AN OPEN, TRANSPARENT, INDUSTRY-DRIVEN APPROACH TO SAFETY

Jack Weast
Sr. Principal Engineer, Intel
VP Automated Vehicle Standards, Mobileye
An Automated Vehicle must be

SAFE

USEFUL

TRANSPARENT

SCALABLE
3 FACETS OF AUTOMATED DRIVING

**SENSE**
- Perception of the complete environment
- The raw material

**PLAN**
- Decision-making
- Analyze the raw material, and what action to take

**ACT**
- Execute the plan
- Control acceleration, braking, steering
How would you demonstrate that an automated vehicle is safe?
ISO 26262 guides electric, electronic, and software quality

• Reduce chance of system faults, mitigate those that do occur

• Essential, but not the full picture
NORMATIVE SAFETY STANDARDS

Algorithm-level safety

• Process to identify classes of safety violations not covered by ISO 26262
• Open to interpretation, which would result in different definitions of “safety”
What does “safety” mean for an autonomous vehicle and how can we define it in a way that is satisfactory to society?
HOW WOULD YOU DEFINE “SAFETY” FOR AN AV?

First try

Self-driving cars should be statistically better than a human driver
The more miles I drive, the safer I am

Probability $\rho$ of fatality / 1 hour of driving in U.S.

To demonstrate $\rho$ an AV must drive

Averaging 30mph, that amounts to

$10^{-6}$

$\frac{1}{\rho}$ hours

$\sim 30m$ miles

To build trust, we need to be better by 2-3 orders of magnitude

Not Safe

THE STATISTICAL APPROACH TO SAFETY

The more miles I drive, the safer I am

For society to accept AVs, $\rho$ should be

Averaging 30mph, that amounts to

100 cars driving 24/7/365 would take

$10^{-9}$

$\sim 30b$ miles

Over a millennium

Not Safe

Not just once: Every update of hardware & software

Not Scalable

The more miles I drive without a crash, the safer I am

Miles driven here

Not the same as here
DISENGAGEMENTS

Minimize the number of times the ADS fails and requires a takeover

Why it's insufficient

• Similar to miles driven, depends on where & when
• Incentive to avoid the tough environments likely to trigger disengagements
HOW WOULD YOU DEFINE “SAFETY” FOR AN AV?

Second try
Develop other machine-friendly methods to define and prove safety
OTHER METHODS: SIMULATION

Why simulation alone cannot fully validate planning

- While sensing validation thrives in simulation, planning faces limitations.
- Driving is a multi-agent system, to simulate it accurately is to simulate human behavior.

WE CANNOT PROVABLY ACCURATELY SIMULATE THE REAL WORLD
OTHER METHODS: SCENARIOS

Expose the AV to the complete set of driving scenarios

Why it's insufficient

• Have to generalize; my list covers any other similar but omitted scenarios
• Difficult to draw the appropriate line between abstract & concrete scenarios
• Incents industry to build to the test

Pre vs. Post Deployment

• Pre-deployment testing assumes that it’s possible to test everything
• And that nothing new will come up post-deployment
OTHER METHODS: PROPRIETARY

Trust me!

THE BLACK BOX
OF AI DECISION-MAKING

...PAY NO ATTENTION TO THAT MAN BEHIND THE CURTAIN
HOW WOULD YOU DEFINE “SAFETY” FOR AN AV?

Third try
The AV only needs to strictly obey the rules of the road
SHOULD THE AV “FOLLOW THE RULES OF THE ROAD”? 

- Traffic light
- Right of way
HOW WOULD YOU DEFINE “SAFETY” FOR AN AV?

Fourth try

Avoid accidents at all costs
THE AV MUST AVOID ACCIDENTS AT ALL COSTS
THE AV MUST AVOID ACCIDENTS AT ALL COSTS

https://www.youtube.com/watch?v=ctoBivu2NSE
THE AV MUST AVOID ACCIDENTS AT ALL COSTS
THE AV MUST AVOID ACCIDENTS AT ALL COSTS
WE NEED SOMETHING BETTER
And we’re not the only ones who think so

ACADEMIA

“Specify unsafe regions for safety, specify safe regions for functionality. A ‘safety envelope’”
– Prof. Philip Koopman, CMU

THINK TANKS

“There is currently no accepted, industry-wide approach to [safety] demonstration”
– Measuring Automated Vehicle Safety, RAND Corporation

GOVERNMENT

“The metrics that are most widely used by self-driving car developers -- miles driven and the frequency of human intervention -- alone are insufficient to demonstrate the safety of an autonomous automobile.”
– Derek Kan, Undersecretary of Transportation for Policy

---

HOW DO HUMANS DO IT?
An AV should, at all times, drive carefully enough so it will never be the cause of an accident, and drive cautiously enough such that it should be able to compensate for reasonable mistakes of others.
RESPONSIBILITY SENSITIVE SAFETY (RSS)

An open and transparent industry standard that provides a verifiable safety check for AV decision-making.

