators' attribution pipelines will differ until standardized.

Residual tragedy. No policy prevents every death. The claim is legitimacy under scrutiny: after a fatal crash the operator can state what the vehicle was trying to do, prove it did that, show that the unavoidable state is entered rarely and less often each quarter, and point to compensation already paid.

10 Conclusion and Recommendations

The crash dilemma is unsolvable as posed and need not be solved as posed. Reframed as a policy under uncertainty, it yields a six-rule doctrine with a decisive residual answer that does not depend on knowing who anyone is. Reframed as a measurable state, it yields a leading indicator, DIR, that any operator can compute from existing logs and publish within a quarter. Reframed as a

governance problem, it yields a disclosure-for-safe-harbor exchange with clear precedent.

We recommend that autonomous-vehicle operators, jointly or separately:

1. Publish a crash doctrine in the form of Rules 1–6, adapted to their stacks, including $\mathcal{E}$, p^* and the severity scale.
2. Add DIR, split self/other, to their safety reporting, and compute a human-driver DIR on a public naturalistic dataset under the same parameters as the benchmark.
3. Adopt a common 10-second pre-event retention standard including candidate trajectories and per-candidate harm.
4. Propose the safe-harbor and no-fault fund to NHTSA and state legislatures, with the per-mile levy priced.

The operator that does this first will stop being asked whom its car would kill. It will be asked what its DIR was this quarter, and it will have an answer.

References

[1] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The Moral Machine experiment," *Nature*, vol. 563, pp. 59–64, 2018.

[2] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," *Science*, vol. 352, no. 6293, pp. 1573–1576, 2016.

[3] Federal Ministry of Transport and Digital Infrastructure (Germany), *Ethics Commission: Automated and Connected Driving, Report*, June 2017.

[4] M. Taylor, "Self-driving Mercedes-Benzes will prioritize occupant safety over pedestrians," *Car and Driver*, Oct. 2016; and subsequent Daimler clarification.

[5] National Highway Traffic Safety Administration, *Early Estimate of Motor Vehicle Traffic Fatalities*, most recent annual release. (Verify current figure before citation.)

[6] World Health Organization, *Global Status Report on Road Safety 2023*, Geneva, 2023.

[7] Waymo LLC, *Safety Impact Hub*: rider-only crash comparison to human benchmarks. (Verify current miles and percentages before citation.)

[8] N. J. Goodall, "Machine ethics and automated vehicles," in *Road Vehicle Automation*, G. Meyer and S. Beiker, Eds. Springer, 2014, pp. 93–102.

[9] P. Lin, "Why ethics matters for autonomous cars," in *Autonomous Driving: Technical, Legal and Social Aspects*, M. Maurer et al., Eds. Springer, 2016, pp. 69–85.

[10] S. Nyholm and J. Smids, "The ethics of accident-algorithms for self-driving cars: an applied trolley problem?" *Ethical Theory and Moral Practice*, vol. 19, pp. 1275–1289, 2016.

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

[12] E. Rosén and U. Sander, "Pedestrian fatality risk as a function of car impact speed," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 536–542, 2009.

[13] B. C. Tefft, "Impact speed and a pedestrian's risk of severe injury or death," *Accident Analysis & Prevention*, vol. 50, pp. 871–878, 2013; AAA Foundation for Traffic Safety technical report, 2011.

[14] D. C. Richards, "Relationship between speed and risk of fatal injury: pedestrians and car occupants," UK Department for Transport, Road Safety Web Publication No. 16, 2010.

[15] L. Sha, "Using simplicity to control complexity," *IEEE Software*, vol. 18, no. 4, pp. 20–28, 2001.

[16] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," in *Proc. AAAI Conference on Artificial Intelligence*, 2018.

[17] National Childhood Vaccine Injury Act of 1986, 42 U.S.C. § 300aa; U.S. Health Resources and Services Administration, *National Vaccine Injury Compensation Program*.

[18] Volvo Car Group, "US urged to establish nationwide federal guidelines for autonomous driving," press release including liability statement, Oct. 2015.

[19] Virginia Tech Transportation Institute, *SHRP 2 Naturalistic Driving Study*; Waymo LLC, *Waymo Open Dataset*.

[20] National Highway Traffic Safety Administration, *Standing General Order 2021-01: Incident Reporting for Automated Driving Systems and Level 2 ADAS*, as amended.