ISSN 2754-1118 (Online)
ISSN 2753-6599 (Print)

Breaking the Tyranny of Net Risk Metrics for Automated Vehicle Safety

Philip Koopman[1] and William H. Widen[2]

1. Carnegie Mellon University, Pittsburgh PA, USA
2. University of Miami School of Law, Miami FL, USA

Abstract

An inquiry into how safe might be "safe enough" for automated vehicle technology must go far beyond the superficial "safer than a human driver" metric to yield an answer that will be workable in practice. Issues include the complexities of creating a like-for-like human driver baseline for comparison, avoiding risk transfer despite net risk reduction, avoiding negligent computer driver behaviour, conforming to industry consensus safety standards as a basis to justify predictions of net safety improvement, avoiding regulatory problems with unreasonably dangerous specific features despite improved net safety, and avoiding problematic ethical and equity outcomes. In this paper we explore how addressing these topics holistically will create a more robust framework for establishing acceptable automated vehicle safety.

1 Introduction

1.1 Overview

The obvious answer to how safe automated vehicles (AVs) should be is "net at least as good as a human driver". The AV industry typically argues that an expectation of a national reduction in net risk of fatalities warrants deployment of self-driving cars on public roads. This narrative dominates policy decisions, public messaging, and other aspects of stakeholder discussions of safety.

We argue that this focus on a net fatality risk metric is counterproductive to long-term success of the technology because that metric cannot be measured accurately at the early stages of deployment. Moreover, additional considerations will impact broad-based stakeholder acceptance. Reduction in net harm is a highly desirable goal, but over-overemphasizing this metric will likely back-fire on the industry due to neglecting other crucial metrics, ultimately contributing to a loss of trust in the technology. To meet with public approval in both the short and long terms, the AV industry must break free from the tyranny of a narrow net-risk metric approach.

In this paper, we explain that acceptably safe AVs must satisfy criteria along multiple different dimensions simultaneously. These criteria include:

- Achieving a Positive Risk Balance (PRB) for comparable conditions
- Mitigating risk transfer onto vulnerable populations

- Avoiding negligent computer driver behaviour
- Conforming to industry consensus safety standards
- Meeting regulatory requirements for risk on a fine-grain basis
- Addressing ethical and equity concerns

1.2 Previous Work

The topic of "how safe is safe enough" for autonomous vehicles remains a continuing discussion.

The BMVI (2017) Ethics Commission Report performed early influential work, creating a set of ethical rules for automated and connected vehicular traffic. Specific contributions included: requiring a positive balance of risks (no worse than human drivers), establishing a responsibility for regulators to ensure safety, minimizing risk to vulnerable road users, prohibiting decision logic regarding which specific road users to harm in an unavoidable crash, imposing responsibility on manufacturers for automated driving system safety rather than human vehicle occupants/drivers, and prioritizing preservation of human life vs. damage to animals and property. The BMVI report also addressed security and privacy concerns, along with other important considerations such as a need for people to retain some measure of control over equipment operation in appropriate circumstances. Our current work builds on that prior work, refining principles based on a half-decade of experience during which deployment of the technology has expanded.

The European Commission (2020) report on Ethics of Connected and Automated Vehicles, which we shall refer to as the "EC Ethics Report" sets forth twenty recommendations that cover not only safety, but also other issues such as informational privacy. It is a policy-oriented document which covers some of the same ground as our work, with sections on road safety risk, data/algorithm ethics, and aspects of responsibility. Its discussions are generally compatible with our work. We emphasize identifying different aspects of risk and safety that should be addressed in creating "safe enough" criteria that in some cases go beyond the scope of the EC Ethics Report, such as responder role contributions to safety.

A recent book focuses on this topic (Koopman 2022), emphasizing technical aspects of interest to a manufacturer wishing to determine that deployment is acceptably safe. Here we extend the scope to incorporate a broader range of concerns, including legal ones.

The SASWG (2022) published safety assurance objectives for all autonomous systems, generally emphasizing technical system properties in a safety engineering context, concentrating on computation, autonomy architecture, and system platform. That SASWG document also contains appendices with extensive cross-comparisons across other previous work to which we refer the reader. The SASWG work and its sources emphasize technical safety engineering considerations (e.g. how to ensure a predetermined level of safety). Here we add regulatory, legal, and a broader range of ethical issues into the scope of considerations.

The UK Centre for Data Ethics and Innovation published a study on the topic of Responsible Innovation for Self-Driving Vehicles (CDEI 2022), emphasizing legal and regulatory frameworks. That document does not directly present a framework for deciding whether a particular vehicle is safe enough for public road operation, though it does identify significant policy issues.

Chapter 5 of a Law Commission (2022) report proposes two paths for authorization for an AV to operate on public roads. One is Type Approval per international (UNECE[1]) standards. Another requires assurance of "at least an equivalent level of safety and environmental protection". Chapter 4 of that report additionally recommends avoiding risk transfers to vulnerable groups, which we also discuss in this paper.

Burton et al. (2020) cover a range of assurance considerations including engineering, ethical and legal, identifying a series of gaps: semantic, responsibility, and liability. They propose a framework including a safety case, dynamic assurance, soft law, and regulations that we believe is compatible with our findings, but takes a somewhat different approach.

This paper continues by covering each identified aspect of acceptable safety. It concludes with a composite statement of the relevant factors that should be accounted for in setting criteria for an acceptably safe autonomous vehicle deployment. We identify additional previous work specific to individual topic areas in the corresponding sections.

1.3 Terminology and Legal Framework

Here is a summary of some key technical terms and abbreviations used in this paper.

Automated Vehicle (AV): a vehicle with a computer driver which can completely carry out the driving task. A vehicle which has no requirement at all for a human driver might be called an "autonomous vehicle". However, in this paper we consider automated vehicles regardless of whether there is a human driver present, and focus solely on the safety of the computer driving function.

Computer Driver: a computer which controls steering and other aspects of motion control for a vehicle on public roads. A computer driver might or might not have a human backup driver who is monitoring operation, but who is not exercising sustained control of steering. (We sometimes use the term "AV" to refer to the behaviour of the vehicle as controlled by the computer driver to improve readability.)

Negligence: behaviour that fails to meet the level of care that someone of ordinary prudence would have exercised under the same circumstances (LII 2023).

Positive Risk Balance (PRB): the proposition that a computer driver should be no less safe (and ideally safer than) a human driver.

This paper is written based on the authors' knowledge of US laws. We understand that laws in other Western countries generally employ analogous principles. While some aspects of non-US law are addressed, that treatment should not be considered comprehensive. The treatment of legal issues is at level of abstraction such that differences should not alter our conclusions.

[1] The United Nations Economic Commission for Europe

2 Positive Risk Balance

2.1 Background

The notion of Positive Risk Balance (PRB) was a key contribution of the BMVI Ethics Commission report (BMVI 2017), captured in its rule number 2. While it notes that the long-term objective is to completely prevent harm to people, the report states that the "licensing of automated systems is not justifiable unless it promises to produce at least a diminution in harm compared with human driving, in other words a positive balance of risks". Moreover, BMVI's rule number 3 states that technologically unavoidable residual risks can be acceptable so long as PRB is achieved.

Since that time, the German car industry in particular has used PRB as its guiding star for acceptable safety.

