Comments for Docket No. NHTSA-2020-0106

David Gelperin
david@clearspecs.com

This Notice focuses on ways the Agency could approach the performance evaluation of ADS through a safety framework, containing a variety of approaches and mechanisms that, together, would allow NHTSA to identify and manage safety risks related to ADS (*Automated Driving System*) in an appropriate manner. NHTSA anticipates focusing this framework on **the functions of an ADS that are most critical for safe operation**.

NHTSA seeks comments on how to select and design the structure and key elements of a (safety) framework.

The Agency also wishes to hear from the public on whether ADS-specific regulations are appropriate or necessary **prior to** the broad commercial deployment of the technology (page 11) – **Definitely**

Engineering measures are those aspects that can be **readily determined through the testing** of a finished motor vehicle or system and **establish the level of safety performance**. (page 20)

> **This definition is extremely out-of-date**. For motor vehicles controlled by sensors and software there are NO engineering measures satisfying this definition. Safe performance of such vehicles can't be *readily determined* by black-box testing. In fact, their safe performance can't be determined by black-box testing alone. [I hope, manufacturers understand this.]

"these four primary functions serve as the core elements NHTSA is considering" (page 9)

> Safety should be THE core quality attribute that NHTSA is considering.
>
> Safety in NOT a functional element. It is a **quality attribute** that crosscuts all functional elements. Mitigation functions or components may arise from requirements for safety hazard mitigation.
>
> One approach to framing safety is to begin by considering the question: How can an ADS kill i.e., what are the **hazards arising from use** of an ADS? This will result in a list of safety-critical ADS misbehaviors e.g., unintended acceleration. The initial list is NOT concerned with the likelihood or causes of the misbehaviors. It is only concerned with listing ADS misbehaviors that might lead to death or serious injury. More such hazards might be identified by asking: How can each functional component of an ADS kill?
>
> Death by carbon monoxide poisoning will not be on the hazard list, because the ADS would not contribute to this hazard.

The next step is to associate mitigations with each ADS misbehavior. Some mitigations will entail the need for mitigation functions or components. For example, if the engine dies and the vehicle is in motion, an “emergency motion stopping function” (see appendix) must decide between “emergency parking” and “emergency stopping”.

To promote safety in the current competitive environment, **the NHTSA should provide lists of core hazards both to and from an ADS with an unconstrained domain, along with their mitigation alternatives**.

The Agency anticipates that the safety framework would include both process and engineering measures to manage risks. (page 9)

Process and engineering measures **alone** are inadequate. Safety-critical **product measures** e.g., complexity, are also important.

The engineering measures should include **skeptical reviews and thorough (white-box) testing of risk mitigation functions and components** e.g., verification of an emergency motion stopping function.

NHTSA requests comment on which (other safety functions) the Agency should prioritize as it continues the research necessary to develop a safety framework. (page 24)

All mitigation functions arising from identification of potential ADS hazards.

NHTSA, as part of the Department’s broader efforts, has begun the research to explore potential ways the Agency can assess the safety of ADS. (page 24)

ADS safety assessment must include skeptical reviews/inspections of: (1) specified safety-critical ADS hazards, (2) specified mitigations of safety-critical ADS hazards, (3) software designs focusing on the mitigations of safety-critical ADS hazards, (4) results of thorough (i.e., white-box) testing of safety-critical mitigation functions and components.

An FMVSS might also require that the planning and control functions of an ADS be programmed to adhere to a **defensive driving model** (page 52)

This is a great idea, if there were such a verifiable, comprehensive, consensus model. [I couldn’t find one.] Maybe NHTSA should lead the development of one.

Appendix

An Emergency Motion Stopping Function (EMSF)

Automated Driving Systems (ADSs) **must have** *an emergency motion stopping function* or its equivalent mainly to mitigate detected software failures and detected misbehaviors e.g., driving through a series of red lights.

Along with misbehavior monitors, every component should check its own logic e.g., with assertions. Inputs, outputs, intermediate results, and invariants should be checked. If a self-check fails, the emergency motion stopping function should be invoked.

The emergency motion stopping function **must be invokable by passengers**. Passenger access should be protected from accidental invocation (similar to a manual fire alarm).

Because of its safety-critical nature and complexity, this motion stopping function should NOT be embedded in a more general control function. To avoid a single point of failure, the function must be housed on an isolated platform that is continually checked for liveness by the main control platform. In addition, the isolated platform must continually check the main control platform for liveness. A copy of the function must also exist on the main control platform.

The emergency motion stopping function would be invoked by the:

- main ADS controller when:
  - monitors detect misbehavior e.g., unintended acceleration
  - assertions fail
  - the isolated platform does not respond to a liveness check
- isolated copy when the main control platform does not respond to a liveness check
- passengers when they detect significant misbehavior

An emergency motion stopping function must first check a common memory indicator and then set it if it is not set or cease execution if it is set. This prevents both copies of the stopping function from executing when both liveness checks fail because communication between the platforms is lost or when passengers invoke the function when it is already executing.

The emergency motion stopping function should perform **emergency parking** if possible and safe, otherwise it should perform **emergency stopping**, unless stopping is too dangerous e.g., when immersed in heavy, fast-moving traffic and the trigger event carries unknown risk e.g., liveness check of the isolated platform fails. When neither emergency parking nor emergency stopping are performed, the function should wait until emergency parking is both possible and safe or emergency stopping is safe “enough”.

To enable a more accurate assessment of passenger risk, **passengers might indicate the criticality of their trip**. For example, when the criticality is high as when a passenger is being rushed to the hospital because of a heart attack and a system or software problem is detected, an ADS **might consult passengers** or decide to continue a trip even when a riskier trip may result.

Note the critical nature of the sensors and sensor processing logic. If either fails, emergency stopping might stop an ADS controlled vehicle on an unguarded railroad track.

Monitoring ADS behavior could be done by hardware, software, and passengers, as well as other vehicles, and remote observers. I think it's too early to rule out any form of monitoring.

I suggest that monitoring be tied to hazard management. An inventory of hazards from an ADS could be developed and then mitigations for some would start with their detection by monitoring or self-checking.

I believe (1) detailed lists of hazards both to an ADS e.g., icy roadways, and from an ADS e.g., UA, along with their corresponding mitigations are essential to achieving safety and (2) detailed demonstration tactics mapped to the identified mitigations are essential to safety verification.

Assuming the existence of a "black box" and sensible information capture, events triggering an emergency motion stopping invocation should be determinable.

EMSF Glossary

Glide path – safe path from current location to "much" safer location with steering, but no power

Powered path– safe path from current location to "much" safer location with steering and power

Active roadway – part of roadway where traffic flows

Emergency parking – stopping off the active roadway with hazard lights and emergency communication

Emergency stopping – stopping on the active roadway with hazard lights and emergency communication

Emergency planning – in an emergency, seeks to identify a glide or powered path

Emergency motion stopping function – entails both emergency parking and emergency stopping and their prerequisites

Due to the complexity of the ADS control tasks and the typical denial of software problems by some automobile companies (e.g., https://www.consumeraffairs.com/news/florida-police-force-bmw-to-a-stop-after-driver-claims-gas-pedal-got-stuck-021618.html), the need for such a function is obvious. Its need does not require some threshold number of deaths for justification. Similar motion stopping functions are described in BSI PAS 1880: 2020.

Failure to **require** this safety function or its equivalent, means the NHTSA will be partially responsible for all deaths resulting from its absence.