2019-01-0123 / 19AE-0135

Safety Argument Considerations for Public Road Testing of Autonomous Vehicles

Philip Koopman & Beth Osyk
Carnegie Mellon University; Edge Case Research

Abstract

Autonomous vehicle (AV) developers test extensively on public roads, potentially putting other road users at risk. A safety case for human supervision of road testing could improve safety transparency. A credible safety case should include: (1) the supervisor must be alert and able to respond to an autonomy failure in a timely manner, (2) the supervisor must adequately manage autonomy failures, and (3) the autonomy failure profile must be compatible with effective human supervision.

Human supervisors and autonomous test vehicles form a combined human-autonomy system, with the total rate of observed failures including the product of the autonomy failure rate and the rate of unsuccessful failure mitigation by the supervisor. A difficulty is that human ability varies in a nonlinear way with autonomy failure rates, counter-intuitively making it more difficult for a supervisor to assure safety as autonomy maturity improves. Thus, road testing safety cases must account for both the expected failures during testing and the practical effectiveness of human supervisors given that failure profile. This paper outlines a high level safety case that identifies key factors for credibly arguing the safety of an on-road AV test program. A similar approach could be used to analyze potential safety issues for high capability semi-autonomous production vehicles.

Introduction

Autonomous vehicle (AV) developers are testing extensively on public roads, potentially putting other road users at increased risk. The proper ratio of simulation, closed course testing, and road testing is a matter of debate. However, at some point any autonomous vehicle will have to undergo some form of road testing, if for no other reason than to demonstrate that a notionally "perfect" design based entirely on simulation and off-road testing really is sufficiently close to perfect in practice.

In other words, there will have be on-road AV "test flights" regardless of design methodology. And, most likely, the safety argument for test flights will be based on having a human supervisor who ensures safety in the event of an autonomy failure. Thus, it is essential to understand how to make human-supervised road testing of AV technology appropriately safe.

Scope

This paper explores the factors that affect whether public on-road AV testing will be sufficiently safe. We assume that testing is conducted on public roads, and that at least one human ***supervisor*** (sometimes called a "safety driver") is tasked with ensuring AV test platform safety (safety of the ***vehicle***). The vehicle includes an autonomy system being tested that is commanding an automotive or truck platform. On-road AV testing (referred to simply as "***testing***" in this paper) is assumed to be part of a larger autonomy validation approach that should include many other activities before road testing is performed.

We assume that the main argument that public testing is safe is that a supervisor is able to assume control of the vehicle and recover from autonomy malfunctions. This must happen with a sufficiently high probability that the testing does not present undue risk to the public, including other road users. While setting the risk threshold is ultimately a public policy question, we use an example threshold of the same risk as an average human driving a conventional, non-autonomous vehicle.

If a human supervisor is responsible for ensuring safety, it is essential that the supervisor be able to recognize a system malfunction when it occurs and intervene to effectively maintain safe vehicle operation whenever necessary. Specifically, the observed vehicle failure rate P_{Loss} will be related to the autonomy failure rate $P_{AutonomyFails}$ and the failure rate of a human supervisor to mitigate autonomy failures (which includes the degree to which the system can be effectively supervised in practice), $P_{SupervisionFails}$:

$$(1) \quad P_{Loss} = (P_{AutonomyFails} * P_{SupervisionFails}) + P_{HumanMistake}$$

An additional term has been added for the probability that a supervisor performs an unsafe action in the absence of an autonomy failure. This might include an unsafe intervention when no intervention was actually needed (e.g., a startle response or inadvertent disengagement without takeover).

It is well understood that humans are imperfect in general. An underappreciated fact is that, as discussed below, the probability of a supervisor failure can increase as the autonomy failure rate decreases. One way to look at this is that as failure rates decrease, the practical ***supervisability*** of the system also decreases. In other words, it is difficult for a human to effectively supervise very good, but imperfect autonomy.

Published Version: https://doi.org/10.4271/2019-01-0123

Therefore, the total failure rate of the human-AV system could increase, at least initially, as the autonomy failure rate decreases. Whether the increase poses an unacceptable risk depends upon various factors we describe.

We describe a safety argument intended for use with SAE Level 4+ autonomy systems [31] that are being field tested, and therefore require in-vehicle and/or remote human supervision. The safety argument described should also be largely applicable to deployed operation of partially autonomous vehicles (e.g., SAE Level 2-3 vehicles) and other partial autonomy schemes that rely upon human supervision to ensure production vehicle safety. However, the emphasis of this paper is on factors specific to test platforms operated by trained supervisors. We do not address practical issues of partially autonomous full scale deployment such as intentional misuse, driver training, and maintenance neglect.

Previous Work

This paper builds upon previous work on safety arguments using Goal Structuring Notation (GSN). [13] Previous work on AV safety arguments has focused on arguing the safety of automobiles [28] or the autonomy system itself. [4][11][37]

Road testing safety publications generally fall into two categories: regulatory publications and tester disclosures. Regulatory approaches typically emphasize the non-technical aspects of testing such as the permitting process, reporting, driver credentials, liability assignment, and so on. [1][5]

Pennsylvania has published guidelines [29] that describe a safety argument having to do with supervisor effectiveness and vehicle response to disengagements built upon our previous work, which also forms a basis for this publication. [16] Elements of those policies include requiring two safety drivers at speeds above 25 miles per hour, having a written plan to address driver ability to respond to autonomy faults, and written testing procedures.

Victoria Australia has passed legislation and published guidelines that require a safety argument, but do not specify a particular technical approach. [34][35]

A previous US Federal policy document dealt primarily with the safety of underlying vehicle test platforms (e.g., whether test platforms must comply with Federal Motor Vehicle Safety Standards), but were largely silent regarding how to ensure road testing safety beyond requiring that supervisors have conventional vehicle driving credentials. [22] A newer supplemental document acknowledges that it is important to ensure that safety drivers maintain vigilance, but does not provide details. [23]

Current metrics published for autonomous vehicle testing largely deal with the logistics of testing, such as number of vehicles deployed and miles driven. The most widely reported statistical data is so-called disengagement reports [5], which are not a sufficient basis for establishing safety. [2]

A fatal mishap graphically illustrated that merely placing a safety driver in a vehicle is insufficient to avoid testing fatalities. [27] A more robust argument of safety required to ensure that testing safety is achieved in practice.

Few companies doing road testing have made public statements on this topic beyond arguing that they have a “safety driver” present in the vehicle. Uber has announced that it will increase driver training, rotate two safety drivers periodically, and take additional measures to improve safety. [33] We are not aware of any published safety argument specifically addressing the issue of managing the nonlinear interaction between autonomy failure rate and human supervisor effectiveness.

A Probabilistic High Level Safety Argument?