2.2 How Much Better than a Human Driver?

While it seems intuitive that AVs should be at least as safe as human drivers, that is not the only possible approach. A long-view utilitarian argument might hold that the expected eventual reduction in fatalities from reduced road deaths morally justifies an increased fatality rate in the short term to speed up the development of a safety panacea. The AV industry might be said to obliquely take that approach via its efforts to minimize regulatory oversight and speed deployment based on arguing there is a moral imperative to reduce road deaths. Such a narrative glosses over the ethical problems with potentially near-term increasing risk to road users while hoping technology matures to provide promised benefits. A testing fatality in 2018 (NTSB 2019), well before one might have expected such an event based on human driver fatality rates, illustrates that dynamic, and justifies concerns over elevated near-term risks due to hurried development efforts.

Some data suggest that a modest PRB might not be sufficient. Liu et al. (2018) found that while a risk of 4 to 5 times as safe as a human driver might be tolerable, a broadly acceptable risk goal would be a hundred-times improvement in safety. It remains unclear whether public stakeholders will accept only a modest net safety improvement. One report popularized the notion that 10% better than human drivers should dictate a decision to deploy the technology (Kalra and Groves 2017), essentially arguing that any measurable decrease in fatalities creates an imperative to deploy sooner rather than later.

2.3 Multi-dimensional PRB Comparisons

While a safety metric of "better than a human driver" sounds intuitively appealing, actually measuring such an outcome requires establishing a comparable baseline for a human driver.

> ***Hypothetical example:*** *A new AV with high-end active safety features being driven in fair weather on empty city streets is compared to a statistical average human driver baseline. That baseline human driver is operating a statistically average lower-trim vehicle 12 years of age, with less capable, outdated safety features. The baseline human driver is also operating in conditions that include dangerous secondary roads, twilight, snow, and with some aggregate fraction of impaired driving.*

Defining and characterizing a human driver baseline can be surprisingly complex, involving contributions from factors such as (Koopman 2022):

- Driving environment (city/urban, light/dark, dry/wet/snow/ice, road maintenance condition, class of roadway, prevalence of vulnerable road users, local driving customs, etc.)
- Vehicle type (weight, installed passive safety features, installed active safety features, maintenance condition, etc.)
- Driver demographics (driver age, driver experience, any driver impairment, any driver distraction, any violation of road rules by driver, etc.)
- Victim demographics (age, pedestrian/bicyclist/motorcyclist/driver/passenger, etc.)

These factors can have a dramatic impact on the baseline crash rate and expected fatalities for human drivers, often comparable to the excess safety factor of 4 or 5 times safer that might otherwise have been thought adequate. As a single example, roadway type accounts for approximately a factor of 5 difference in fatality rate per mile in Pennsylvania when comparing the safest roadway (the Pennsylvania Turnpike) to the most dangerous type (non-Interstate system highways) (PennDOT 2022, page 16).

To be a fair comparison, a PRB calculation would need to include a weighted average of contributions to harm from data accounting for various contributing factors. Some factors might be excluded for policy reasons, such as not including risk contributions from drunk drivers as part of a policy decision to make the baseline an unimpaired driver.

Creating a human driver baseline would require detailed data for human driver rates of harm in various combinations of circumstances. While this is technically achievable, it is far from a simple comparison of national average fatality rates. It is possible that a very large safety margin (an order of magnitude or more) could simplify the comparison process so that minor factors could be neglected in analysis, but only at the cost of demanding much higher AV safety performance than might be strictly necessary to achieve credible PRB.

2.4 PRB and Risk Subsidy

A more subtle issue has to do with the possibility of risk subsidy.

> ***Hypothetical example:*** *AVs have a 10% reduced net fatality rate, accounting for all relevant environmental and driver type PRB conditions, but not equipment type. Upon review, it becomes apparent that the entirety of this safety improvement is attributable to the installation of an Automatic Emergency Braking (AEB) feature on all AVs. In comparison, the older human-driving vehicle fleet that includes a high proportion of lower-cost vehicles has a comparatively low installation rate of AEB. Moreover, it is found that turning on an AV computer driver actually reduces safety compared to manual operation of that same vehicle in comparable conditions (but not enough to be as bad as human driver outcomes in the average vehicle fleet). The computer driver itself reduces safety, but this is masked by the AEB feature safety improvement when compared to a fleet of predominantly non-AEB-equipped vehicles.*

This type of scenario is one possible outcome for automating driving that requires continuous human driver supervision. Data analysis from Goodall (2023) shows that early adopters of an automation feature saw an 11% adjusted estimated crash rate increase with the automation feature on vs. off. However, net safety with the automation feature on was

still said to be much better than for the US vehicle fleet which, on average, lacked comparable active safety features.

We characterize such a situation as a risk subsidy: some active safety technology is added to increase net safety. Then a vehicle automation feature is added that increases total harm substantively, but not enough to make the vehicle worse than it would be without that additional active safety technology. Public stakeholders might consider an AV safety claim based on a risk subsidy to be a misrepresentation of the safety benefits of AV features.

2.5 PRB as One of Many Metrics

Even if manufacturers go to the significant lengths required to implement a PRB metric with detailed weightings to account for the various factors involved, it is unlikely that this metric will show conclusive evidence of fatality reductions during initial deployments in the first several years of operation. The reason is a simple statistical significance issue. Current outcomes are approximately one fatality per 74 million miles driven (NHTSA 2023a), and one fatality per 192 million miles in the UK (UK DfT 2022). Given that at the time of this writing robotaxi companies are claiming perhaps one to three million miles each (Bidarian 2023), there might be two more orders of magnitude mileage accumulation to get a meaningful understanding of fatality outcomes. Accounting for miles driven by different ADS models remains problematic, as does accounting for accumulation of miles across multiple software upgrades.

Because property damage and injury crashes are much more prevalent than fatalities in human driver data, statistical confidence as to PRB outcomes for those less severe instances of harm will come earlier. However, it remains to be seen how accurate fatality predictions will be based on less severe crash rates, even with a sophisticated “inverting the diamond” analysis and prediction approach such as that used by Waymo (Victor et al. 2023). A significant confounder for early predictions is the potential for common cause failures, such as a defective software update causing a cluster of high-severity crashes before the update could be rolled back.

Additionally, there are risks from large disruptions in the operational environment causing failures due to the general brittleness of machine learning technology to novel situations. A large power outage affecting traffic signals (Fleischer 2023), overloaded mobile data networks (Hawkins 2023), a bodged software update, or other adverse event that affects hundreds of cars simultaneously could potentially cause large numbers of severe common cause mishaps in ways that human drivers are unlikely to experience, potentially invalidating forecasting approaches based on an assumption of random independent failures. An overarching concern is that while putting the same computer driver in every vehicle provides a basis for fleet-wide learning and improvement, it also introduces a source of common cause failures across a deployed fleet.

Another potential issue with pure PRB approaches is that in practice they emphasize aggregate total risk, and aggregate total harm. The intuitive appeal is that if we can reduce total road fatalities by even 10% over a relevant human driver baseline, that will be an improvement in highway safety and should be considered a victory for adoption of AV technology. However, there are possible outcomes of net PRB that might still be societally unacceptable.

Hypothetical example: *Total fatalities are reduced by 50%, but every single person harmed is a child boarding or debarking a school bus. (Scenario inspired by a Tesla mishap (Krisher 2023)).*

Hypothetical example: *Total pedestrian fatalities are reduced by 90%, but every single fatality is due to an AV rolling through stop signs. (Scenario inspired by a Tesla recall (Gitlin 2022)).*

While perhaps too specific to be likely in practice, these hypothetical outcomes are intended to illustrate the point that there are aspects of socially acceptable safe outcome characteristics that go beyond pure aggregate PRB. As one researcher noted, "*An AV that kills 1,000 fewer car occupants but 100 additional pedestrians may not be acceptable, even though 900 net lives are saved*" (Goodall 2021).

