A Safety Standard Approach for Fully Autonomous Vehicles

Philip Koopman[1,2], Uma Ferrell[3], Frank Fratrik[1], Michael Wagner[1]

[1] Edge Case Research, Pittsburgh PA 15201, USA
[2] Carnegie Mellon University, Pittsburgh PA 15213, USA
[3] The MITRE Corporation, McLean VA, 22102, USA
koopman@cmu.edu,uferrell@mitre.org,ffratrik@ecr.ai,mwagner@ecr.ai

Abstract. Assuring the safety of self-driving cars and other fully autonomous vehicles presents significant challenges to traditional software safety standards both in terms of content and approach. We propose a safety standard approach for fully autonomous vehicles based on setting scope requirements for an overarching safety case. A viable approach requires feedback paths to ensure that both the safety case and the standard itself co-evolve with the technology and accumulated experience. An external assessment process must be part of this approach to ensure lessons learned are captured, as well as to ensure transparency. This approach forms the underlying basis for the UL 4600 initial draft standard.

Keywords: Self-driving cars, autonomous vehicles, safety standard, UL 4600.

1 Introduction

Self-driving cars are (eventually) coming, and could have a profound impact on transportation [13]. On-road testing is underway in a number of locations across the world, and announcements regularly proclaim that cars will be able to operate without a human driver "soon" (or perhaps later [13]). Overall, safety looms as a significant concern.

Standards that address computer-based system safety for conventional vehicles have existed for decades [3][9], with ISO 26262:2018 [4] being a recent incarnation. A more recent standard addresses Advanced Driver Assistance System (ADAS) Safety Of The Intended Function (SOTIF) [5].

These existing standards are essential, but do not achieve comprehensive coverage of how to ensure that deployed fleets of Highly Autonomous Vehicles (HAVs) will operate safely. While safety standards from other domains such as aviation and military systems can provide additional insight, designers in those domains also struggle with issues unique to building safe autonomous systems. Additionally, HAV technology can benefit from an agile, iterative approach to ensuring and regulating safety [14].

This position paper seeks to outline a number of issues that must be addressed in a comprehensive HAV safety standard. The strategy described is the basis of the draft "UL 4600 Standard for Safety for the Evaluation of Autonomous Products" [16] that is intended to cover HAVs and eventually other related domains.

WAISE 2019 Preprint. Approved for public release.
Distribution unlimited. Case number 19-1358.

2 Current Standards

2.1 ISO 26262

Traditionally, automotive designers have based their overall safety strategy on a principle that a human driver is ultimately responsible for safety. This has resulted in, among other things, a focus in ISO 26262 [4] on functional safety.

Broadly speaking, functional safety ensures the system has a capability to mitigate failure risk sufficiently for identified hazards. The amount of mitigation required depends upon the severity of a potential loss event, operational exposure to hazards, and human driver controllability of the system when failure occurs. These factors combine into an Automotive Safety Integrity Level (ASIL) per a predetermined risk table. The assigned ASIL for a function determines which technical and process mitigations must be applied, including specified design and analysis tasks that must be performed. ISO 26262 is consistent with safety standards such as IEC 61508 [2] on items such as:

- Specifies a V-based process reference model
- Addresses software, hardware, and system aspects using integrity levels
- Includes lifecycle topics such as production, operation, support, and tools
- Specifies approach to safety incorporating hazards, safety goals, and ASILs
- Specifies analysis, design, and verification techniques based on ASIL

In summary, the emphasis on ISO 26262 is on avoiding design faults (e.g., via software quality requirements) and mitigating the effect of equipment faults during operation (e.g., via failsafes).

2.2 ISO/PAS 21448 (SOTIF)

More recently, the automotive industry has created a safety standard for driver assistance functions that could fail to operate properly even if no equipment fault is present. The ISO/PAS 21448 “Safety of the Intended Functionality” (SOTIF) standard [5] addresses those issues. It primarily considers mitigating risks due to unexpected operating conditions (the intended function might not always work in these due to limitations of sensors and algorithms) and gaps in requirements (lack of complete description about what the intended function actually is). Highlights of this standard include covering:

- Insufficient situational awareness
- Foreseeable misuse and human-machine interaction issues
- Issues arising from operational environment (weather, infrastructure, etc.)
- An emphasis on identifying and filling requirement gaps (removing “unknowns”)
- In practice, an emphasis on enumerating operational scenarios (e.g., [10])

In summary, ISO 21448 extends the scope of ISO 26262 to cover ADAS functionality. Both explicitly permit extending scope further. But as a pair they are not architected to cover the full extent of HAV safety. (A pending, not-yet-public revision of ISO 21448 aims go further.)