It is desirable to have zero deaths and zero injuries attributable to road testing of autonomous vehicle technology. However, it is impracticable to achieve absolutely zero risk when operating any vehicle on a public road. Thus, there will always be at least some very low probability of a loss event. It is important to actively minimize the expected number and severity of loss events, especially when exposing the general public to a risk from testing performed by commercial interests who are strongly incentivized to deploy a new technology quickly.

We do not believe a precise risk calculation is feasible based on currently available data for the novel and quickly evolving technology of autonomous vehicles. However, a generic mathematical formulation can inform a safety argument. The probability of a loss event during on-road AV testing is:

$$(2)\quad P_{Loss}(i) = P_{Failure}(i) * ((1–P_{Detection}(i))+(1–P_{Mitigation}(i))) + P_{HumanMistake}$$

$$(3)\quad Risk = Sum(P_{Loss}(i) * Severity(i))$$

In other words, the probability of a loss event during AV testing due to a particular autonomy malfunction of type ***i*** is the probability that the autonomy fails *and* the combined probability that either the supervisor fails to detect the failure in a timely manner or the supervisor fails to execute a mitigation action in a timely manner. There is an additional contribution to the probability of failure if the supervisor performs an unnecessary but unsafe action, such as an unsafe intervention. The risk is the sum of all the probabilities of such loss events times the severity of each possible loss event.

As is commonly the case, a probabilistic formulation of risk has some issues. In this case, the interrelationship between $P_{Failure}$ and $P_{Detection}$ is non-linear. Autonomy that fails on a frequent basis keeps the supervisor alert, whereas autonomy that fails very infrequently can result in supervisor boredom and dropout. [38] Thus, assuming a constant human failure rate is incorrect.

The probability and consequence of a loss is also highly dependent upon the Operational Design Domain (ODD), which defines the general operational environment the system is intended to operate within. [22] For example, a low-speed shuttle operating in a benign environment with few obstacles, cooperative adult pedestrians, and light vehicular traffic could reasonably be expected to have comparatively low risk. A fully loaded truck operating in high-speed suburban rush hour traffic on secondary roads might be expected to have a comparatively higher risk for a given autonomy maturity level.

Enumerating all possible failures and the probability of timely and correct human supervisor response seems like an overwhelming task in the face of scarce historical on-road data for partially autonomous human supervision. (What data is available indicates that operators are imperfect in at least some situations. [26]) Getting more realistic data without incurring additional risk via public road testing is likewise difficult. Finally, the probability of failure of novel AV technology is generally a moving target, if it is known at all.

The reality is that AV testing is already occurring on public roads with little expectation of robust data for probability calculations in the near term. So while a probabilistic approach might be useful in the long term, in the near term it seems that a better use for the equation is to help identify areas that contribute to risk, even if exact numbers aren't yet available.

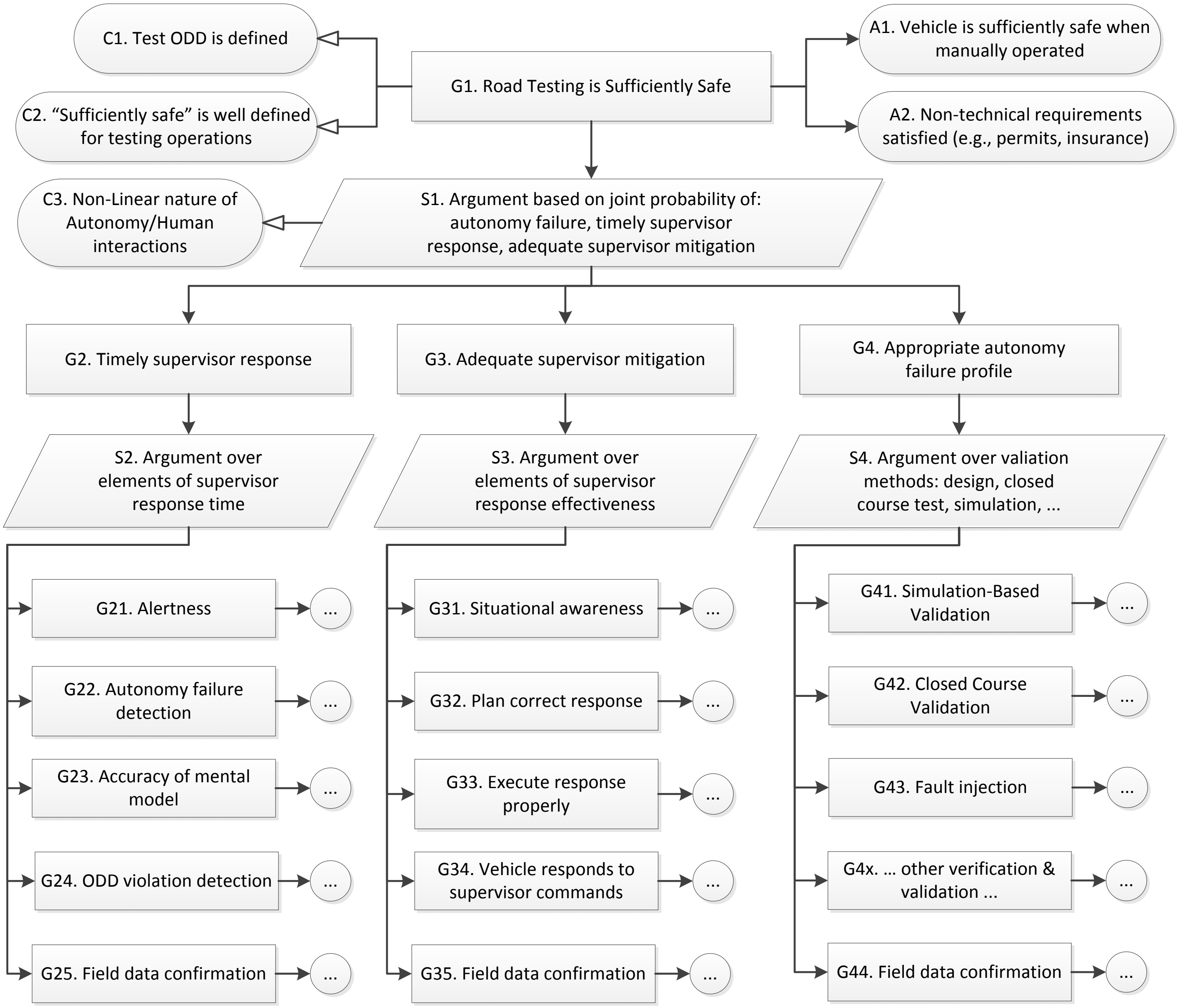
C1. Test ODD is defined
C2. "Sufficiently safe" is well defined for testing operations
G1. Road Testing is Sufficiently Safe
A1. Vehicle is sufficiently safe when manually operated
A2. Non-technical requirements satisfied (e.g., permits, insurance)
C3. Non-Linear nature of Autonomy/Human interactions
S1. Argument based on joint probability of: autonomy failure, timely supervisor response, adequate supervisor mitigation
G2. Timely supervisor response
G3. Adequate supervisor mitigation
G4. Appropriate autonomy failure profile
S2. Argument over elements of supervisor response time
S3. Argument over elements of supervisor response effectiveness
S4. Argument over valiation methods: design, closed course test, simulation, ...
G21. Alertness
G22. Autonomy failure detection
G23. Accuracy of mental model
G24. ODD violation detection
G25. Field data confirmation
G31. Situational awareness
G32. Plan correct response
G33. Execute response properly
G34. Vehicle responds to supervisor commands
G35. Field data confirmation
G41. Simulation-Based Validation
G42. Closed Course Validation
G43. Fault injection
G4x. ... other verification & validation ...
G44. Field data confirmation
...