PRB summary: Positive Risk Balance is a necessary but insufficient condition for safety outcomes that are likely to be broadly acceptable to a wide range of stakeholders. Other considerations such as the demographic profile of victims and whether loss events are associated with the AV violating road rules are likely to be relevant as well. Moreover, computing PRB is complicated because of the need for a comparison to a detailed human driver baseline that accounts for varied factors.

3 Risk Transfer

3.1 Preamble

Even if net harm is reduced, it might well be that transfer of risk from one population or demographic segment onto another is considered societally unacceptable. The EC Ethics Report (2020) addresses this with its recommendation 1, noting that no category of road user should be at risk of increased harm, even if net harm to all road users has been reduced. That report goes further, suggesting in its recommendation 5 that AVs should adapt their behaviour to redress existing risk inequalities.

3.2 Statistical Risk Transfer

Hypothetical example: *A computer driver is involved with a series of crashes at emergency response scenes, causing both injuries and fatalities. Public pressure forces the manufacturer to deploy a software update to address the issue even before data is available as to whether this is a worse risk than that presented by human drivers in similar situations. (Scenario inspired by Tesla investigation by NHTSA (Hawkins 2022).)*

A more general issue is that if net harm is decreased, but harm to a distinguishable group of road users is increased compared to a relevant baseline, that risk transfer seems unlikely to be acceptable to public safety stakeholders. This applies especially if the group seeing an increase in harm from AVs is considered particularly vulnerable. Examples of such population segments might include: road workers, people with sensory impairments, people with mobility restrictions, children, the elderly, pedestrians, bicyclists, wearers of distinctive

ethnic clothing styles, people with darker skin tones, and historically disadvantaged groups. This issue will apply not only to harm from operation, but also potential harm that might occur during public road testing of the technology (Widen 2022). A similar risk transfer occurs between present road users and future road users if present road users are exposed to an increased risk of harm on the expectation that the lessons learned from earlier deployments will increase safety for future road users — an ethically controversial deployment decision (Widen 2023). This situation also raises the question of the appropriate discount rate for anticipated future benefits, both for time value of money and likelihood of realization (Johnsson and Voorneveld 2018).

A related issue is if there is an identifiable pattern in loss events that might be possible to avoid with a design improvement. AV proponents argue that the perfect should not be the enemy of the good (Kalra and Groves 2017), meaning that the technology should be deployed as soon as there is a net statistical reduction in harm. However, it is predictable that systematic losses that might be mitigated at reasonable cost will produce adverse reactions from at least some public safety stakeholders — even if the net harm is reduced.

3.3 Intentional Risk Transfer

A different sort of risk transfer might happen intentionally based on design decisions. The classic example of this is the so-called Trolley Problem applied to AVs. In that version of the Trolley Problem, a computer driver is said to be presented with a binary no-win situation in which somebody will necessarily die (Koopman et al. 2021). The question is which victim(s) the computer driver should choose, often couched in terms of some multi-dimensional utilitarian calculation of the number of lives and the comparative value of each life involved. Burton et al. (2020) discuss why such a framework is an unhelpful analogy for AV safety. EC Ethics Report (2020) recommendation 6 is to proactively address such situations.

BMVI (2017) rejects the notion that a calculus of any sort should be used to offset one life against another in an unavoidable crash, especially in its rule number 9. On a per-crash basis, BMVI rule 9 also requires no transfer of harm onto those not directly benefiting from vehicle automation. For example, an innocent bystander pedestrian should not be sacrificed in hopes of avoiding a crash that would be fatal to multiple vehicle occupants. However, BMVI also admits that there might be some circumstances in which reducing the total number of innocents harmed in an unavoidable crash is justifiable, so this is still an open topic to some degree.

While the question of intentional risk transfer can raise serious moral questions, for the near term those incidents should be a small fraction of the total number of mishaps. (If this is not true, likely the computer driver has much more serious safety issues.) A relatively simple behavioural constraint might suffice pending further study of the problem.

> ***Hypothetical example:*** *An AV is designed so that it tries to avoid harming any person. However, once its software deems it reasonably likely to inflict harm on one or more people, its manoeuvring options are restricted to avoid harming any additional people who would not already be harmed by the existing trajectory plan.*

With this example design approach, the computer driver is permitted to act to reduce the severity or attempt to avoid harm, but is not permitted to intentionally harm any individual who is not already expected to be harmed when it becomes apparent that a loss event is unavoidable. This policy would, for example, prohibit swerving to avoid striking a tree if

that swerve would instead run over a single pedestrian — even if doing so would be likely to save multiple vehicle occupants from severe harm. While such a strategy is likely not optimal in at least some utilitarian sense, there is no general agreement on what "optimal" might really mean in practical situations. Until consensus is reached on a better strategy (if that is even possible), it is a defensible strategy aligned with BMVI guidelines (BMVI 2017) that is relatively straightforward to define and implement.

Risk transfer summary: AVs should minimize or eliminate risk transfer. This is likely a practical requirement for public safety and equity stakeholders, even if doing so arguably limits the potential safety benefit of the technology by requiring constrained loss mitigation strategies.

4 Lack of Negligent Driving Behaviour

4.1 Negligence

An additional limitation on AV safety is likely to be a prohibition against negligence. We define *negligence* in this context as driving behaviour which, if exhibited by a human driver for a relevant set of conditions, would be considered negligent (or reckless) according to applicable laws, statutes, ordinances, and regulations that might apply. The scope of negligence must include both tort law and criminal law. Simply put, this is holding computer drivers to the same standards of negligent and reckless driving behaviour that already apply to human drivers.

4.2 An example of Negligent Computer Driving

We do not here attempt a comprehensive treatment of legal issues. Rather, we use an example to illustrate relevant concerns.

> ***Hypothetical example:*** *An autonomous vehicle is driving on city streets. A licensed driver who initiated autonomous operation an hour earlier is present in the vehicle, but asleep in a reclined seat as provided by the manufacturer. ("Take a nap and leave the driving to us!") The computer driver runs a red light and strikes a pedestrian in a marked crosswalk, killing the pedestrian instantly. The police arrest the sleeping vehicle occupant for negligent homicide or a similar offence.*

In this example, if the human driver had been driving, that human driver likely would be found negligent in the absence of significant extenuating circumstances such as brake failure, and potentially charged with negligent homicide. This is due to the situation in which violation of a traffic law (running a red light) directly led to a fatality under a presumption of negligence *per se*.[2]

The question of how such a situation is handled when a computer is driving is unsettled. The authors have proposed that the computer driver should be held to the same standards as

[2] "A defendant who violates a statute or regulation without an excuse is automatically considered to have breached her duty of care and is therefore negligent as a matter of law" … "The most common application of negligence *per se* is traffic violations, where the driver is automatically considered negligent for violating the traffic code." Legal Information Institute. (2020 update). Definition of "*negligence per se*". Wex (Cornell). Available at: https://www.law.cornell.edu/wex/negligence_per_se

a human driver, with the manufacturer held to be the responsible party in such a situation (Widen and Koopman 2023a).

Regardless of how negligence might be handled legally, as a practical matter, it seems highly desirable that computer drivers exhibit vanishingly small amounts of negligent driving behaviour. Deploying vehicles that ignore traffic signals because the designers think that it is safe to do so per their design should not permitted. Computer drivers should obey traffic laws. If manufacturers find traffic laws overly restrictive, they should lobby governments to change the traffic laws instead of flouting them[3]. Regulators seem to agree with this statement. For example, Tesla vehicles were recalled for not coming to a full and complete stop at stop signs (NHTSA 2022a).

