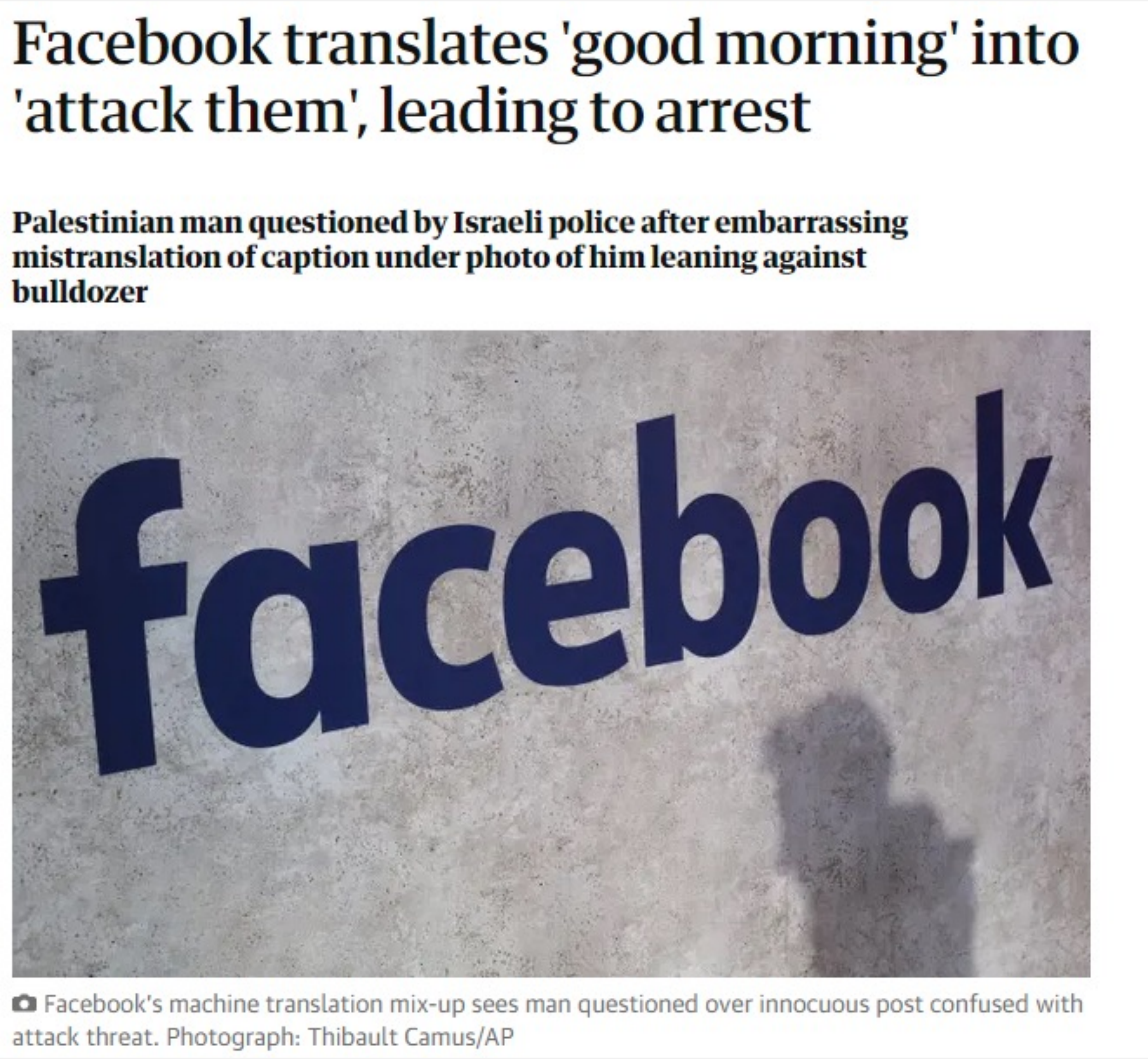
DESIGNING AND BUILDING SAFE INTELLIGENT SYSTEMS

Roman

SAFETY AND INTELLIGENT SYSTEMS- EXAMPLES

- Flagship example: self-driving cars causing accidents
- In practice, we are surrounded by systems we trust
- What are other examples of high impact safety failures in intelligent systems ?

- There are a lot of other examples!
- There is a lot of overlap in safety-related challenges between applications
- The challenges are highly multidimensional!



Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

Facebook's machine translation mix-up sees man questioned over innocuous post confused with attack threat. Photograph: Thibault Camus/AP

The Guardian headline of October 24,2017

AI FOR DRIVING: CURRENT LANDSCAPE

- In most cases, a human is responsible for supervising the AI
- Example: Waymo autonomous cars drove >6M miles in 2019-2020
- What difference does it make that a human is expected to supervise the AI?

- Because the human is responsible, the AI can make mistakes
- Expecting the AI to be responsible would massively increase safety requirements.