2.3 Other Safety Standards

There are numerous other safety standards from other domains including: IEC 61508 [2] for chemical process control; CENELEC EN 50128 [1] for rail systems; MIL-STD-882E [15] for military systems; and SAE ARP 4754A [11] as well as SAE ARP 4761 [12] for aviation. While these provide additional safety perspective, none covers the full range of HAV issues. Mainly this is due to assumptions of human operator availability (e.g., aircraft pilots), complete requirements identification, and/or significantly simplified operational environment compared to HAVs (e.g., protected rail right of way). While these standards provide valuable insight and principles, more is needed to provide thorough guidance for HAVs.

3 Constraints On Acceptable Standards

The need to have an HAV safety standard is urgent. Companies regularly promise to deploy HAVs to production without "safety drivers." While we could wait for the usual decade(s) of field experience for designs to converge before writing a safety standard, it is highly desirable to have a standard sooner rather than later.

Despite the excellent foundation provided by current standards, significant challenges await any would-be standard authors for HAVs. These include both the type and immaturity of the technology being used. However, they also include some profound implications of removing the human driver from the vehicle safety equation.

3.1 Novel Technology

HAVs as currently envisioned use technology that is inherently incompatible with legacy safety standards approaches. A standard must address at least [6]:

- Use of Machine Learning (ML) technology. A significant advantage of using ML is using a training-based approach to resolve intractable design situations. However, that same lack of requirements impedes traceability and ability to do design reviews.
- Use of unpredictable algorithms. Randomized algorithms and other so-called Artificial Intelligence (AI) techniques tend to behave in an unpredictable way, generally characterized as being non-deterministic. This complicates creating repeatable tests.

Traditional safety standards employ update cycles of perhaps 5 to 10 years, but HAV technology is evolving much more rapidly. Premature standards could inhibit innovation. Additionally, a traditional consensus-based standard approach is difficult when developers are still figuring out how to make the technology work acceptably well. Any standard will need an unprecedented level of flexibility to be viable.

3.2 No Human Driver

The contents of any standard will have to address fundamental changes in system-level fault management. Controllability evaporates with an HAV, because there is no human driver to exercise control. Therefore, autonomy itself must manage vehicle failures.

An additional issue with the removal of the human driver is that a large number of other operational and lifecycle activities beyond the actual driving must also be covered. This includes safely interacting with humans such as potentially unruly passengers and emergency personnel. Moreover, autonomy might need to mitigate risk due to operational faults (e.g., passenger evacuation in a car fire) and handle lifecycle faults.

4 A Safety Case Approach

We believe that the difficult constraints of creating a safety standard for HAVs can be met with an approach that combines: use of a safety case for the overarching structure, specifying breadth of safety case scope, incorporating lessons learned, updating for a changing environment, and using a multi-layered feedback approach that includes independent assessment. This approach accounts for not only managing the risk presented by unknowns, but also the evolving technology and changing operational environment.

4.1 Safety Cases with Specified Scope

Safety case approaches have been used previously (e.g., [4][8]). We believe that rather than being just a part of the safety package, the safety case should be the primary overarching structure containing essentially everything. This approach permits keeping items such as tools to be used and engineering processes to be followed flexible. By the same token, this means that the safety case must not only present fully substantiated arguments that appropriate and necessary processes and practices have been used, but also that the selection choices are in fact sufficient to ensure safety.

As long as the other elements of our approach are followed, in principle the standard need not specify any particular tool or process step approach. Rather, it can require that certain high-level claims and argumentation be present. As an example, the standard can require that all hazards and associated risks be identified, but not what techniques must be used to accomplish that. To avoid unnecessary effort and expense, credit can be taken for conformance to ISO 26262, ISO 21448, and other relevant standards to the degree conformance is credible and actually applies to HAV safety.

A potential concern is the creation of a safety case that is lacking in depth or evidence. The draft standard requires a certain level of depth by enumerating required subclaims and safety case coverage. (As an example, hazards associated with the supply chain must be identified.) At a high level, we have identified the following topics that must be specifically addressed for HAVs beyond the level of detail in other standards:

- Definition of Operational Design Domain (e.g., weather, scenarios [7])
- Machine learning faults (e.g. training data gaps, brittleness)
- External operational faults (e.g., other vehicles violating traffic rules)
- Faulty behavior by non-driver humans (e.g., pedestrians, lifecycle participants)
- Non-deterministic, variable system behavior (e.g., test planning, acceptance criteria)
- High residual unknowns (e.g., requirements gaps and post-deployment surprises)
- Lack of human oversight (e.g., operational fault handling, passenger handling)

- System-level safety metrics (e.g., use of leading and lagging metrics)
- Transitioning the system to degraded modes and minimum risk conditions

4.2 Ongoing Risk Assessment

Considering the novelty, complexity, and consequences involved with HAV deployment, challenges are expected in creating a bulletproof initial safety case. Rather than adopting a fiction that mere conformance to a standard at deployment results in flawless risk mitigation, instead it is important to continually evaluate and improve the residual risk present in the system. Identifying latent and emergent risks is essential to enable identifying, implementing, and verifying additional mitigation measures.