BMVI (2017) holds that product liability should be the governing principle for harm caused by computer drivers. While product liability should be a possible avenue for collecting compensation for harm, conventional fault-based tort law also should provide an avenue for recovery because it is a more efficient and cost-effective process for attributing liability in garden-variety accident situations in which a computer driver might have behaved in a clearly negligent manner.

This issue goes beyond arguing for a general respect for the rule of law. Tort law and criminal law do not recognize statistical safety arguments to determine the fact of culpability. To put it bluntly, someone who saves 1000 lives does not get a coupon for one free homicide. Someone who has a perfect driving record for 40 years does not get a free pass for causing a crash by running a red light while drunk. While a person's history and character might be weighed in determining an appropriate penalty, guilt is determined by the facts in the particular case, not the history of the individual. For example, a first time Driving Under the Influence offender may qualify for a diversion program unavailable for repeat offenders (Bieber 2023).

By the same token, a computer driver that is documented to cut fatality rates in half should not get a free pass on negligent driving in a particular case, unless the public policy approach is to pre-empt tort law and criminal liability for computer drivers. Even so, it seems unlikely that public stakeholders will welcome AVs that cause fatalities associated with egregious traffic rule violations that they perceive would not have been excused for a competent human driver, even if total fatalities are reduced.

4.3 Responder-Role Safety

Lack of negligent behaviour is a highly desirable aspect of safety, but seems unlikely to ensure acceptable safety on its own. In particular, blaming other drivers for easily preventable crashes is likely to degrade safety outcomes if blame for contributory negligence or comparative fault results in the exclusion of data from safety metrics.

> ***Example:*** *An autonomous vehicle is making an unprotected turn across oncoming traffic on a multi-lane road at night. There is an oncoming vehicle that is speeding. The AV calculates that the oncoming vehicle must slow down to make a turn onto a side street as required by a pavement "turn only" lane marking, and that this gives enough time for the AV to turn into that same street ahead of the oncoming car. The oncoming car does not slow down, and does not make the turn. The AV, sensing a failed plan, executes an emergency safety stop in the oncoming car's lane, with a subsequent*

[3] EC Ethics Report Rule 4 reasonably suggests considering revision of traffic rules and setting non-compliance policies for AVs.

collision. The AV company states that since the oncoming car was more at fault due to speeding and not making the required turn, the multiple injuries from the collision should not count against the AV's safety record. (Inspired by a robotaxi mishap (CA DMV 2022).)

This example involving both vehicles as contributors to the incident illustrates important aspects of the interaction of blame and safety.

If the other vehicle had not been speeding, more blame might have fallen on the AV instead. There is no indication in the crash report that the speeding was a factor in the computer driver's decision making. A crash still would likely have happened if similar timing issues had been involved with a non-speeding oncoming vehicle. This suggests that the decision to make the left turn in this case was unduly risky, based as it was on an assumption that another road user would manoeuvre despite lack of indication from that other vehicle of a manoeuvring intention[4].

Arguably the reaction of the AV to perform an immediate stop when it calculated that it might crash is problematic. There are circumstances in which making an immediate stop might increase risk rather than reduce it, such as making a sudden stop during an unprotected left turn with high-speed oncoming traffic. This illustrates that the action taken by a computer driver when there is a problem will sometimes need to be more than a simple emergency stop. Stopping in ways that block emergency responders provides an additional illustration of the problem with assuming that stopping is always a safe behaviour (Nicholson et al. 2023).

Finally, the AV would have been better off if it had calculated that there was likely uncertainty in the behaviour of an oncoming vehicle, and waited for that vehicle to pass before initiating the left turn. Occurring late at night as this mishap did, it is likely that a paucity of traffic gave plenty of opportunity to make the turn safely after that oncoming speeding car had passed.

A more general view of this topic is that the role of a responder matters for safety (Victor et al. 2023). Even though one vehicle might behave in a negligent manner (whether by a human or computer driver), harm might still be avoided by other responding vehicles taking evasive actions. For example, the computer driver might have determined that the speeding car might not slow down, and instead waited to make the left turn. In another scenario a responder might slow down to delay entering an intersection when it detects cross traffic is likely to proceed through a red traffic signal. In practice, responders avoiding crashes that might otherwise be blamed on negligent driving by the other driver can make a substantial contribution to road safety.

From a public stakeholder perspective, it is likely unacceptable for AVs to make seemingly "stupid" driving decisions (based on what public perception of a reasonable human driver behaviour might be) such as turning in front of an oncoming speeding car, and then attempting to claim innocence because the other driver was found more than 50% at fault, despite a substantive fraction of blame being assigned to the AV in a crash investigation.

This has three practical implications for setting acceptable safety characteristics:

[4] We note that the October 2, 2023 loss event in San Francisco which resulted in the suspension of activities by Cruise LLC may present an issue of comparative fault. We do not use that example because, at the time of this writing, the facts of that case remain under investigation.

1. Computer drivers should have robust skills in the role of an incident responder. This roughly corresponds to having good defensive driving skills.
2. As a more specific part of defensive driving behaviours, computer drivers should have robust models of other road user potential behaviours that includes other vehicles violating traffic rules in readily foreseeable ways (speeding, rolling stops, using an incorrect lane, entering an intersection shortly after their traffic signal turns red) and pedestrians not using official crosswalks, which contributed to the Uber ATG pedestrian fatality (NTSB 2019).
3. Risk analysis of safety manoeuvres such as in-lane stops should account for potential harm that might occur after the stop, such as being hit by other road vehicles or trains, and disrupting emergency response services.

4.4 The Role of Blame in Safety

Care must be exercised when invoking the notion of blame in setting acceptable risk criteria. Lack of legal fault for AV behaviour does not equal an acceptable safety outcome when AV behaviour is a contributing factor to an accident, even if not the primary factor.

> ***Hypothetical example:*** *A robotaxi fleet is hit from behind frequently, each time blaming the trailing driver for the crash. The net result is, however, that the robotaxis are involved in twice as many rear-end crashes as a human driven baseline, even though not a single crash is blamed on the AV. (Scenario inspired by (Stewart 2018).)*

Computer drivers being hit from behind at low speeds is commonly attributed to driving behaviour deviating from the expectations of a trailing human driver, sometimes being characterized as the AV being overly cautious in comparison to prevailing human driving norms. A commonly cited reason for such crashes is the AV displaying so-called "phantom braking" behaviour in which an AV panic stops for no reason that is discernible to trailing vehicles. While the blame by default is imposed on the trailing car in the crash, blame assignment on its own seems unlikely to improve a trend in rear-end crashes in those circumstances.

While it is desirable to avoid at-fault crashes for AVs, safety metrics should not exclude crashes that are not the AV's fault as determined by legal standards. Rather, they should show that both (a) at-fault crashes are better than an ordinary at-fault human driver baseline, and (b) total crashes (both at-fault and not-at-fault) are also better than an overall human driver baseline. Proportional fault should not be used to claim that an AV was not at fault unless the proportion of fault assigned is negligible (perhaps less than a few percent), and not merely 50% or less fault as found in some comparative fault systems[5].

The EC Ethics Report (2020) recommendation 19 goes further, recommending a fair system for attribution of moral and legal culpability. That report's recommendation 20 is to establish a fair and effective mechanism for granting compensation to those harmed by an AV.