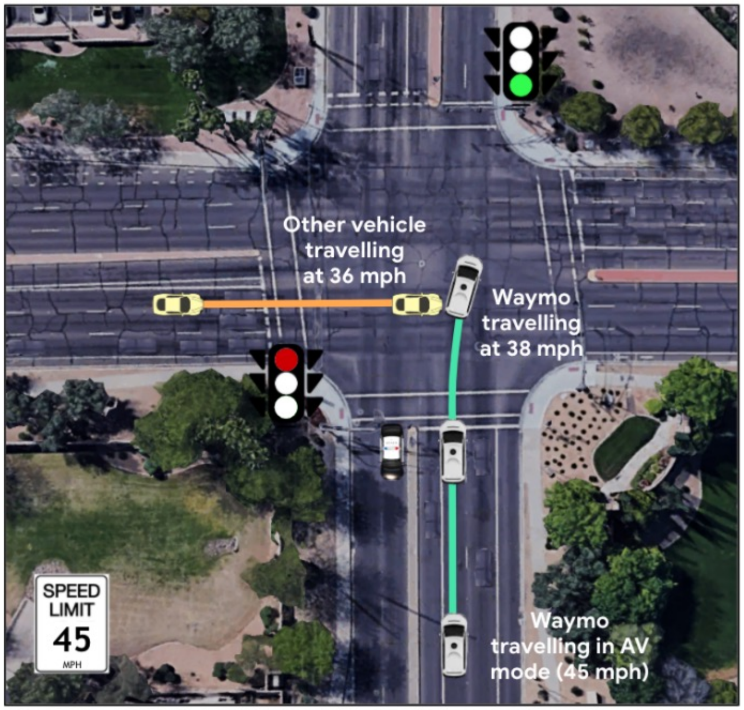

Other vehicle travelling at 36 mph
Waymo travelling at 38 mph
Waymo travelling in AV mode (45 mph)
SPEED LIMIT 45 MPH

Figure 6. Event E - A collision in which the other vehicle passed through a red stop light

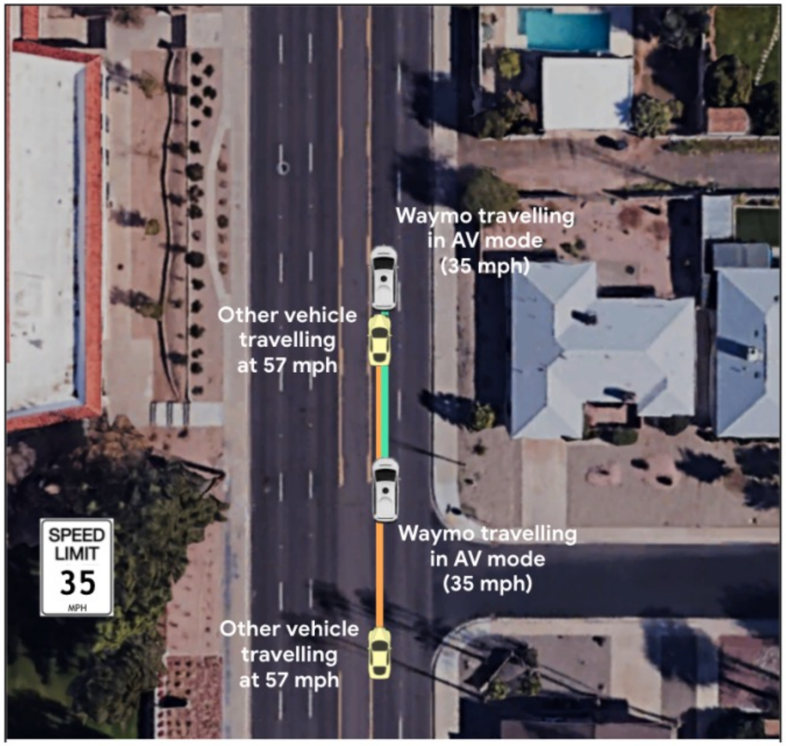

Waymo travelling in AV mode (35 mph)
Other vehicle travelling at 57 mph
SPEED LIMIT 35 MPH
Waymo travelling in AV mode (35 mph)
Other vehicle travelling at 57 mph

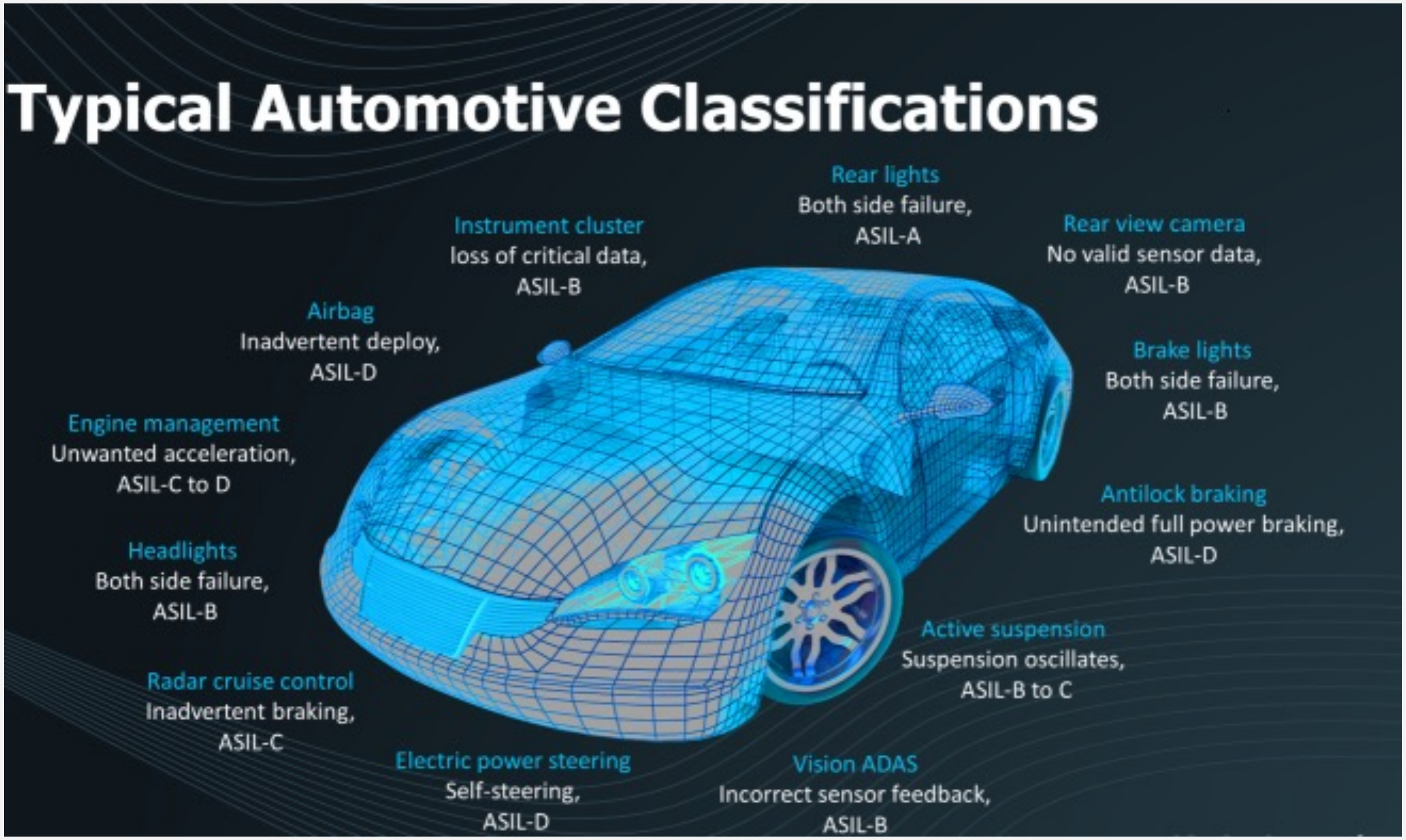
Figure 4. Event C: A rear end collision that resulted in airbag deployment for the vehicle that struck the Waymo vehicle.