Figure 1. GSN high-level on-road testing safety argument.

Thus, we propose a safety argument approach that enumerates factors affecting risk. Our approach permits identifying, monitoring, and reducing factors that increase risk without necessarily having a precisely quantified risk value. In other words, it is a framework for managing risk. A significant aspect of the argumentation is accounting for practical limits of supervisability. It can be deployed as part of a safety argument that might, for example, argue that each element of risk has been made As Low As Reasonably Practicable (ALARP) and that continuous monitoring of risk will permit correcting any issues or gaps in the argument as soon as they are discovered during field testing.

A Safety Case for Autonomous Vehicle Testing

Figure 1 shows a high level Goal Structuring Notation (GSN) safety argument for on-road AV testing with a human supervisor that is aligned with the elements of the previously discussed risk computation. Each element is labelled according to its role in the argument: **G**oal, **C**ontext, **S**trategy, and **A**ssumption. Unlabeled circles are used as place holders for more detailed arguments and evidence to support the argument for a particular system.

Assumptions and Top-Level Goal (G1)

The top level goal G1 assures that road testing is sufficiently safe. For the sake of this safety argument, we define the context as a particular set of ODDs selected for road testing (C1).

We assume that the vehicle is sufficiently safe for road use when under non-autonomous manual operation, just like any other vehicle authorized to operate on public roads (A1). We also assume that non-technical requirements such as permitting, insurance, driver licensing procedures and so on have been taken care of (A2). If any of these elements are not true (for example, due to using a bespoke test platform that is not street legal), additional safety case support would be needed.

Selecting a metric for “sufficiently safe” (C2) for at-scale deployment of AVs should properly be a matter of societal agreement. However, for practical purposes we note that any on-road testing program is likely to involve far fewer than 100 million actual vehicle road-miles. Additionally, the intense public scrutiny placed on AV mishaps likely motivates a desire to have a zero-mishap test program if at all possible. Therefore we consider a strategy for test organizations that want to set a safety goal of zero fatalities, zero serious injuries, and a relatively low probability of less severe mishaps during a comparatively small public road testing program. We further assume they will want to continually improve safety in response to field incidents and near misses.

Accepted practices for on-road testing, and in particular for driving training and workload management should be followed. See, for example, SAE J3018 [32], which can be used to provide argumentation support.

Argument Strategy (S1)

Once assumptions have been shown to be valid, the main safety argument S1 is that the joint probability of autonomy failure and a slow or ineffective supervisor response is sufficiently low. This results in sub-goals and arguments. We’ll discuss timely supervisor response (G2/S2) and adequate supervisor mitigation (G3/S3) to set the stage, and then explain why the autonomy failure rate (G4/S4) matters even though a human supervisor is assuming responsibility for safety.

Timely Supervisor Response (G2/S2)

If the autonomy fails in a way that is not self-mitigated, the supervisor must respond before a mishap can occur. While the root cause fault is important for engineering purposes, the pressing issue while operating on the road is ensuring that the vehicle is made safe. Thus, this analysis emphasizes supervisability, including human detection and responsiveness, rather than defects in the autonomy system. (Note: sub-goal numbers under G2 skip to G21 using a hierarchical scheme.)

Limits to Human Alertness (G21)

It is well known that humans have trouble remaining engaged as supervisors, even if they are trained and otherwise well qualified for the task. Historical attentiveness experiments are typically relatively short in duration, and do not account for the realities of months of long, boring supervisor shifts. In real-world conditions it is all too easy for supervisors to drift off task to an even larger extent. A recent example of an autonomous vehicle testing mishap included video of a supervisor looking away from the road and allegations of significant driver distraction contributing to the mishap. [18]

Having a second supervisor in the vehicle can help, but even then it is possible for two highly trained and motivated supervisors to not only lose attention, but even completely fall asleep. For example there are airline incidents involving allegations of both airline pilots having fallen asleep during commercial passenger flights. [24] There are several other recorded incidents of one or both pilots apparently falling asleep during flight. [25]

Driver monitoring systems are one approach to mitigating supervisor inattention. However, the accuracy of the technology should be assessed in practice. There will almost certainly be a non-zero monitoring system failure rate. And, even if monitoring works, it is possible that the driver will sleep through an alarm (e.g., as alleged in [9]).

In general, it is unrealistic to simply assume that even well trained supervisors will be completely engaged 100% of the time simply because they have been instructed to remain alert. A credible safety argument must allow for and mitigate the possibility of supervisors becoming disengaged, distracted, or even potentially falling asleep during testing sessions.

Detection of Autonomy Malfunction (G22)

Assuming the supervisor is alert, that supervisor must continuously monitor vehicle behavior and detect a malfunction when one occurs. That detection must happen fast enough to leave enough time for recovery. Therefore, the acceptable detection latency is a factor of the type of malfunction and the amount of time it will take to plan and execute a mitigation maneuver (G3/S3, discussed later).

It can be helpful to have self-diagnosis functions to help speed the detection time for detected autonomy faults. However, it is important to realize that some faults might not be detected. Worse, a supervisor becoming accustomed to a high probability of automated fault detection could result in over-reliance on that fault detection. That in turn could increase the latency for the supervisor realizing an undetected autonomy failure has occurred. The non-linear effects of high, but incomplete, self-diagnosis coverage interacting with supervisor complacency as well as any false alarm rates resulting in alarm fatigue should be considered as part of the safety argument.

Accuracy of Mental Model (G23)

There are at least two ways for a supervisor to approach autonomy failure detection. The first way is for the driver to intervene whenever the vehicle does anything that might possibly be incorrect, leaving very wide safety margins. If the autonomy system is prone to dramatically unsafe actions, such as commanding sudden turns into oncoming traffic, the supervisor might well keep the vehicle on a very tight leash – with substantial justification. This approach essentially involves the supervisor imposing a high false alarm rate on autonomy failure detection in exchange for minimizing the risk from potential failures.

However, as the autonomy function becomes more mature and demonstrates it is not prone to frequent, violent misbehavior, there will likely be a supervisor tendency to switch to a more permissive approach of waiting until the autonomy does something that is clearly unsafe before intervening. That involves giving the autonomy more latitude so long as it is reasonably well behaved. This more permissive approach seems especially likely if there is pressure to minimize supervisor interventions, such as when development progress is measured by reduction in reportable disengagement frequency.