Negligence and Blame Summary: AVs should minimize the rate of crashes caused by computer driver behaviour that would be considered negligent it performed instead by a human driver. In particular, an improvement of net statistical safety should not be used as

[5] A pure comparative fault system considers all fault from whatever source and in whatever percentage. A modified comparative fault system does not allow recovery if the plaintiff is 50% or more at fault. A contributory negligence system allows no recovery if the plaintiff is found to have any fault.

an excuse to forgive negligent computer driver behaviour. Net safety metrics must include all crashes, not just those in which blame has been assigned to the computer driver, to encompass the contribution of defensive driving skills to net safety outcomes.

5 Standards Conformance

A significant challenge in defining acceptable AV safety is validating the accuracy of leading indicators (Kalra and Groves 2017) to predict safety outcomes that might not be statistically measurable until many years in the future. Sophisticated predictive approaches based on human-centric safety improvement techniques can be applied to provide improved confidence. However, in the final analysis there are assumptions and threats to validity to any predictive technique. Net risk prediction accuracy, especially for fatalities, will not be established until hundreds of millions of miles of real-world road usage have been accumulated.

The AV industry as a whole has been focussed on public road testing as a way to predict eventual safety[6]. However, in other industries a primary method of assuring safety deployment is following industry consensus standards. It is remarkable that the automotive industry, in sharp contrast to other life-critical technology industries, has historically not been required to follow its own consensus safety standards to deploy on US public roads[7].

One way to improve confidence in predictions of eventually acceptable safety would be for AV manufacturers to follow industry consensus standards. While the list of potentially relevant standards is large and growing, some key international standard candidates include: ISO 26262:2018, ISO 21448:2022, ANSI/UL 4600:2023, ISO/SAE 21434:2021, and SAE J3018_202012. Additionally, a Safety Management System should be in place, perhaps structured using advice from the Automated Vehicle Safety Consortium (AVSC 2021).

A typical industry talking point against standards is that they are said to "stifle innovation." This is the sort of thing that tends to be said by those who see safety practices as an impediment to risky innovation, such as the creator of the Titan mini-sub who perished in an implosion event (Musumeci and Guenot 2023), or companies simply looking for an excuse to evade regulatory oversight.

Standards tend to be written in the metaphorical blood of past mishaps. Moreover, the standards referred to above do not constrain the specific technology used to implement AVs. Rather, they require the use of hazard and risk analysis approaches with accompanying mitigation approaches. Those standards also require accounting for known issues with life critical systems such as risks posed by common cause failures, and encourage addressing problems that are foreseeable enough that they have been included in an international standard.

Standards Summary: There is no need for the AV industry to relearn lessons the hard way via loss events that could be avoided. Standards conformance is a key technique other industries use to establish a justifiable belief in acceptable safety before deployment.

[6] The EC Ethics Report (2020) points out in its recommendation 3 that road testing introduces its own risks.

[7] The authors are not aware of an ISO 26262 conformance requirement for any country for conventional road vehicles. It seems that ISO 26262 is strongly encouraged in Germany for fully automated vehicles, e.g. per Appendix 1-Autonomous Vehicles Approval and Operation Ordinance (AFGBV), Part 1, 1.3 Planning of routes and speeds; 7.2.1 Hazard Analysis Available at: https://www.buzer.de/Anlage_1_AFGBV.htm?m=26262#hit. The authors thank Gabi Escuela for bringing this to our attention.

6 Regulatory Requirements

6.1 Safe Enough?

Any AV will additionally need to meet regulatory requirements specific to the country or region in which it is being deployed. The rules differ by region, but conformance to relevant rules is a required part of the characterization of acceptable safety.

Regulators also need a criterion by which to decide what might be "safe enough", which is commonly expressed in terms of an acceptable risk threshold. In the US and the EU, the prevailing criterion is "Absence of Unreasonable Risk" (AUR), whereas in the UK the primary criterion is "As Low As Reasonably Practicable" (ALARP).

6.2 Testing-centric Requirements

Testing-centric requirements base a compliance determination on a test that can be replicated by an independent party on a series production vehicle. In the US, the Federal Motor Vehicle Safety Standards (FMVSS) serve this role (NHTSA 2023b).

The FMVSS suite requires specific safety-related functionality and safety feature performance. Topics range from tire pressure warning indicators to rear-view cameras to passenger crash safety features and more. FMVSS compliance is self-certified by manufacturers, and subject to audits by regulators after deployment. There is no pre-release regulatory approval process in the US for ground vehicles.

Of particular concern for AVs is that the FMVSS collection was created with many test procedures explicitly requiring the presence of a driver seat, steering wheel, and other equipment which might not be present in a cargo AV or completely driverless passenger AV. While the FMVSS 200 series on crash safety has been modified to address this constraint (NHTSA 2022b), work continues on other portions of FMVSS to remove inapplicable dependencies on human driver support equipment. A topic of intense political lobbying by the industry is increasing the number of FMVSS exemptions to permit scaled-up deployment until those other aspects of FMVSS can be addressed (Congress 2023). Granting a large number of exemptions can amount to deregulation, and not merely a deviation for small-scale controlled testing.

Europe has a type approval approach, with acceptance tests performed by independent parties under contract to manufacturers in a process known as homologation. That process results in a government-issued certificate granting permission to sell a particular vehicle (European Commission n.d.).

An additional aspect of testing-centric requirements is the NCAP[8] process, which involves a star rating customer information approach rather than a hard regulatory requirement for a particular level of performance. The theory is that consumers will favour vehicles with higher safety ratings disclosed at the time of sale. The US has come under criticism for its NCAP system lagging behind EURO-NCAP in adopting tests relevant to automated driving features (NTSB 2022).

It is possible for test-based regulations to extend to vehicle automation features, which was the case with the UN ECE #157 regulation on Automated Lane Keeping Systems (ALKS)

[8] New Car Assessment Program

(United Nations 2021). This is a regulatory approach for European approval of low-speed highway traffic jam pilot vehicle features.

A particular challenge to all types of regulatory approaches is the recent increase in frequency and safety relevance of over-the-air software updates that change a vehicle's software via a remote software update rather than a dealer visit. These are increasingly used not only to deliver remedies for safety recalls, but also to deploy new features which might introduce safety problems of their own. Regulators are still struggling to address over-the-air updates within their regulatory frameworks (Stumpf 2021).

6.3 Risk Reduction

Another concept used in regulations that can put a constraint on the acceptable boundaries of AV performance is that unreasonable risk should not be present, with a prevalent threshold being an Absence of Unreasonable Risk (AUR). AUR is a concept determined by a multi-point set of criteria in the US including: the utility of the product, the level of exposure to the risk, the nature and severity of hazards, and the likelihood of resulting harm. Also relevant are the state of the art, the availability of alternate designs, and the feasibility of eliminating the risk (US CFR 1992).

The US regulator — the National Highway Traffic Safety Administration (NHTSA) — uses AUR as a cause for pursuing regulatory enforcement action in addition to failure to comply with FMVSS (NHTSA n.d.).

In practice, AUR operates primarily at the feature level, not the vehicle level. The main mechanism available to NHTSA is a recall, which is a regulatory action taken only after a particular vehicle feature has been deployed on public roads. It is commonly the case that a recall is only pursued after a substantial number of incident complaints have been filed, and sometimes only after multiple reports of harm. NHTSA is, however, quietly experimenting with a regulatory strategy requiring immediate crash reporting and directing rapid recalls to change defective ADS software (Wansley 2022).

While AUR might be used as an overall design goal by manufacturers, in practice the application of AUR by regulators tends to be recalls for specific, documented dangerous behaviours or design deficiencies. It would require a dramatic expansion of their historical regulatory approach to ban an entire vehicle based on a failure to achieve net PRB or the like based on overall crash rates. While one initial somewhat broader recall has been promulgated (NHTSA 2023c), this seems to be breaking new ground for regulators.