**FORMALIZE**
Human notions of safe driving

**IDENTIFY**
A Dangerous Situation

**EXECUTE**
The Appropriate Response

Keep a safe distance longitudinally & laterally

Safe distance compromised in both directions

Brake to restore safe longitudinal distance
RSS is:

• A mathematical model that formalizes a “common sense” interpretation of safe driving
  • What is a Dangerous Situation?
  • What is the proper response to a Dangerous Situation?
  • What does it mean to be reasonably cautious?
  • What assumptions can the AV make about the behavior of others?
WHERE DOES RSS FIT?

SENSE
- Analyze the complete environment
- Consider actions
- Make a decision

PLAN
- Analyze the raw material
- Consider actions
- Make a decision

ACT
- Execute the plan
- Control acceleration, braking, steering

RSS IS A CHECK FOR PLANNING SAFETY

PLANNING
- Planning gets you from point A to point B

RSS
- RSS helps keep you safe along the way
RSS can be used in any mechanism for verification.
BALANCING SAFETY AND USEFULNESS
As a result, our driving becomes useless, and arguably less safe. When merging with a busy highway, driving too cautiously can cause problems. If we don't take assertive, but reasonable action, we may never complete the merge.
THE BALANCING ACT BETWEEN SAFE & USEFUL

We have a tight window, but we have a reasonable expectation that car behind us will adjust 

Brakes to keep safe distance Before continuing
Safe action space: the set of all possible actions the AV can take that are safe

Ideally: the AVs driving policy aligns and can propose any action within that space

How to maximize the safe actions available to the driving policy
Driving policies learn with a Reward Function

Motives/weights dictate what kind of driving experience the AV produces

Without incorporating safety, some proposed actions will fall outside our safe action space
What if we add safety to the Reward Function?

- Adding safety to the Reward Function constrains the safe action space
- Safety now a competing interest in decision-making
- Now policy is overly-conservative, and still potentially unsafe
How (or whether) an AV gets from point A to point B should be a proprietary differentiator.

Safety cannot be left to proprietary chance.

- Safety should be an open, transparent industry standard.
- Time to Destination
- Efficiency
- Comfort

SAFE ACTION SPACE

SAFE ACTION SPACE

SAFE ACTION SPACE

SAFE ACTION SPACE

SAFE ACTION SPACE

SAFE ACTION SPACE
• Decouple safety from decision-making
• RSS becomes safety-check layer between driving policy and actuation
• RSS acts as the filter that defines safety
SAFE ACTION SPACE

RSS is our missing layer

- Decouple safety from decision-making
- RSS becomes a standard safety-check layer between proprietary driving policy and actuation
- RSS acts as the filter that defines safety for the industry
BASIC PRINCIPLES OF A SAFE AV

1. Keep a safe distance from the car in front of you
2. Leave time and space for others in lateral maneuvers
3. Exhibit caution in occluded areas
4. Right-of-Way is given, not taken
5. If you can safely avoid an accident without causing another you must do so
Define Safe Longitudinal Distance

\[ d_{\text{min}} = \left[ v_r \rho + \frac{1}{2} \alpha_{\text{max}} \rho^2 + \frac{(v_r + \rho \alpha_{\text{max}})^2}{2 \beta_{\text{min}}} - \frac{v_f^2}{2 \beta_{\text{max}}} \right]_+ \]
\[ d_{\text{min}} = \left [ v_r \rho + \frac{1}{2} \alpha_{\text{max}} \rho^2 + \frac{(v_r + \rho \alpha_{\text{max}})^2}{2 \beta_{\text{min}}} - \frac{v_f^2}{2 \beta_{\text{max}}} \right ]^+ \]

- \( v_r \): Rear car (\( c_r \)) velocity
- \( v_f \): Front car (\( c_f \)) velocity
Define Safe Longitudinal Distance

\[ d_{\text{min}} = \left[ v_r \rho + \frac{1}{2} \alpha_{\text{max}} \rho^2 + \frac{(v_r + \rho \alpha_{\text{max}})^2}{2 \beta_{\text{min}}} - \frac{v_f^2}{2 \beta_{\text{max}}} \right]^+ \]