In the case of a driver attempting to minimize false alarm disengagements, it is likely that the driver will, over time, learn the general behaviors of the vehicle, and intervene only when something unusual occurs or there is a clear and imminent danger. This is especially true if the AV generally drives along the same route in consistent driving conditions. However, there is a risk that accompanies this normal human response, which is that the supervisor might not have an accurate model of what is really going on inside the autonomy system, and might not realize that latent faults are present that just haven't been activated yet. In other words, the supervisor needs to build an expectation of vehicle behavior, but that expectation is likely to be invalid in some way due to the legibility problem. [8] An especially difficult problem will be if a particular situation is novel to the autonomy in some way that the human supervisor does not perceive, or does not consider to be a remarkable difference from previous experience. If the supervisor mis-diagnoses or fails to diagnose a vehicle failure, that provides an opportunity for a supervisor mistake to result in a mishap.

As an example, consider a fatal scenario in which a slow moving street cleaner partially obstructs a drive lane with a partially autonomous guidance system engaged (e.g., as in [3] and [7]). A cautious human driver might well reduce speed and attempt a complete lane change immediately upon seeing this anomalous situation. However, a supervisor trying to give the autonomy a chance to sort out the situation and respond appropriately might wait to act until too late and suffer a mishap. (It is unclear whether this is what actually happened in the cited crash. The citation is merely a concrete illustration of a potential mishap scenario.)

An additional problem is that unless the autonomy is specifically designed for transparency to human supervisors, it is likely that the human will have insufficient information available to judge the intentions of the system. For example, consider a vehicle approaching a traffic signal. Absent a heads-up or similar display, the supervisor has no way to know if the vehicle sees that the signal is red, or even if it sees the signal at all. While the supervisor can infer that the vehicle sees the signal if it slows down, it might be slowing down for an unrelated reason, or might not slow down at all despite a track record of previously detecting that same traffic signal many times in the past. [19]

An augmented reality display or other approach can help the supervisor understand the internal state of the autonomy. This can, for example, make it easier to detect if a traffic signal has been detected or if a pedestrian in the middle of the road has been missed by vehicle sensors. However, using this type of system carries its own costs and risks. If poorly designed it can cause supervisor cognitive overload. Additionally, there is a risk of complacency and accompanying delayed intervention time if the vehicle shows the supervisor that it perceives an obstacle or plans to follow an obstacle-free path, but then fails to react accordingly.

ODD Violations (G24)

Any practical autonomy system has limitations on its expected deployment environment, or Operational Design Domain (ODD). While a robust autonomy system should be able to detect violations of its ODD assumptions, a system under test might well have defects in that capability. Therefore the supervisor must not only monitor vehicle behaviors, but also undetected excursions from the intended ODD.

Some ODD violations might be relatively straightforward, such as geographic limitations, a constraint on weather conditions, or a constraint to only test in daylight. Other constraint violations might be subtle and perhaps difficult for the supervisor to detect, such as sensor degradation or subtle environmental factors such as operation in a thin haze.

Sensor degradation and object perturbations that do not particularly trouble humans can cause autonomy failures. For example, malicious alteration of road signs can cause autonomy failures. [10] A more subtle problem is that environmental degradation that is almost imperceptible to humans, such as slight amounts of haze or camera blur, can cause autonomy failures. [30]

There might be types of objects out of scope for an ODD that show up during test drives. For example, designers might have considered people wearing animal costumes to be out of scope. While October 31 might be excluded from the ODD to avoid encountering costumed pedestrians in a country that celebrates Halloween, it would then be up to the supervisor to terminate testing due to ODD violation when encountering a pack of children in costume going to a pre-Halloween party on an earlier date.

A specific challenge is that the supervisor must be cognizant in significant detail about the various aspects of the ODD. This includes which types of objects the system has been trained on, exactly what weather conditions are permissible, and other limitations to training data to detect ODD violations.

The supervisor must also monitor false alarm hazard reactions that surprise other human drivers who reasonably expect the test vehicle to act as if there is no hazard. As an example, panic braking for a non-existent obstacle (a false positive) could cause a rear-end collision with a trailing vehicle whose driver reasonably believed there was no obstacle in the road, and therefore was following too closely to avoid a crash. While such a mishap might be blamed on the trailing human driver, it is nonetheless a mishap involving the test vehicle.

Finally, it might be desirable for the supervisor to monitor the vehicle for erratic, inconsistent, or other behaviors that degrade public confidence. There might be a need to mitigate such behaviors to manage public perception of autonomy capabilities even if they do not pose a significant risk of a mishap.

Field Data Confirmation (G25)

Assuming the goal of a test program is zero fatalities and zero major crashes, the safety argument should include data feedback from elements contributing to the safety argument rather than waiting to react to a real-world collision.

Even if a safety plan is sound in principle, real-world issues can be expected to result in surprises or degraded safety performance, revealing unexpected coverage gaps in the safety case. Even the best-trained and qualified supervisors might be distracted by personal issues, slowed down by the onset of a minor illness, or simply have bad days. Emergent autonomy defects could cause nearly unrecoverable failures that take supervisors by surprise. Unexpected, potentially subtle ODD violations will crop up that take a while for supervisors to recognize. And so on.

To detect and eventually mitigate emergent problems and gaps in the safety case, any credible safety argument will need to have data fed back from real world experience to validate assumptions, calibrate residual risk expectations, and detect problems that crop up so that they can be mitigated before loss events occur.

Here are some examples of potentially useful feedback data for the supervisor response component of the safety argument:

- Fraction of time primary supervisor is alert with eyes on road, including distribution of time lengths when eyes were off road. (This cannot be 100% eyes on road, if for no other reason than eye fatigue and blinking. It seems likely to be substantially worse in practice without continual feedback.)
- Mean and distribution of time to detect autonomy failures, potentially including faults injected for training and measurement purposes.
- Analysis of incidents to determine the role of inaccurate supervisor mental model in contributing to near misses or false alarm disengagements.
- Sampling of data recordings to determine rate at which supervisors fail to recognize risky situations (near misses).
- Data on arrival rates and distributions of ODD violations

The goal for metric collection should not be perfection, but rather to have a realistic understanding of what areas of performance can and should be improved to minimize the risk of a mishap.

Adequate Supervisor Mitigation (G3)

Once the supervisor has detected an autonomy failure, the next step is planning and executing a mitigation procedure, which generally means performing a vehicle maneuver to put the vehicle in a safe state until the malfunction can be resolved. Accomplishing this requires situational awareness, planning, and execution of a safing maneuver.