In the UK, and some other countries, a general regulatory approach requires risks to be As Low As Reasonably Practicable (ALARP) or the equivalent (HSE n.d.).

At a high level of abstraction, ALARP has more similarities than might be apparent with an AUR approach. In implementation, the question comes down the how the cost/benefit decision is made in terms of how much risk reduction is "reasonable" or "reasonably practicable" for a particular AV in its expected operational environment. The regulatory emphasis in practice in both cases ends up being whether any particular risk or specific potential design defect could have or should have been mitigated further, rather than the overall statistical rate of harm.

Regulatory Summary: Regulations provide a combination of vehicle-level testing requirements and retrospective requirements to mitigate risks that emerge after deployment.

Neither of these are directly aligned with a PRB metric, and so create additional requirements to achieve acceptable safety.

7 Ethical and Equity Concerns

7.1 Preamble

While it is typical for a discussion of acceptable safety levels to emphasize technical and regulatory requirements, from a public acceptance point of view ethical and equity issues are also important. We consider the "Moral Crumple Zone" issue, broader ethical imperatives, and equity considerations that should factor in to ensuring acceptable safety.

7.2 Moral Crumple Zone

The concept of a Moral Crumple Zone is a design strategy to shield a technological system from blame at the expense of the nearest convenient human, regardless of whether it is reasonable to expect that human to have been able to avoid or mitigate loss events (Elish 2019). It deflects attention from system shortcomings which place humans in unrecoverable situations or otherwise set them up to fail.

The classical example of a moral crumple zone in the AV domain is the story of the tragic Uber ATG fatality.

> ***Real example:*** *An AV test vehicle strikes and kills a pedestrian at night. The driver is said to have been distracted by a mobile phone at the time of the crash, and did not notice the pedestrian in time to stop, even though adequate sight line and time were available. The technical cause of the mishap was a combination of tracking software defects, an expectation that pedestrians would only cross at official crosswalks, and the disabling of a manufacturer AEB system. A National Transportation Safety Board (NTSB) investigation found that the driver was not paying attention, but also found that profound deficiencies in safety culture at Uber ATG set the stage for the mishap. The company, Uber ATG, settled with survivors but was not charged criminally. The test driver was sentenced to supervised probation after a plea deal involving a felony endangerment charge (NTSB 2019) (Smiley 2023).*

A significant issue with the Uber ATG fatality is that one can argue the safety drivers were set up for failure due to lack of supervision, onerous working conditions, immature technology, and the inevitability of automation complacency. It seems likely a matter of when, and to which test driver — rather than if — such a crash would happen. While it is reasonable to hold test drivers accountable for a failure to diligently perform their duties, it does not serve the interests of safety to let companies use those same test drivers as a defensive shield against accountability for dangerously run test programs.

In the wake of the Uber ATG fatality, the industry updated the SAE J3018_202012 test driving safety standard to incorporate lessons learned. However, no currently operational company publicly states that it conforms to that industry consensus standard[9].

[9] Argo AI was independently assessed to conform to SAE J3018 (PRNewswire 2021). However they have recently terminated operations.

A related issue is Tesla's use of retail customers as so-called "beta testers" for obviously immature automation technology. Tesla has succeeded in laying the blame for crashes in court on drivers for failing to mitigate driving errors made by the technology in an initial case (Roy et al. 2023). A criminal case similarly saw the driver take responsibility for a fatality that occurred while Autopilot was in use as part of a plea agreement (Dazio and Krisher 2023). It is unclear whether this trend of blaming drivers will continue despite concerns of the NTSB regarding automation complacency for this technology (NTSB 2019).

7.3 Ethical and Equity Considerations

There are many other ethical and equity considerations that should be incorporated as appropriate into acceptability criteria for a safety program. Some of them might be considered to stray a bit far afield from traditional functional and system safety. However, they can directly impinge upon and constrain design choices, activities, and deployment decisions that are relevant to safety. Examples include:

- Whether developmental testing of potentially defective prototype vehicles is allowed in historically disadvantaged areas, where it can be argued that residents harmed are likely to receive lower compensation from defendants than would residents of other areas (Widen 2022).
- Whether states or municipalities should be pre-empted by higher layers of government from instituting limitations on testing or deployment of AVs responsive to specific hazards in their local communities, such as a prohibition on testing in sensitive locations such as active school zones (Widen and Koopman 2022).
- The level of public transparency afforded to mishap reports from manufacturers that are required by regulators. Current national-level mishap reports (NHTSA 2023d) have substantive redactions of arguably non-technical data such as crash locations in the name of manufacturer proprietary data secrecy.
- Whether an argument that an aspirational goal of eventual reduction in harm should be permissible as a justification for deploying immature technology on public roads when it is not known when (or even if) PRB and other acceptable safety criteria will be fulfilled.
- Who, if anyone, should be held accountable for negligent driving behaviour on the part of a computer driver. Current US laws are anything but clear on this topic (Widen and Koopman 2023b).
- The degree to which disruption of public safety functions and emergency responders (Nicholson et al. 2023) should be factored into PRB baselines and other safety metrics.

Conformance to the IEEE 7000-2021 standard might provide a way to identify and address ethical considerations from a wide range of stakeholders, but is not currently required by any jurisdiction of which the authors are aware.

Ethics and Equity Summary: A number of questions regarding ethics and equity remain open, but should be addressed by any proposed set of criteria for acceptable safety. A particular area of practical concern is avoiding the use of test drivers or retail customer drivers as Moral Crumple Zones to shield manufacturers from accountability for potentially defective, immature driving automation features.

8 Summary

Recapping the section summaries, acceptable safety criteria for a computer driver should encompass all of the following areas:

- Positive Risk Balance using a baseline that accounts operating conditions and vehicle safety features like-for-like.
- Minimized risk transfer, with no net risk transfer onto disadvantaged or vulnerable groups.
- Vanishingly small instances of negligent computer driver behaviour, with an additional burden to perform competently at defensive driving skills.
- Conformance to industry consensus safety standards.
- Absence of unreasonable risk (AUR) and/or ALARP risk mitigation on a behaviour-by-behaviour basis.
- Address ethical and equity issues, including transparency, accountability, and absence of a moral crumple zone strategy.

One could argue that this sets a template for other applications of safety critical automation beyond automotive that involve impinging on a human operator's ability to meet their duty of care to other stakeholders.

References

ANSI/UL 4600. (2023). *Evaluation of Autonomous Products*. ANSI/UL 4600, 3rd Edition, approved by the American National Standards Institute, 2023. Underwriters Laboratories, Chicago. https://www.shopulstandards.com/ProductDetail.aspx?productid=UL4600 Accessed 16th January 2024.

AVSC. (2021). *AVSC Information Report for Adapting a Safety Management System (SMS) for Automated Driving System (ADS) SAE Level 4 and 5 Testing and Evaluation*. Automated Vehicle Safety Consortium (AVSC) AVSC000007202107. https://www.sae.org/standards/content/avsc00007202107/. Accessed 16th January 2024.

Bieber C. (2023). *First Offense DUI: Everything You Need To Know*. Forbes Advisor https://www.forbes.com/advisor/legal/dui/first-offense-dui/. Accessed 16th January 2024.

Bidarian N. (2023). *Regulators give green light to driverless taxis in San Francisco*. CNN.COM, August 11, 2023. https://www.cnn.com/2023/08/11/tech/robotaxi-vote-san-francisco/index.html. Accessed 16th January 2024.