Waymo scientists analyze every collision the Waymo cars are a part of

Source: Schwall et al, *Waymo Public Road Safety Performance Data*

ROBOTICS PERSPECTIVE

- What is a safe autonomous system?
- How safe would an autonomous car need to be for you to use it?
  - As safe as the average human driver?
  - Safer than the average human driver?
  - Safer than the top 5% of human drivers?
  - Safer than any human driver?
  - As safe as an aircraft? (1B hours per catastrophy)
- Perfect safety is impossible!
  - Solution: system confidence level? Human assistance?



Overview of different automotive safety classifications for automotive vehicles defined under ISO 26262, the current international standard for vehicles. ASIL: Automobile safety integrity level (risk classification scheme)

Image credit: Mentor

ARTIFICIAL INTELLIGENCE PERSPECTIVE

- Validating inductive reasoning is difficult (Hume 1748)
  - Not trivially suited for assessing respect of requirements.
  - How can we provide performance guarantees?
- Real world data cannot be guaranteed not to differ from training. Example?
  - We can do our best to have the best training possible.
  - Data is part of the safety concern.
- Salay et al: ISO 26262 only allows machine learning for components, but not the full system. How can we build a new standard for AI–powered intelligent systems?
- Do you believe it is possible to guarantee a certain level of safety with ML?

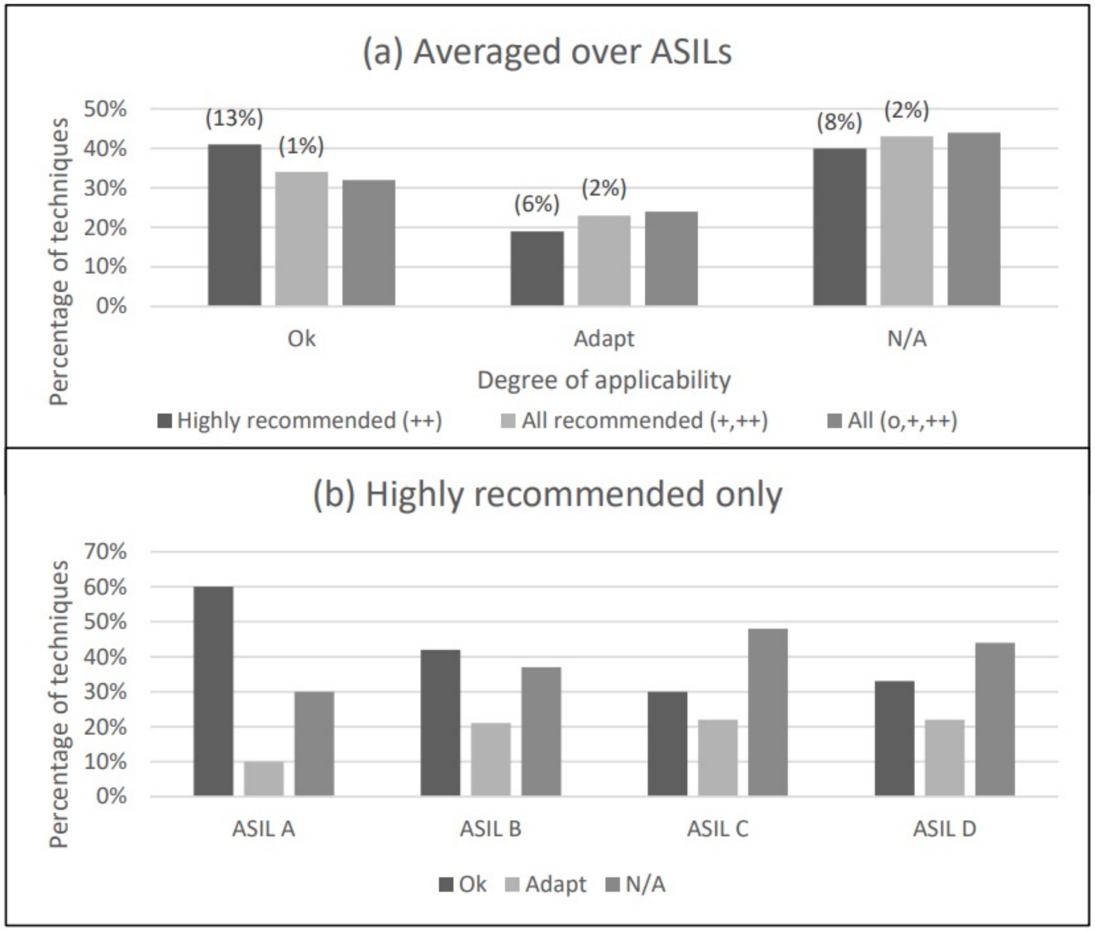


Fig. 3. Percentage of unit-level software techniques applicable to ML components: (a) values averaged over the four ASIL levels with standard deviation shown in parentheses; (b) values for each ASIL when only highly recommended techniques are considered.