By the same token, it is important to address known safety issues before exposing testers and the public to undue safety risk. Developers should strive for a culture of responsible safety risk identification and ownership rather than simply checking boxes. This includes taking ownership of development mistakes as well as gaps in design, test, and the safety case itself. Honest self-assessment and iteration over the system development and deployment lifecycle is vitally important to mature the safety case.

We also believe that independent assessment is essential. This is especially true in light of the high-stakes, high pressure environment of HAV development. Beyond providing essential checks and balances on system safety, independent assessment can provide a way to share lessons learned without revealing proprietary design details.

4.3 Feedback and Lessons Learned

Rather than treat the rapid evolution of HAV technology as an obstacle, we intend to embrace it. Neither waiting until the dust settles (which might not ever really happen) nor prematurely freezing the standard seem viable. Instead, we plan to evolve the standard in tandem with the technology. Here is how we believe it can work (Fig. 1):

- Seed the initial standard with required essential practices and anti-patterns that have proven value (e.g., identifying hazards, avoiding known unsafe design patterns) based on stakeholder inputs.
- Require essential elements of the safety case (e.g., pick and adapt any reasonable hazard analysis approach from your favorite safety standard).

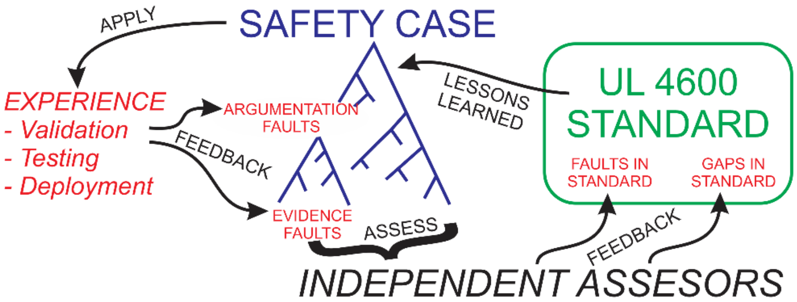

Fig. 1. The standard informs safety case construction. Field experience provides feedback.

- Include a list of safety case acceptable patterns and excluded anti-patterns.
- Require plausible argumentation that residual risk and "unknowns" will be tracked.
- Require feedback paths based on root cause analysis of incidents and loss events during both development and deployment to identify weak spots in the process:
  - Gaps in enumerated lists in the safety case (e.g., a new hazard)
  - Gaps in safety case evidence (e.g., an "impossible" failure occurs in the field)
  - Flaws in argumentation and assumptions (e.g., real world assumption violations)
  - Gaps in patterns, anti-patterns, and required elements in the safety standard itself
  - Adoption of new practices that have proven to provide value into the standard

5 Conclusions

We believe that a goal-based safety case approach with pre-seeded feedback paths is a practical way to create a safety standard for HAVs. This can encourage the use of accepted safety practices at first, yet still evolve and mature along with the industry. A potential outcome is an agile alternative to inflexible regulations for ensuring safety.

Disclaimer. We are subject matter experts working with UL to create an initial draft version of UL 4600 using this approach. The final standard may differ.

Acknowledgements. The authors wish to thank the UL 4600 drafting team participants from UL and Edge Case Research for their support and thoughtful comments.

References

1. CENELEC, "Railway applications - Communication, signaling and processing systems - Software for railway control and protection systems," EN 50128:2011.
2. IEC, "Functional safety of electrical/electronic/programmable electronic safety-related systems," IEC 61508:2010.
3. ISO, "Road Vehicles – Functional Safety" ISO 26262:2011.
4. ISO, "Road Vehicles – Functional Safety" ISO 26262:2018.
5. ISO, "Road Vehicles – Safety of the Intended Function" ISO/PAS 21448:2019.
6. Koopman, P. & Wagner, M., "Toward a framework for highly automated vehicle safety validation," SAE 2018-01-1071, 2018.
7. Koopman, P. & Fratrik, F. "How many operational design domains, objects, and events?" SafeAI 2019.
8. Ministry of Defence, "Safety Management Requirements for Defence Systems," Defence Standard 00-56, 2017.
9. MISRA, Development Guidelines for Vehicle Based Software, November 1994.
10. Pegasus Project, https://www.pegasusprojekt.de/en/home, accessed 4/21/2019.
11. SAE, Guidelines for Development of Civil Aircraft and Systems, ARP4754A, 2010.
12. SAE, Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems and Equipment, ARP4761, 2012.
13. US Dept. of Commerce, "The employment impact of Autonomous Vehicles," Aug. 2017.
14. US Dept. of Commerce, https://www.commerce.gov/issues/regulatory-reform, 7 June 2019.
15. US DoD, "Standard Practice: System Safety", MIL-STD-882E, 11-May-2012.
16. Yoshida, J., "UL Takes Autonomy Standards Plunge," EE Times, 4/16/2019.