BMVI. (2017). *Ethics Commission Report: Automated and Connected Driving*, Bundesministerium für Verkehr und digitale Infrastruktur, the Federal German Ministry of Transport and Digital Infrastructure. https://perma.cc/6UBX-KH5G. Accessed 6th January 2024.

Burton S., Habli I., Lawton T., McDermid J., Morgan P., and Porter Z. (2020). *Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective*. Artificial Intelligence, Vol. 279, Feb. 2020, 103201. Preview available at: https://doi.org/10.1016/j.artint.2019.103201. Accessed 6th January 2024.

CA DMV. (2022). *Report of Traffic Collision Involving an Autonomous Vehicle, June 10, 2022*. State of California Department of Motor Vehicles. https://www.dmv.ca.gov/portal/file/cruise_060322-pdf. Accessed 6th January 2024.

CDEI. (2022). *Policy Paper: Responsible Innovation in Self-Driving Vehicles*. UK Department for Science, Innovation and Technology Centre for Data Ethics and Innovation. 19 August 2022. https://www.gov.uk/government/publications/responsible-innovation-in-self-driving-vehicles. Accessed 6th January 2024.

Congress. (2023). *Hearing: Self-Driving Vehicle Legislative Framework: Enhancing Safety, Improving Lives and Mobility, and Beating China*. Subcommittee on Innovation, Data, and Commerce, Committee on Energy and Commerce. US House of Representatives, July 26, 2023. https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=116277. Accessed 16th January 2024.

Dazio S., and Krisher T. (2023). *As a criminal case against a Tesla driver wraps up, legal and ethical questions on Autopilot endure*. Associated Press, August 15, 2023. https://apnews.com/article/tesla-autopilot-los-angeles-d02769ba359cf6381dc1176c3f5a72a5. Accessed 16th January 2024.

Elish M. C. (2019). *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*. Engaging Science, Technology, and Society (ISSN 2413-8053), Volume 5, pp 40–60 https://estsjournal.org/index.php/ests/article/view/260/177. Accessed 16th January 2024.

European Commission. (2020). *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility*. European Commission Directorate-General for Research and Innovation. https://data.europa.eu/doi/10.2777/035239. Accessed 6th January 2024.

European Commission. (n.d.). *Frequently Asked Questions — Type Approval of Vehicles.* European Commission Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs https://single-market-economy.ec.europa.eu/sectors/automotive-industry/technical-harmonisation/faq-type-approval-vehicles_en. Accessed 6th January 2024.

Fleischer M. (2023). *Watch S.F. traffic officers try to get this stuck autonomous Cruise car to move*. San Francisco Chronicle, August 3, 2023. https://www.sfchronicle.com/opinion/article/san-francisco-police-self-driving-cars-cruise-18277009.php. Accessed 16th January 2024.

Gitlin J. (2022). *Tesla recalls 53,822 cars because they won't stop at stop signs*. Ars Technica, February 1, 2022. https://arstechnica.com/cars/2022/02/tesla-recalls-53822-cars-because-they-wont-stop-at-stop-signs/. Accessed 16th January 2024.

Goodall N. (2021). *Potential Crash Rate Benchmarks for Automated Vehicles*. Transportation Research Record, 2675(10), pp. 31–40. https://doi.org/10.1177/03611981211009878. Accessed 16th January 2024.

Goodall N. (2023). *Normalizing crash risk of partially automated vehicles under sparse data*. Journal of Transportation Safety & Security, 16:1, pp. 1–17. https://www.tandfonline.com/doi/full/10.1080/19439962.2023.2178566. Accessed 16th January 2024.

Hawkins A. (2022). *The federal government's Tesla Autopilot investigation is moving into a new phase.* TheVerge,com, June 9, 2022. https://www.theverge.com/2022/6/9/23161365/tesla-autopilot-nhtsa-crash-investigation-emergency-vehicle. Accessed 16th January 2024.

Hawkins A. (2023). *Robotaxis are driving on thin ice*. TheVerge,com, August 15, 2023. https://www.theverge.com/2023/8/15/23831170/robotaxi-cpuc-sf-waymo-cruise-traffic-halt. Accessed 16th January 2024.

HSE. (n.d.). *ALARP "at a glance"*. UK Health and Safety Executive. https://www.hse.gov.uk/enforce/expert/alarpglance.htm. Accessed 16th January 2024.

IEEE 7000. (2021). *IEEE Standard Model Process for Addressing Ethical Concerns during System Design* (September 15, 2021). Institute of Electrical and Electronics Engineers, Piscataway, NJ.

ISO 21448. (2022). *Road vehicles — Safety of the intended functionality*. ISO 21448:2022. International Organization for Standardization, Geneva. https://www.iso.org/standard/77490.html. Accessed 16th January 2024.

ISO 26262. (2018). *Road vehicles — Functional safety*. ISO 26262, in 12 parts, 2nd Edition, 2018. International Organization for Standardization, Geneva. https://www.iso.org/standard/68383.html. Accessed 16th January 2024.

ISO/SAE 21434. (2021). *Road vehicles — Cybersecurity engineering*. ISO/SAE 21434:2021. SAE International, Pittsburgh and International Organization for Standardization, Geneva. https://www.iso.org/standard/70918.html. Accessed 16th January 2024.

Jonsson A., and Voorneveld M. (2018). *The Limit of Discounted Utilitarianism*. Theoretical Economics 13, pp. 19–37. https://onlinelibrary.wiley.com/doi/pdf/10.3982/TE1836. Accessed 16th January 2024.

Kalra N., and Groves D. (2017). *The Enemy of Good: Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles*. RAND Corporation Research Report RR-2150-RC. https://www.rand.org/pubs/research_reports/RR2150.html. Accessed 6th January 2024.

Koopman P., Kuipers B., Widen W., and Wolf M. (2021). *Ethics, Safety, and Autonomous Vehicles*. IEEE Computer, December 2021, pp. 28–37. https://ieeexplore.ieee.org/document/9622307. Accessed 16th January 2024.

Koopman P. (2022). *How Safe Is Safe Enough?: Measuring and Predicting Autonomous Vehicle Safety*. Independently Published, September 2022. ISBN: 979-8848273397.

Krisher T. (2023). *US probes crash involving Tesla that hit student leaving bus*. Associated Press, April 7, 2023. https://apnews.com/article/tesla-school-bus-student-hurt-firetruck-d282a5dd63874f22f5e1a6fc8168801b. Accessed 16th January 2024.

Law Commission. (2022). *Automated Vehicles: Joint Report*. Law Commission of England and Wales, and Scottish Law Commission, HC 1068 SG/2022/15. https://www.scotlawcom.gov.uk/files/4616/4313/7041/Automated_vehicles_joint_report_cvr_24-01-22.pdf. Accessed 6th January 2024.

LII. (2023). Negligence. Definition from Legal Information Institute, Cornell Law School. https://www.law.cornell.edu/wex/negligence. Accessed 6th January 2024.

Liu P., Yang R., and Xu Z. (2019). *How Safe Is Safe Enough for Self-Driving Vehicles?* Risk Analysis, Vol. 39 No. 2, pp. 315–325. https://doi.org/10.1111/risa.13116. Accessed 6th January 2024.

Musumeci N., and Guenot M. (2023). *OceanGate wrote that certifying its sub would block innovation. A longtime sub expert says the 'exact opposite is true'*. Insider, July 25, 2023. https://www.insider.com/oceangate-ceo-certification-innovation-sub-expert-opposite-is-true-2023-7. Accessed 16th January 2024.