In this paper, the authors look at ASIL testing techniques for safety at the software level and their applicability to ML-based systems

Chart a) groups ASIL verification technique by level of applicability to ML (from OK to non-applicable).

Chart b) focuses on the importance of that technique to the global safety verification process, from 0 to ++. ASIL level represents the level of risk, from A to D.

HARDWARE PERSPECTIVE

- Automotive systems rely on their hardware to function
- All hardware may crash if given enough time
  - Redundancy
  - Fault tolerance
- Core issue: undetected faults
- Regular inspection and maintenance may prevent multiple failures that could not be accounted by redundancy and fault tolerance.
- Would you be comfortable with your AI driver "only" having a 0.001% chance to suddenly fail?

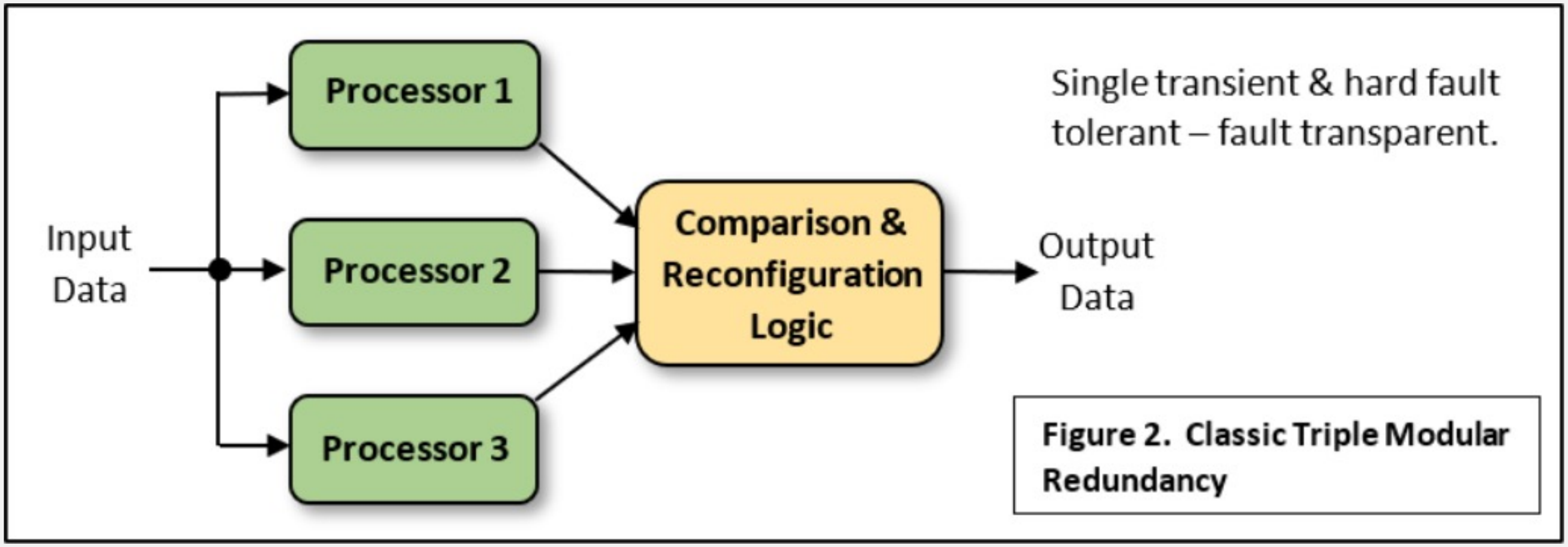


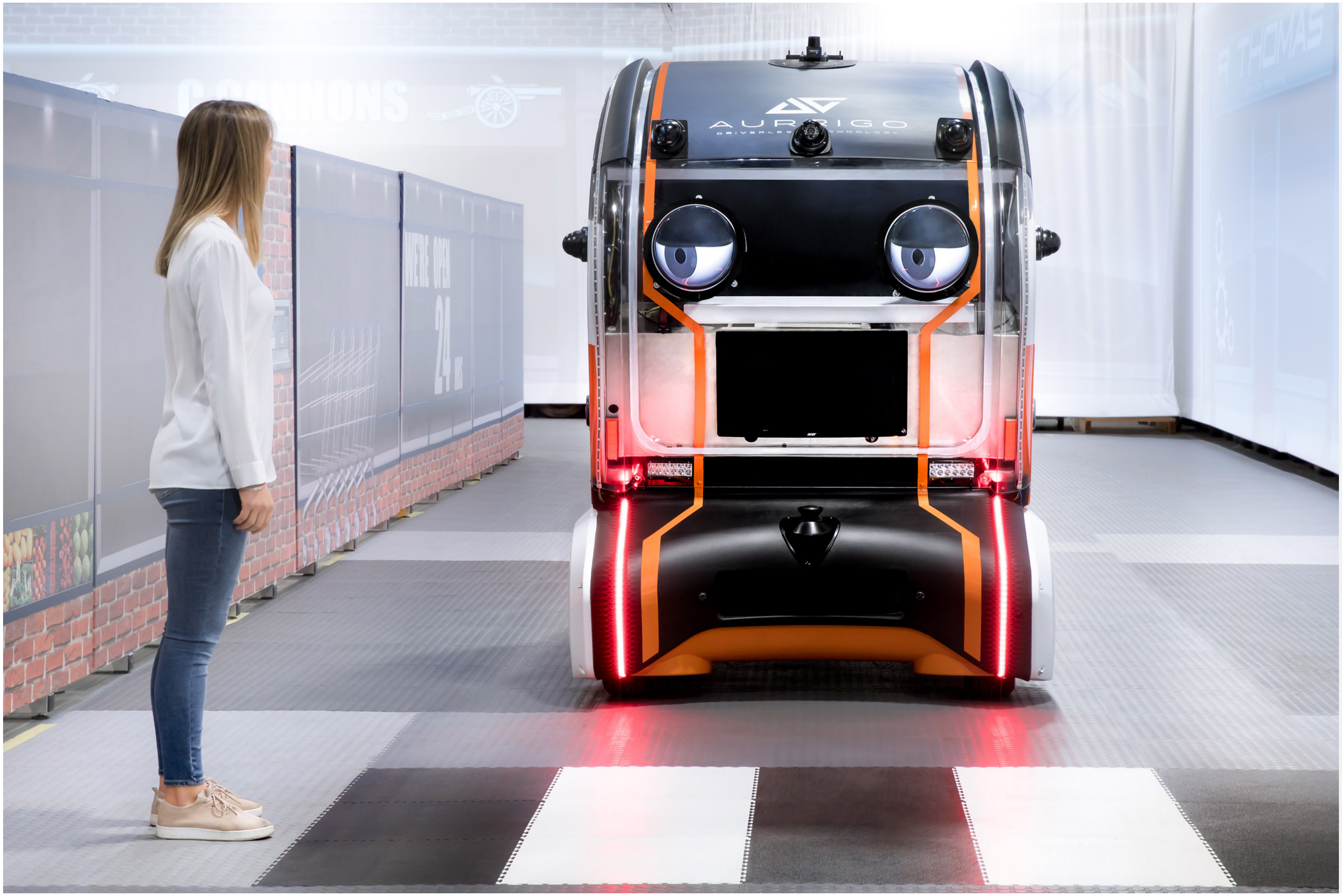
Figure 2. Classic Triple Modular Redundancy

Example of hardware redundancy via voting

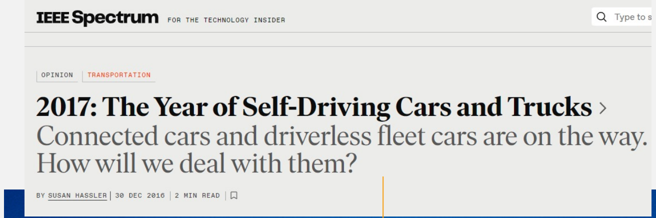
Source: Bill Marshall, PhD on TMR

HUMAN COMPUTER INTERACTIONS

- Self-driving vehicles serve no purpose if no one is willing to use them
- Autonomous vehicles must interact with cars, pedestrians..
  - Some humans are unpredictable
  - Some humans are straight up self destructive (children, drunk people, etc…)
- Other humans need to interact with the system!
  - Knowing self-driving cars are safer on average, would you feel safer crossing the street in front of a driverless vehicle than in front of a human driver?
- This is a difficult problem, maybe just as hard as the technical development.