Situational Awareness (G31)

The supervisor must remain aware of vehicle condition, environmental condition, other vehicles, obstacles, and situational aspects during operation. This includes both normal and abnormal situations. For example, the vehicle under test might suffer a mechanical breakdown or collision (e.g., autonomy equipment failure, tire blowout, road debris strike, side-swipe by another vehicle). Or it might be that another vehicle suffers an unexpected failure (e.g., cargo spill onto roadway, crash that blocks road), there is a roadway failure (e.g., rock slide, bridge collapse), or there is a sudden departure from the ODD that results in sudden loss of visibility (e.g., heavy rain squall, whiteout from wind-blown snow).

At the time an autonomy failure happens, the supervisor should already have a reasonably accurate mental model of the environment, other road users, and other relevant factors as an input into the response planning process. A lack of situational awareness can be expected to delay the response process while the supervisor spends time gathering information about the surrounding environment, or can result in an incorrect response if the supervisor feels an urgent need to react immediately despite inadequate situational awareness. As a simple example, if an in-lane obstruction suddenly appears, the supervisor must know if there are vehicles in adjacent lanes before initiating a swerving maneuver. Even for an in-lane braking maneuver, the supervisor should be aware of following traffic to manage the risk of a rear-end collision due to panic braking.

Plan and Execute Correct Response (G32, G33)

Ultimately it is the supervisor's job to either avoid a mishap or minimize the consequences if avoidance is infeasible.

Given adequate situational awareness, the supervisor must select an appropriate response for an autonomy failure. Depending upon the circumstances this response can range from simply re-engaging the autonomy after a false alarm disengagement to complex avoidance maneuvers to recover from a dangerous situation.

The likelihood of correctly planning and executing a sufficiently safe response will be influenced by situational awareness, training, and general experience of the supervisor. For example, a supervisor lacking situational awareness might take incorrect action or over-compensate for an undesirable vehicle motion as part of a startle response.

Planning and execution can take a substantial amount of time in novel, unexpected situations. There is data showing that humans can make incorrect decisions in the first few seconds of a high-consequence system failure. This is especially true for scenarios that have not been the subject of training. In some circumstances the chance of a human performing the wrong supervisory action in a short time window for unexpected, unrehearsed, high-consequence scenarios approaches 100%. (e.g., [36] Table 16.5)

A credible safety case involving human supervisors recovering from autonomy failures will need substantial data indicating that human supervisors are not only trained, but also exposed to effective in-service proficiency maintenance for high consequence autonomy failures.

Vehicle Response to Supervisor Commands (G34)

Once the supervisor takes control actions, the test vehicle must correctly respond. This can be complicated by modifications that might have been made to the test vehicle's underlying platform to enable autonomous operation.

Since the autonomy system has by definition failed when a supervisor intervenes, it is dangerous to assume that the autonomy's disengagement mechanism will actually work unless a specific safety argument has been made regarding those mechanisms. It is generally desirable to have an independent safety-rated disengagement mechanism that can override the autonomy system even if the autonomy is erroneously attempting to maintain control of vehicle operation. Such a safety mechanism should be designed in accordance with an accepted safety standard such as ISO 26262. [15]

The switchover from autonomous operation to manual supervisor operation brings with it a number of potential issues that can lengthen or degrade the ability to bring the vehicle to a safe state. Factors that should be addressed as part of the safety argument include:

- Time delay between initiating a disengagement and fully restoring driver control
- The update of vehicle control inputs to match driver inputs (e.g., matching vehicle internal state values to physical accelerator and brake pedal positions)
- Whether an autonomy malfunction can degrade or prevent supervisor control (e.g., a steering assist unit controlled by the autonomy system might be able to overpower human driver steering force)
- Whether transient signals during the disengagement process cause unexpected or exceptional behaviors such as control loop destabilization or program execution faults in underlying vehicle control software.

While the specifics will vary for each vehicle, it is difficult to imagine a safe design in which the primary autonomy function is itself entirely responsible for safety disengagements. Similarly, a system in which all human driver commands go through the autonomy system is likely to be flawed if an autonomy failure mode can cause the autonomy software to simply ignore supervisor control inputs.

An additional concern is false alarm takeovers by the supervisor. For example, consider a situation in which a supervisor is startled by an external event, or otherwise reacts to a perceived hazard in a way that causes a loss event when it is believed the autonomous vehicle itself would have avoided a mishap (e.g., [17]). While it is tempting to simply blame such outcomes on human supervisor mistakes, it is important to remember that humans are imperfect, and the inevitability of these types of events must be accounted for in the safety case.

Field Data Confirmation (G35)

As with supervisor responses, field data should be taken to quantify the effectiveness of supervisor mitigation. This data should encompass the various factors affecting the supervisor's ability to plan and respond to autonomy faults.

Appropriate Autonomy Failure Profile (G4)

An essential observation for the preceding sections for goals G2 & G3 is that they rely upon imperfect human responses to provide safety. There is some non-zero probability that the supervisor will not react in a timely fashion, and some additional probability that the supervisor will react incorrectly. Either of these outcomes could be an incident or mishap. Such a non-zero probability of unsuccessful failure mitigation means it is necessarily the case that the frequency of autonomy failures will influence on-road safety outcomes.

However, lower failure rates are not necessarily better. The types and frequencies of autonomy failures will affect the supervisability of the system. Therefore, the field failure rate and types of failures must be compatible with the measures being taken to ensure supervisor engagement. Thus, the failure profile must be "appropriate" rather than low.

In practice it might be desirable to induce intentional autonomy failures in low-risk scenarios. Such emergency drills could help maintain driver vigilance and malfunction response proficiency.

Non-Linear Autonomy/Human Interactions (C3)

A significant difficulty in reasoning about the effect of autonomy failure on safety is that there is a non-linear response of human attentiveness to autonomy failure. We propose that there are five different regions of supervisability of autonomy failures, with two different hypothetical scenarios based on comparatively lower and higher supervisability trends illustrated in Figure 2.

1. **Autonomy fails frequently in a dangerous way.** In essence this is autonomy which is not really working. A supervisor faced with an AV test platform that is trying to run off the road every few seconds should terminate the testing and demand more development. We assume that such a system would never be operated on public roads in the first place, making a public risk assessment unnecessary. (Debugging of highly immature autonomy on public roads seems like a bad idea, and presents a high risk of mishaps.)

2. **Autonomy fails moderately frequently but works or is benign most of the time.** In this case the supervisor is more likely to remain attentive since an autonomy failure in the next few seconds or minutes is likely. The risk in this scenario is probably dominated by the ability of the supervisor to plan and execute adequate fault responses, and eventual supervisor fatigue.

3. **Autonomy fails infrequently.** In this case there is a real risk that the supervisor will lose focus during testing, and fail to respond in time or respond incorrectly due to loss of situational awareness. This is perhaps the most difficult situation for on-road testing, because the autonomy could be failing frequently enough to present an unacceptably high risk, but so infrequently that the supervisor is relatively ineffective at mitigation. This dangerous situation corresponds to the "valley of degraded supervision" in the upper half of Figure 2.