- \( \rho \): Vehicle response time
- \( \beta_{\text{min}} \): Min braking for \( c_r \) to apply to avoid colliding with \( c_f \)
Define Safe Longitudinal Distance

\[
d_{\min} = \left[ v_r \rho + \frac{1}{2} \alpha_{\max} \rho^2 + \frac{(v_r + \rho \alpha_{\max})^2}{2 \beta_{\min}} - \frac{v_f^2}{2 \beta_{\max}} \right]_+
\]

\(d_{\min}\)  \(c_r, d_{\min}, c_f\)

\(\alpha_{\max}\)  Max acceleration during response time (for \(c_r\))

\(\beta_{\max}\)  Max braking applied by \(c_f\)

not physical limits, but upper bounds on reasonable behavior
SAFE LONGITUDINAL DISTANCE (OPPOSITE DIRECTIONS)

\[ d_{\text{min}} = \left( \frac{\nu_1 + \nu_{1,\rho}}{2} \right) \rho + \frac{\nu_{1,\rho}^2}{2\beta_{1,\text{min}}} + \left( \frac{|\nu_2| + \nu_{2,\rho}}{2} \right) \rho + \frac{\nu_{2,\rho}^2}{2\beta_{2,\text{min}}} \]

\( c_1 \) traveling with velocity \( \nu_1 \),  
\( \nu_1 \geq 0 \)

\( c_2 \) traveling with velocity \( \nu_2 \),  
\( \nu_2 < 0 \)
SAFE LONGITUDINAL DISTANCE (OPPOSITE DIRECTIONS)

\[
d_{min} = \left(\frac{v_1 + v_{1,\rho}}{2}\right)\rho + \frac{v_{1,\rho}^2}{2\beta_{1,min}} + \left(\frac{|v_2| + v_{2,\rho}}{2}\right)\rho + \frac{v_{2,\rho}^2}{2\beta_{2,min}}
\]

\[
v_{1,\rho} = v_1 + \rho \alpha_{max}
\]

\[
v_{2,\rho} = |v_2| + \rho \alpha_{max}
\]

Change in velocity during response time \(\rho\)
The silver car has reached the Danger Threshold
($t_d$ is the last safe time before we enter a dangerous situation)
Though the silver car initiated the dangerous situation, the blue car still ought to brake to return to a safe distance.
PROPER RESPONSE - OPPOSITE DIRECTION

If traveling in opposite directions, both cars must apply the brakes to a full stop.

\[ \beta_{1,\text{min}} \quad \mathrel{\Longleftarrow} \quad \mathrel{\longrightarrow} \quad \beta_{2,\text{min}} \]
BASIC PRINCIPLES OF A SAFE AV

Rules we formalize in RSS

1. Keep a safe distance from the car in front of you
2. Leave time and space for others in lateral maneuvers
3. Exhibit caution in occluded areas
4. Right-of-Way is given, not taken
5. If you can safely avoid an accident without causing another you must do so
Cars usually perform small lateral movements, Driving perfectly straight is impossible
Define Safe Lateral Distance

\[ d_{\text{min}} = \mu + \left[ \left( \frac{v_1 + v_{1,\rho}}{2} \right) \rho + \frac{v_{1,\rho}^2}{2\beta_{1,\text{lat},\text{min}}} \right] - \left[ \left( \frac{v_2 + v_{2,\rho}}{2} \right) \rho + \frac{v_{2,\rho}^2}{2\beta_{2,\text{lat},\text{min}}} \right] \]

Given car’s lateral position, \( l \) is the lateral location at time \( t \)

\( \mu \) represents our current lateral velocity
If \( t \in [t_d, t_d + \rho) \)
Both cars must limit lateral acceleration

\[ |\alpha| \leq \alpha_{\text{lat, max}} \]
PROPER RESPONSE - LATERAL DANGER

If $t \geq t_d + \rho$
Both cars must react

$\alpha_{lat, min}$

$-\alpha_{lat, min}$
Time $t$ is dangerous for cars $c_1, c_2$ if both longitudinal and lateral distances between them are non safe.
Given a dangerous time $t$, its Danger Threshold, $t_d$, is the earliest non-dangerous time such that all times in the interval $(t_d, t]$ are dangerous.
BASIC PRINCIPLES OF A SAFE AV

Rules we formalize in RSS

1. Keep a safe distance from the car in front of you
2. Leave time and space for others in lateral maneuvers
3. Exhibit caution in occluded areas
4. Right-of-Way is given, not taken
5. If you can safely avoid an accident without causing another you must do so
When sensing capabilities are physically limited, we must exhibit caution.
Both cars assume a reasonable limit on the speed of the other.