Jaguar testing AI eye contact



IEEE Spectrum FOR THE TECHNOLOGY INSIDER

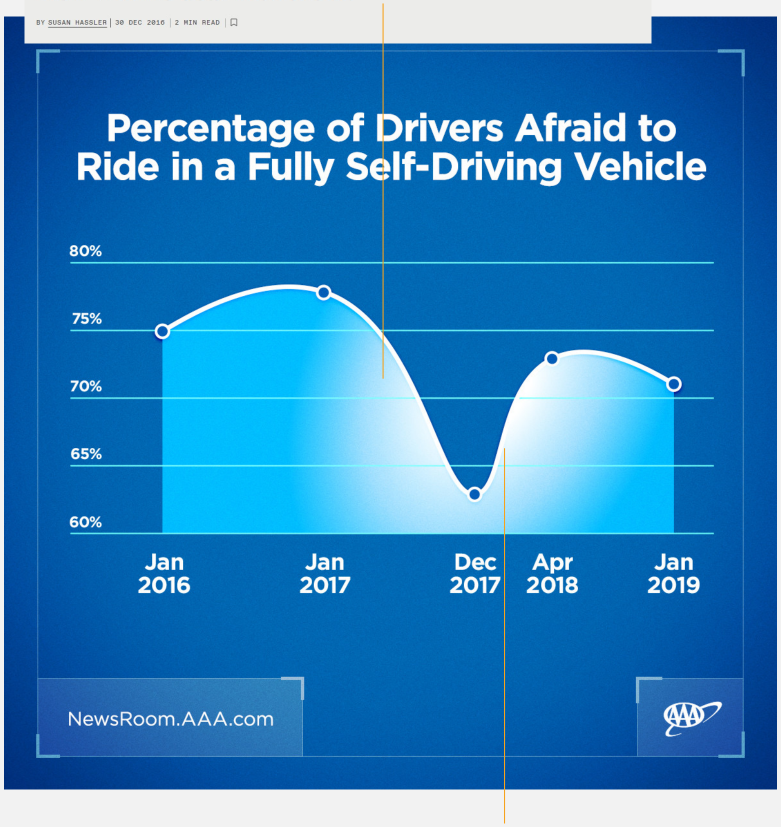
Type to s

OPINION | TRANSPORTATION

2017: The Year of Self-Driving Cars and Trucks ›

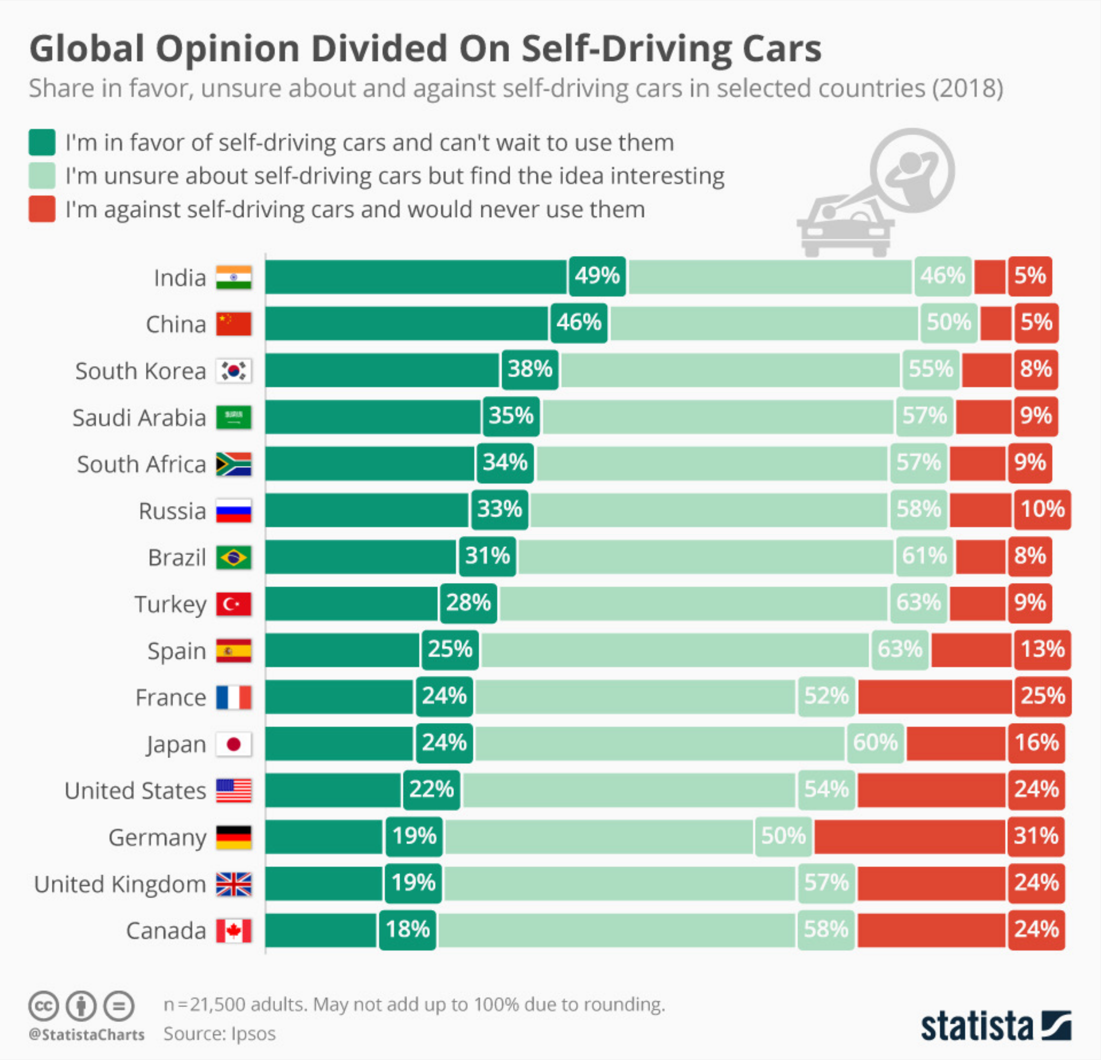
Connected cars and driverless fleet cars are on the way. How will we deal with them?

BY SUSAN HASSLER | 30 DEC 2016 | 2 MIN READ



Percentage of Drivers Afraid to Ride in a Fully Self-Driving Vehicle

Uber self-driving car kills a pedestrian despite a human supervisor being in the car