NHTSA. (2022a). *Untitled*. Part 573 Safety Recall Report 22V-037, January 27, 2022. US DOT National Highway Traffic Safety Administration. https://static.nhtsa.gov/odi/rcl/2022/RCLRPT-22V037-4462.PDF. Accessed 16th January 2024.

NHTSA. (2022b). *Occupant Protection for Vehicles With Automated Driving Systems: Final Rule*. US DOT National Highway Traffic Safety Administration, March 30, 2022. https://www.federalregister.gov/documents/2022/03/30/2022-05426/occupant-protection-for-vehicles-with-automated-driving-systems. Accessed 16th January 2024.

NHTSA. (2023a). *Early Estimate of Motor Vehicle Traffic Fatalities in 2022*. US DOT National Highway Traffic Safety Administration Report No. DOT HS 813 428. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813428. Accessed 16th January 2024.

NHTSA. (2023b). *Federal Motor Vehicle Safety Standards and Regulations*. (Title 49 Code of Federal Regulations) US DOT National Highway Traffic Safety Administration. https://icsw.nhtsa.gov/cars/rules/import/FMVSS. Accessed 16th January 2024.

NHTSA. (2023c). *Untitled*. Part 573 Safety Recall Report 23V-085, April 11, 2023. US DOT National Highway Traffic Safety Administration. https://static.nhtsa.gov/odi/rcl/2023/RCLRPT-23V085-9893.PDF. Accessed 16th January 2024.

NHTSA. (2023d). *Standing General Order on Crash Reporting — For incidents involving ADS and Level 2 ADAS*. Amended April 2023. US DOT National Highway Traffic Safety Administration. https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting. Accessed 16th January 2024.

NHTSA. (n.d.). *Understanding NHTSA's Regulatory Tools: Instructions, Practical Guidance, and Assistance for Entities Seeking to Employ NHTSA's Regulatory Tools*. https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/understanding_nhtsas_current_regulatory_tools-tag.pdf. Accessed 16th January 2024.

Nicholson J., Luttropp D., Jones N., and Friedlander J. (2023). CPUC Status Conference: Safety Issues Regarding Driverless AV Interactions with First Responders. (slides presented at California Public Utilities Commission meeting, August 7, 2023). https://www.sfmta.com/sites/default/files/reports-and-documents/2023/08/2023.08.07_cpuc_status_conference_8.7.2023_final.pdf. Accessed 16th January 2024.

NTSB. (2019). *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018.* National Transportation Safety Board Highway Accident Report NTSB/HAR19/03, November 19, 2019. https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf. Accessed 6th January 2024.

NTSB. (2022). *Automobile Safety Rating System Is Failing Consumers*. National Transportation Safety Board, June 3, 2022. https://www.ntsb.gov/news/press-releases/Pages/NR20220603.aspx. Accessed 16th January 2024.

PennDOT. (2022). *2021 Pennsylvania Crash Facts & Statistics*. Bureau of Maintenance and Operations of the Pennsylvania Department of Transportation https://www.penndot.pa.gov/TravelInPA/Safety/Documents/2021_CFB_linked.pdf. Accessed 16th January 2024.

Roy A., Levine D., and Jin H. (2023). *Tesla wins bellwether trial over Autopilot car crash*. Reuters, April 22, 2023. https://www.reuters.com/legal/us-jury-set-decide-test-case-tesla-autopilot-crash-2023-04-21/. Accessed 16th January 2024.

SAE J3018_202012. (2020). Safety-Relevant Guidance for On-Road Testing of Prototype Automated Driving System (ADS)-Operated Vehicles. J3018, 3rd Edition, 2020. SAE International, Pittsburgh. https://www.sae.org/standards/content/j3018_202012. Accessed 16th January 2024.

SASWG. (2022). *Safety Assurance Objectives for Autonomous Systems*. Safety Critical Systems Club — Safety of Autonomous Systems Working Group, January 2022. https://scsc.uk/r153B:1?t=1. Accessed 6th January 2024.

Smiley L. (2023). *The Legal Saga of Uber's Fatal Self-Driving Car Crash Is Over*. Wired.com (July 28, 2023). Available at: https://www.wired.com/story/ubers-fatal-self-driving-car-crash-saga-over-operator-avoids-prison/. Accessed 16th January 2024.

Stewart J. (2018). *Why People Keep Rear-Ending Self-Driving Cars*. Wired, Oct. 18, 2018. https://www.wired.com/story/self-driving-car-crashes-rear-endings-why-charts-statistics. Accessed 16th January 2024.

Stumpf R. (2021*). Feds Order Tesla to Justify OTA Autopilot Updates Instead of Recalling Cars*. TheDrive.com, October 14, 2021 https://www.thedrive.com/tech/42736/feds-order-tesla-to-justify-ota-updates-instead-of-recalling-cars. Accessed 16th January 2024.

UK DfT. (2022). *National Statistics: Reported road casualties Great Britain annual report: 2021*. UK Department for Transport. https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2021/reported-road-casualties-great-britain-annual-report-2021. Accessed 16th January 2024.

United Nations. (2021). *Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping Systems*. UN Regulation No. 157, ECE/TRANS/WP.29/2020/81. https://unece.org/sites/default/files/2021-03/R157e.pdf. Accessed 16th January 2024.

US CFR. (1992). *Reporting of unreasonable risk of serious injury or death*. U.S. Code of Federal Regulations, 16 CFR § 1115.6. https://www.law.cornell.edu/cfr/text/16/1115.6. Accessed 16th January 2024.

Victor T., Kusano K., Gode T., Chen R., and Schwall M. (2023). *Safety Performance of the Waymo Rider-Only Automated Driving System at One Million Miles*. Waymo LLC. https://storage.googleapis.com/waymo-uploads/files/documents/safety/Safety%20Performance%20of%20Waymo%20RO%20at%201M%20miles.pdf. Accessed 16th January 2024.

Wansley M. (2022). *Regulating Driving Automation Safety*. 73 Emory Law Journal (forthcoming 2024), Cardozo Legal Studies Research Paper No. 689, August 15, 2022. https://ssrn.com/abstract=4190688. Accessed 16th January 2024.

Widen W. H. (2022). *Highly Automated Vehicles & Discrimination Against Low-Income Persons*. North Carolina Journal of Law and Technology, Vol. 24, No. 1, University of Miami Legal Studies Research Paper No. 4016783. https://dx.doi.org/10.2139/ssrn.4016783. Accessed 16th January 2024.

Widen W. H. (2023). *Automated Vehicles, Moral Hazards & the 'AV Problem'*. 5 Notre Dame J. Emerging Tech.1 (2023), University of Miami Legal Studies Research Paper No. 3902217. https://dx.doi.org/10.2139/ssrn.3902217. Accessed 16th January 2024.

Widen W. H., and Koopman P. (2022). *Autonomous Vehicle Regulation and Trust*. UCLA Journal of Law & Technology, Spring 2022, Volume 27, No. 3. http://dx.doi.org/10.2139/ssrn.3969214. Accessed 16th January 2024.

Widen W. H., and Koopman P. (2023a). *Winning the Imitation Game: Setting Safety Expectations for Automated Vehicles*. 25 Minn. J. L., Sci. & Tech. 113. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4429695. Accessed 22nd January 2024.

Widen W. H., and Koopman P. (2023b). *Level 3 Automated Vehicles and Criminal Law*. JURIST — Academic Commentary. https://www.jurist.org/commentary/2023/08/widen-koopman-automated-vehicles-criminal-law. Accessed 16th January 2024.