What is a reasonable assumption on the speed limit of the other?
BASIC PRINCIPLES OF A SAFE AV

Rules we formalize in RSS

1. Keep a safe distance from the car in front of you
2. Leave time and space for others in lateral maneuvers
3. Exhibit caution in occluded areas
4. Right-of-Way is given, not taken
5. If you can safely avoid an accident without causing another you must do so
How do we establish priority on roads with odd geometries?

$r_2$ has more significant lateral velocity compared to $r_1$ so $r_1$ is the primary route.
GRAY AREAS - WORK TO DO

Why this needs to be an open and transparent discussion
What if the front vehicle brakes > max,brake?

Front vehicle brakes harder than $a_{\text{max,brake}}$ and causes collision.

Current proper response contains values that are blame-free but can lead to collision.
Discontinuities in Road Condition

In rain both cars brake softer than respective dry boundaries, but rear car’s braking generates much more displacement than front car’s braking behavior allows.
Consider this:
An object on the road we only detect after its too late, because the silver car changes lanes at the last moment

Should safe distance account for this worst-case scenario?
To keep a safe distance on a highway going ~65mph, a car would need more than 150 feet (~10 car lengths) to stop in time.

Society would likely agree this is unreasonable... so what can the AV assume about others?
Proportional Responsibility
In some places, like the US, it is not always binary
We made the Proper Response, but are not “responsibility free”

Should safe distance account for the potential actions of the rear car?
What are we doing

**INDUSTRY**
Engaging with customers, competitors and consortia to have an open dialogue on the safety assurance of AV’s

**GOVERNMENT / NGO’S**
Understanding government and NHO expectations on transparency and measurable verification of AV’s

**ACADEMIA**
RSS Research Centers at Universities in USA and PRC

**REAL WORLD**
Deploying RSS in our AV Fleet in some of the most challenging environments
On a Formal Model of Safe and Scalable Self-driving Cars

Shai Shalev-Shwartz, Shaked Shammah, Amnon Shashua

Mobileye, 2017

Abstract

In recent years, car makers and tech companies have been racing towards self driving cars. It seems that the main parameter in this race is who will have the first car on the road. The goal of this paper is to add to the equation two additional crucial parameters. The first is standardization of safety assurance — what are the minimal requirements that every self-driving car must satisfy, and how can we verify these requirements. The second parameter is scalability — engineering solutions that lead to unleashed costs will not scale to millions of cars, which will push interest in this field into a niche academic corner, and drive the entire field into a “winter of autonomous driving”. In the first part of the paper we propose a white-box, interpretable, mathematical model for safety assurance, which we call Responsibility-Sensitive Safety (RSS). In the second part we describe a design of a system that adheres to our safety assurance requirements and is scalable to millions of cars.

1 Introduction

The “Winter of AI” is commonly known as the decades long period of inactivity following the collapse of Artificial Intelligence (AI) research after the 1970s. The term “Winter of AI” refers to a period when AI research was at a standstill due to the lack of progress in the field. This period lasted for several decades and ended with the emergence of modern machine learning and deep learning techniques. The “Winter of AI” is now seen as a setback for the field of AI, but it also led to significant advances in other areas of computer science, such as natural language processing and computer vision. The “Winter of AI” also paved the way for the development of new AI techniques, such as deep learning, which have revolutionized the field of AI. The “Winter of AI” has also had a significant impact on the way AI research is conducted today. In the past, AI research was often focused on building large, complex systems that were difficult to understand and evaluate. Today, AI research is often focused on building small, interpretable models that can be easily understood and evaluated. This shift in focus has led to significant advances in the field of AI, and has also helped to increase the public’s understanding of AI and its potential.
RSS IN SUMMARY

An open and transparent industry standard that provides verifiable safety assurance for AV decision-making

• The industry must collaborate with governments and agree on what it means for an AV to drive safely

• RSS provides a starting point for a definition of what it means for an AV to drive safely

• RSS can be formally verified and so solves the statistical verification challenge with an open and measurable metric

• RSS is technology neutral compatible with any AV solution

Join us in this important effort to provide safety assurance for Automated Vehicles!