Global Opinion Divided On Self-Driving Cars

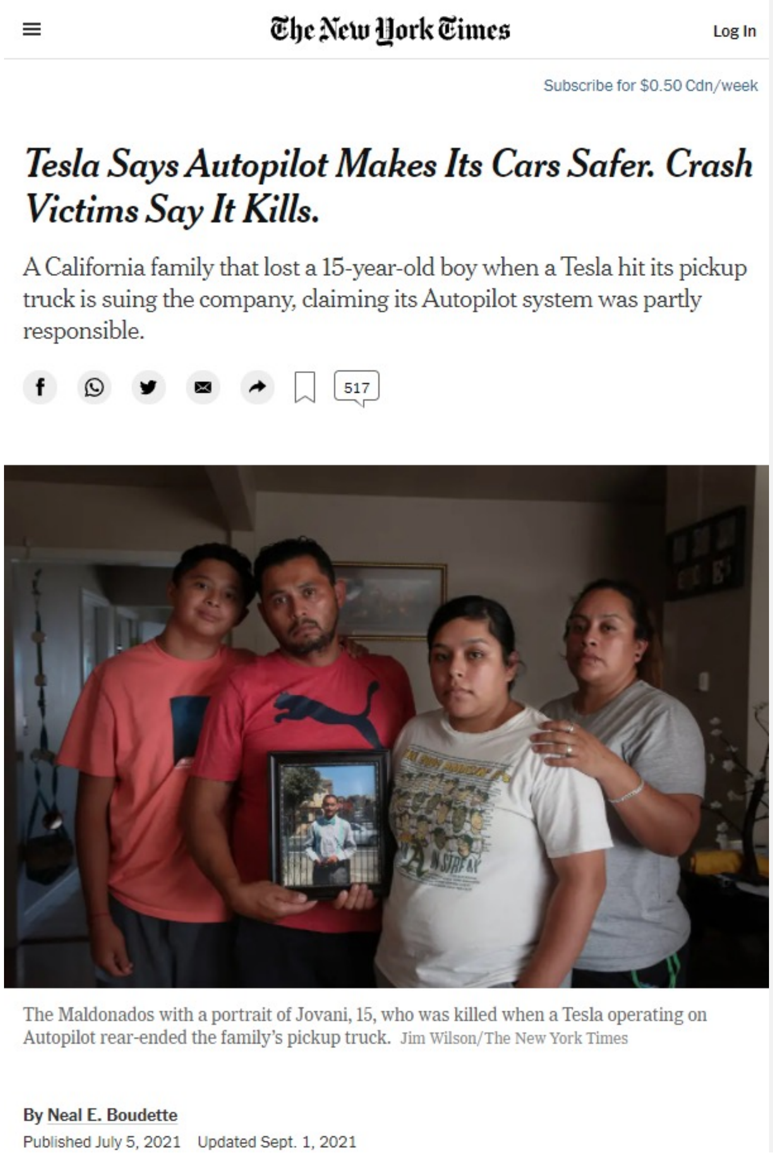
Share in favor, unsure about and against self-driving cars in selected countries (2018)

n = 21,500 adults. May not add up to 100% due to rounding.

Source: Ipsos

LEGAL PERSPECTIVE

- Self-driving vehicles serve no purpose if no one is willing to produce them
- All cars cause accidents
  - **Person A** gets into their car without checking that every system is operational
  - **Person B**, the mechanic, did the last inspection on the car
  - **Person C** crosses the street on a very late green light and gets hit by the car
  - **Company Z** made the car. **Who is responsible?**
- Can we trust the car's logs? We already know at least one thing went wrong!
- How does this work with insurance?



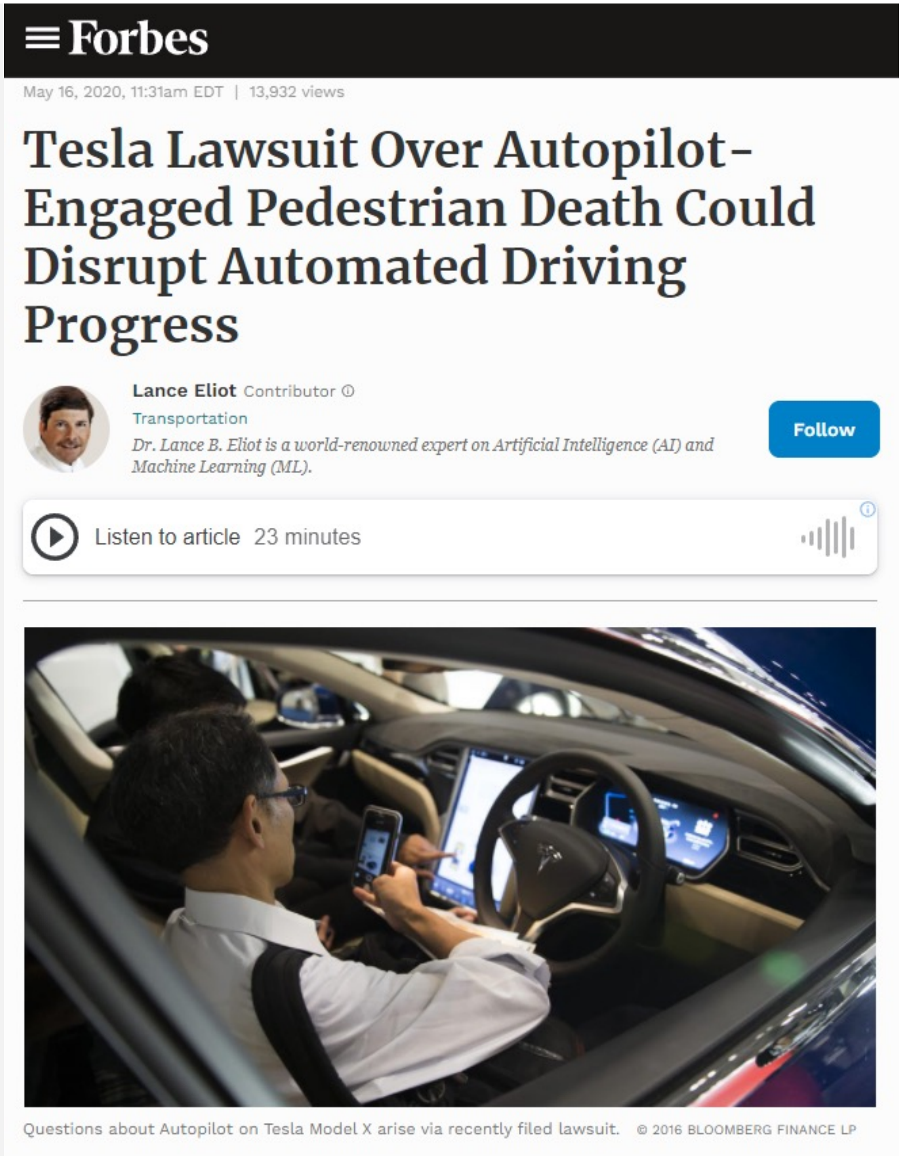
Tesla Says Autopilot Makes Its Cars Safer. Crash Victims Say It Kills.