4. **Autonomy fails very infrequently, with high diagnostic coverage.** At a high level of maturity, the autonomy might fail so infrequently that it is almost safe enough, and even a relatively disengaged driver can deal with failures well enough to result in a system that is overall acceptably safe. High coverage failure detection that prompts the driver to take over in the event of a failure might help improve the effectiveness of such a system. The ultimate safety of such a system will likely depend upon its ability to detect a risky situation with sufficient advance warning for the supervisor to re-engage and take over safely. (This

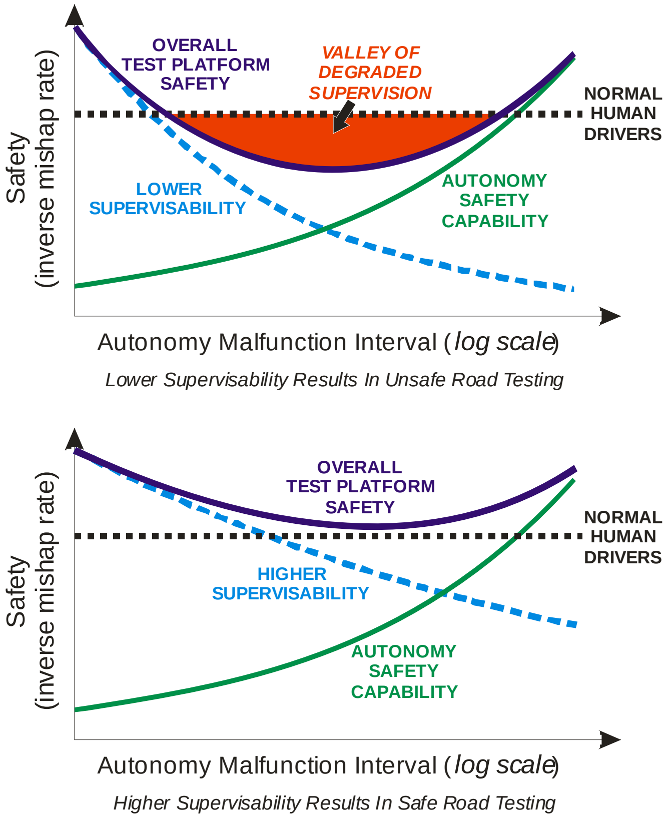

Figure 2. Hypothetical Testing Safety Performance Curves

scenario is generally aligned with envisioned production deployment of SAE Level 3 autonomy.)

5. **Autonomy essentially never fails.** In this case the role of the supervisor is to be there in case the expectation of "never fails" turns out to be incorrect in testing. It is difficult to know how to evaluate the potential effectiveness of a supervisor, other than that the supervisor will have the same tasks as the "very infrequently" preceding case, but is expected not to have to perform them.

Perhaps counter-intuitively, the probability of a supervisor failure is likely to *increase* as the autonomy failure rate decreases from regions 1 to 5 above (from left to right along the horizontal axis of Figure 2). In other words, the less often autonomy fails, the less reliable supervisor intervention becomes. The most dangerous operational region will be #3, in which the autonomy is failing often enough to present a significantly elevated risk, but not often enough to keep the supervisor alert and engaged. This is a well understood risk (e.g., [6]) that must be addressed in a road testing safety case. Figure 2 illustrates this effect with hypothetical performance data that results in an overall test platform safety value in accordance with Eqn. 1. A hypothetical lower supervisability curve results in a region in which the vehicle is less safe than a conventional vehicle driven by a human driver. Safe testing requires a comparatively higher supervisability curve to ensure that the overall test platform safety is sufficiently high, as shown by the lower half of Figure 2.

Because autonomy capabilities are generally expected to mature over time, the safety argument must be revisited periodically during test and development campaigns as the autonomy failure rate decreases from region 2 to 3 above. An intuitive – but dangerously incorrect – approach would be to assume that the requirements for test supervision can be relaxed as autonomy becomes more mature. Rather, it seems likely that the rigor of ensuring supervisors are vigilant and continually trained to maintain their ability to react effectively needs to be *increased* as autonomy technology transitions from immature to moderately mature. This effect only diminishes when the AV technology starts approximating the road safety of a conventional human driver all on its own (regions 4 & 5).

Pre-Test Validation (G41, G42, G43, G44, G4x)

Since there is always some risk in conducting tests on public roads, it is prudent to maximize validation via simulation, closed course testing, and other methods. It should be noted that data collection on public roads is fundamentally different than testing on public roads in that autonomy does not have control authority over the vehicle for data collection.

Robust fault injection campaigns could, and should, be used to characterize the likely consequences of autonomy failures that will inevitably happen during public road testing. Additionally, field data should be collected to determine the type, frequency, and severity of autonomy failures, even if those failures do not result in a mishap.

Implications of the Reference Safety Argument

Implementing a test program in accordance with a safety argument of this type brings with it a need to consider some essential implications and tradeoffs.

Response to Field Data

Even if care is taken to address all factors in the safety argument, there will inevitably be uncertainties, underestimated residual risks, argumentation gaps and other surprises. This is especially true given the inherent non-linearity of the supervisability problem.

It is essential in a robust safety case to properly account for the fact that all the elements of a test platform will have imperfections, including both the autonomy algorithms and the human supervisors. Testing teams should expect to encounter incidents of autonomy equipment failure, autonomy algorithm failure, unexpected environmental factors, unexpected (and even illegal) behavior by other actors, lapses in supervisor attention, errors in good faith supervisor judgement, potential badly behaving supervisors, and even data collection faults. In other words, inevitable failures in the testing process must also be considered as contributing factors to degradation of supervisor and equipment capabilities when evaluating risk.

A mature safety culture does not discount issues, adverse events, and surprises as one-off events. Nor does it find a scapegoat and dismiss that person, declaring all problems with the test program to be fixed by that personnel action. Rather, ensuring and improving safety requires considering every incident, mishap, and near miss as a failure in the testing program safety process. It is crucial to identify and fix the root cause of all safety problems beyond addressing any superficial symptoms.

Autonomy Behavior Permissiveness

Not all legal, safe vehicle behaviors are supervisable. There is an essential tension between testing safety and autonomy behavior permissiveness that ultimately seems likely to limit the effectiveness of any on-road testing program. This is related to the topics G23 mental model of autonomy and G31 situational awareness, as well as the timeliness of supervisor responses in general.

As an illustrative example, consider an AV test platform approaching a red traffic light. It might be that the AV is tuned to perform aggressive, last-second stops so that it can continue at-speed through the intersection in case the traffic light is about to turn green. (Indeed, vehicle-to-infrastructure data transmissions might inform the autonomy that the light will change just before the vehicle reaches the intersection, enabling it to proceed without loss of speed.) The human will experience a vehicle that repeatedly stops at the last second when approaching a red light, or maintains speed when it is sure that the light will change just in time, and learn that this is normal behavior.