A California family that lost a 15-year-old boy when a Tesla hit its pickup truck is suing the company, claiming its Autopilot system was partly responsible.

The Maldonados with a portrait of Jovani, 15, who was killed when a Tesla operating on Autopilot rear-ended the family's pickup truck. Jim Wilson/The New York Times

By Neal E. Boudette

Published July 5, 2021 Updated Sept. 1, 2021



May 16, 2020, 11:31am EDT | 13,932 views

Tesla Lawsuit Over Autopilot-Engaged Pedestrian Death Could Disrupt Automated Driving Progress

Lance Eliot Contributor

Transportation

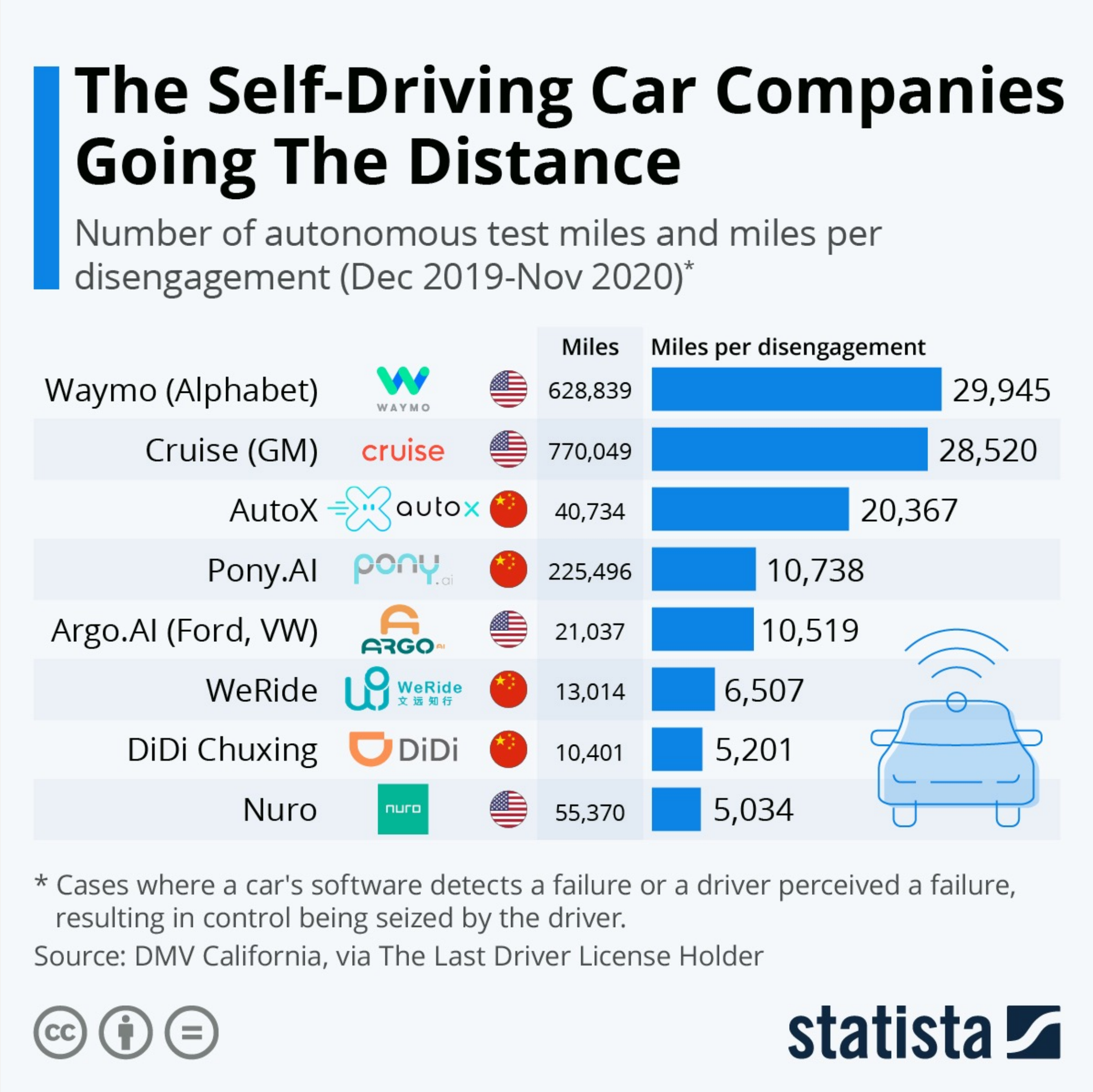
Dr. Lance B. Eliot is a world-renowned expert on Artificial Intelligence (AI) and Machine Learning (ML).

Questions about Autopilot on Tesla Model X arise via recently filed lawsuit. © 2016 BLOOMBERG FINANCE LP

FUTURE DIRECTIONS

- Current certification systems like ISO26262 cannot apply
- Building a new certification system is heavily multi-disciplinary and needs to apply to a wider range of techniques
  - For instance, requirements could focus on intent and behavior rather than specific technical details that are not generalizable.
- Need to combine perspectives and backgrounds to find an optimal tradeoff between performance, cost, risk and ethical considerations.

- Do you think we will have self driving cars for the general public in the next 10 years?



The Self-Driving Car Companies Going The Distance

Number of autonomous test miles and miles per disengagement (Dec 2019-Nov 2020)*

* Cases where a car's software detects a failure or a driver perceived a failure, resulting in control being seized by the driver.

Source: DMV California, via The Last Driver License Holder

And yet they drive..