However, a fault in data transmission, control logic, map data error, or a machine learning functional defect might result in the vehicle running a red light. Since the human has been trained that a last-second stop is normal, it will take extra time for the human to diagnose and react to a failure to stop right before the intersection, potentially resulting in the AV running the light and entering cross traffic flow. Consider that if the vehicle regularly maintains speed knowing that a traffic light will change to green at the last second, it will be impossible for a supervisor to react in time if the light does not change as expected as the vehicle reaches the intersection. In other words, it is unreasonable to expect a supervisor to fully mitigate a malfunction in a behavior that, by design, occurs at the last possible moment to avoid an incident or loss event.

Avoiding these types of incidents might require making the AV less aggressive to improve supervisability by giving the supervisor more time to detect and react to behavioral failures. But, that necessarily comes at the cost of reduced efficiency and behavioral permissiveness. Thus, it seems likely that there are inevitable limits to the aggressiveness of an AV compared to a human driver if a supervisor is being used to ensure safety. The vehicle must leave time for the supervisor to react, necessarily reducing its ability to operate right at the limit of safe behavior for the sake of efficiency.

These limits might be improved to a degree by selecting supervisors who have better-than-average skills and response times. (Such a claim of extraordinary supervisor abilities should be supported by outcome-based evidence, and likely requires constant refresher training, proficiency drills, and performance monitoring.) Additionally, internal AV state information might be supplied to provide improved expectations as to whether the AV is about to make a mistake. But, ultimately, there will be limits to what can be accomplished. It seems likely that ensuring the safety of an AV under test that has human (or beyond human) driving capability will be difficult to achieve via reliance upon a human supervisor.

It might be possible to adjust acceptable autonomy permissiveness via an analysis that parallels ISO 26262 ASIL analysis. (This is not actually conformance to ISO 26262 per se.) Such an approach would involve considering the severity, exposure, and controllability of potential autonomy malfunctions in assessing risk. For the analysis presented herein we have implicitly assumed that severity has the same profile across various malfunctions as for human driver errors, exposure is high with respect to a particular testing ODD, and controllability is high (the system should only be testing in situations for which the supervisor can assure safety via exercising control). One way of considering this subsection is that it explores issues that arise when autonomy behavior compromises supervisor controllability.

Relying Upon ADAS Systems

Some draft safety arguments for road testing that we have seen propose to take credit for off-the-shelf ADAS systems. Doing so is often problematic and can be prone to abuse.

For example, consider an Automatic Emergency Braking (AEB) system. A naïve safety argument is that if AEB is installed in a vehicle, there is no need for a safety driver to have full alertness, because the vehicle will brake itself before hitting anything.

Such an argument has a host of defects. For example, a typical AEB system is only intended to work in certain scenarios. It might not alarm on obstacles that have zero in-line velocity to avoid false alarms from overhead road signs. It is also might not work for problematic scenarios such as negative obstacles (e.g., an uncovered work zone pit), radar-absorptive materials (e.g., detached fabric core tire treads), or low profile critical obstacles (e.g., a person who has fallen and lies prone onto the road surface). For that matter, it might not even be designed to avoid collisions with pedestrians at all. Other ADAS systems have their own limitations.

A significant issue is that to provide value as an ADAS system, an off-the-shelf AEB system need only work most of the time, not all of the time. The argument is that the human has already made a mistake or otherwise been put in an unrecoverable situation before AEB is activated, so anything AEB can do to help is a benefit. In fact, it is expected that the AEB will be tuned to have a low false alarm rate to avoid over-riding human intent unless there is very high probability that the human is making a mistake, even if that means some missed activations. This approach can provide significant value for a last-ditch defense to mitigate the consequence of collisions due to presumably infrequent error of the responsible human driver.

However, a typical ADAS tuning approach is fundamentally incompatible with an argument that AEB will aggressively protect a potentially error-prone immature autonomy system. While for human drivers the AEB should likely be tuned to have a low false-alarm rate, instead for an autonomy testing scenario the AEB should be tuned to tolerate a higher false-alarm rate in exchange for a lower rate of missed activations.

While using off-the-shelf ADAS systems as defense-in-depth mechanisms has some merit, ultimately the safety argument made when designing the ADAS systems was that a human driver ultimately will take responsibility for vehicle safety. Any safety argument that inverts this by replacing a human supervisor responsibility with the ADAS system is therefore inherently suspect, and is likely deficient.

Semi-Autonomous Production Systems

Many of the arguments used here for test platforms will also apply to production semi-autonomous vehicles. For example, the safety argument based on failure frequency, driver response, and appropriate driver mitigation actions still holds in general. So too do the issues regarding the nonlinear interactions between autonomy failure rate and driver attentiveness. We believe that the safety case discussed is a suitable starting point for a semi-autonomous production safety case. However, significant adjustments will need to be made.

It seems likely that the AV testing problem is in many ways easier than the semi-autonomous production deployment problem. That is because for AV testing the scale is smaller, so it can be economically feasible to take extraordinary steps to improve driver abilities, including approaches such as using multiple supervisors, relying upon extremely high supervisor skill levels, requiring frequent rest breaks, using overlay sensor systems beyond normal vehicle equipment for sensor diversity, and using labor-intensive monitoring of system elements such as independent reviews of supervisor attentiveness.

We make the assumption for this analysis that AV testing is conducted by highly qualified, mature professionals who are making every effort to ensure safety and will shut down testing operations rather than compromise safety. A safety case for partially autonomous production systems will need to additionally account for unqualified supervisors, skill loss over time [20], intentional misuse, evasion of driver monitoring mechanisms, lax maintenance procedures, attempted operation in invalid ODDs, and so on.

Completing the Safety Argument

The safety argumentation structure described in Fig. 1 is a starting point, and needs to be elaborated to create a complete safety case. Because there is not yet an industry standard way for designing and road testing autonomous vehicles, the specifics of the safety case will vary among design teams. Design teams should explore available tutorial information on GSN (e.g., [14]) for more information on how to complete the argumentation structure.

A related issue is ensuring that the safety case is complete and sufficiently broad in scope. Other than a requirement that all the points in this paper must be addressed as a minimum, determining completeness will depend upon the system design, ODDs, and other system-specific factors.

Conclusions

A safety argument for human-supervised on-road testing of autonomous vehicles should include sub-arguments that the supervisor will respond in a timely manner, that the supervisor's response will be adequate to mitigate autonomy failures, and that the autonomy fails with an appropriate profile given supervisor capabilities. A significant challenge in successfully providing evidence to support such an argument will be the supervisability problem. Coupling between autonomy failure rate decreases and degradation of supervisor performance will make it difficult for human supervisors to remain attentive when autonomy fails with low frequency, requiring specific mitigation to avoid unacceptable risk.

Because a number of difficult and subtle human performance topics must be addressed in a credible safety argument of this type, it will be essential that field data is collected and continually analyzed to ensure that an autonomy test program achieves its safety objectives. It will be crucial to keep in mind that as the autonomy capabilities start to mature, safe road testing will actually *increase* the performance demands placed upon human supervisors to remain vigilant and effective.

The reference safety case presented is based on lessons learned across multiple autonomous ground vehicle projects over the last several years. Additionally, it is based on discussions with regulators and other industry stakeholders. While each on-road testing program might have its own unique requirements, we consider this to be a solid starting point for ensuring the safety of on-road test programs for autonomous vehicles. We hope that this material can provide the beginnings of a standardized approach to creating transparent, independently assessed safety cases for on-road testing safety.

References

[1] Jurisdictional Guidelines for the Safe Testing and Deployment of Highly Automated Vehicles, AAMVA, May 2018

[2] Banerjee, Jha, Cyriac, Kalbarczyk and Iyer, "Hands off the wheel in autonomous vehicles? A systems perspective over a million miles of field data," DSN 2018.

[3] Boudette, "Autopilot cited in death of Chinese Tesla Driver," NY Times, Sept. 14, 2016.

[4] Burton, S., "Making the case for safety of machine learning in highly automated driving," SAFECOMP, Sept. 2017, pp. 5-16.

[5] Testing of Autonomous Vehicles with a Driver, California Department of Motor Vehicles, 2018. https://goo.gl/vCzNJu

[6] Casner, S., Hutchins, E., & Norman, D., The Challenges of Partially Automated Driving, Comm. ACM, May 2016, pp. 70-77.

[7] China.com "Tesla confirms 'autopilot' engaged in fatal crash in China," China.com, Feb. 28, 2018, http://english.china.com/news/china/54/20180228/1217483.html

[8] Dragan, A., Lee, K. & Srinivasa, S., "Legibility and predictability of robot motion," Human-Robot Interaction (HRI) 2013, pp. 301-308.

[9] Efrati, A., "Waymo collision shows flaws in self-driving car tests," The Information, Oct 2, 2018.

[10] Eykholt et al., "Robust physical-world attacks on deep learning visual classification," https://arxiv.org/pdf/1707.08945.pdf

[11] Gauerhof, Munk & Burton, "Structuring validation targets of a machine learning function applied to automated driving," SAFECOMP 2018, pp. 45-58, 2018.

[12] Gertman, Blackman, Human Reliability and Safety Analysis Data Handbook, New York, John Wiley Interscience, 1994.

[13] GSN Community Standard Version 2, January 2018.

[14] GSN Working Group, http://www.goalstructuringnotation.info/, accessed Dec. 28, 2018.

[15] ISO, ISO 26262-1:2011(en) Road vehicles — Functional safety. International Standardization Organization, 2011.

[16] Koopman, P., Ensuring the safety of on-road self-driving car testing (presentation), PA AV Summit, April 9, 2018.

[17] Krafcik, J., “The very human challenge of safe driving,” Nov. 5, 2018, https://medium.com/waymo/the-very-human-challenge-of-safe-driving-58c4d2b4e8ee

[18] Krisher, Billeaud, "Police: Backup driver in fatal Uber crash was distracted," AP, June 22, 2018.

[19] Lingeman, "Watch Mobileye's self-driving prototype run a red light while news cameras are rolling," Autoweek, May 26, 2018.

[20] Marchau, Heijden, "Policy aspects of driver support systems implementation: results of an international Delphi study," Transport Policy 5 (1998), pp. 249-258.

[21] Meyhofer, E., Self-driving cars return to Pittsburgh roads in manual mode, Medium, July 24, 2018.

[22] NHTSA, Automated Driving Systems: a vision for safety, US Dept. of Transportation, DOT HS 812 442, Sept. 2017.

[23] NHTSA, Automated Vehicles 3.0: Preparing for the Future of Transportation, US Dept. of Transportation, Oct. 2018.

[24] NTSB Identification: SEA08IA080 (aviation incident report)

[25] NTSB, Safety Recommendation (A-09-61 through -66), August 7, 2009

[26] NTSB, "Collision between a car operating with automatic vehicle control systems and a tractor-semitrailer truck, Willison FL, May 7, 2016," NSTB/HAR-17-XX, Sept. 12, 2017

[27] NTSB, Preliminary Report Highway HWY18MH010 (Uber Technologies, Tempe Arizona)

[28] Palin, R. & Habli, I., “Assurance of automotive safety – a safety case approach,” SAFECOMP 2010, LNCS 6351, pp. 82-96.

[29] PennDOT, Automated Vehicle Testing Guidance, Pennsylvania Department of Transportation, July 23, 2018.

[30] Pezzementi, Z., Tabor, T., Yim, S., Chang, J., Drozd, B., Guttendorf, D., Wagner, M., & Koopman, P., "Putting image manipulations in context: robustness testing for safe perception," IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), Aug. 2018.

[31] SAE, Automated Driving (from SAE J3016), http://www.sae.org/misc/pdfs/automated_driving.pdf accessed 10/13/2017.

[32] SAE, Guidelines for Safe On-Road Testing of SAE Level 3, 4, and 5 Prototype Automated Driving Systems (ADS) J3018_201503.

[33] UberATG, A principled approach to safety, Medium, Nov 2, 2018.

[34] Victoria Guidelines for Trials of Automated Vehicles, Victoria Government Gazette, No. S 421, 12 Sept. 2018.

[35] Road Safety (Automated Vehicles) Regulations 2018, S.R. No. 120/2018, Victoria Australia, 4 Sept. 2018.

[36] Villemeur, Reliability, availability, maintainability and safety assessment: Assessment hardware software and human factors, 1992.

[37] Wardzinski, A., "Safety assurance strategies for autonomous vehicles," SAFECOMP 2008, pp. 277-290, 2008.

[38] Young, M. S. & Stanton, N. A. (1997). Automotive automation: Investigating the impact on drivers’ mental workload. International Journal of Cognitive Ergonomics, 1(4), 325-336

Contact Information

Dr. Philip Koopman is an Associate Professor of Electrical and Computer Engineering at Carnegie Mellon University, where he specializes in autonomous vehicle safety, software safety and dependable system design. He is also CTO and co-founder of Edge Case Research. He has affiliations with the Carnegie Mellon University Robotics Institute, National Robotics Engineering Center (NREC) and the Institute for Software Research.
E-mail: koopman@cmu.edu

Dr. Beth Osyk is a Lead Engineer at Edge Case Research, which specializes in software robustness testing and high-quality software for autonomous vehicles, robots, and embedded systems.
E-mail: bosyk@edge-case-research.com

Definitions/Abbreviations

ADAS	Advanced Driver Assistance System
AEB	Automatic Emergency Braking
AV	Autonomous Vehicle
GSN	Goal Structuring Notation
ODD	Operational Design Domain