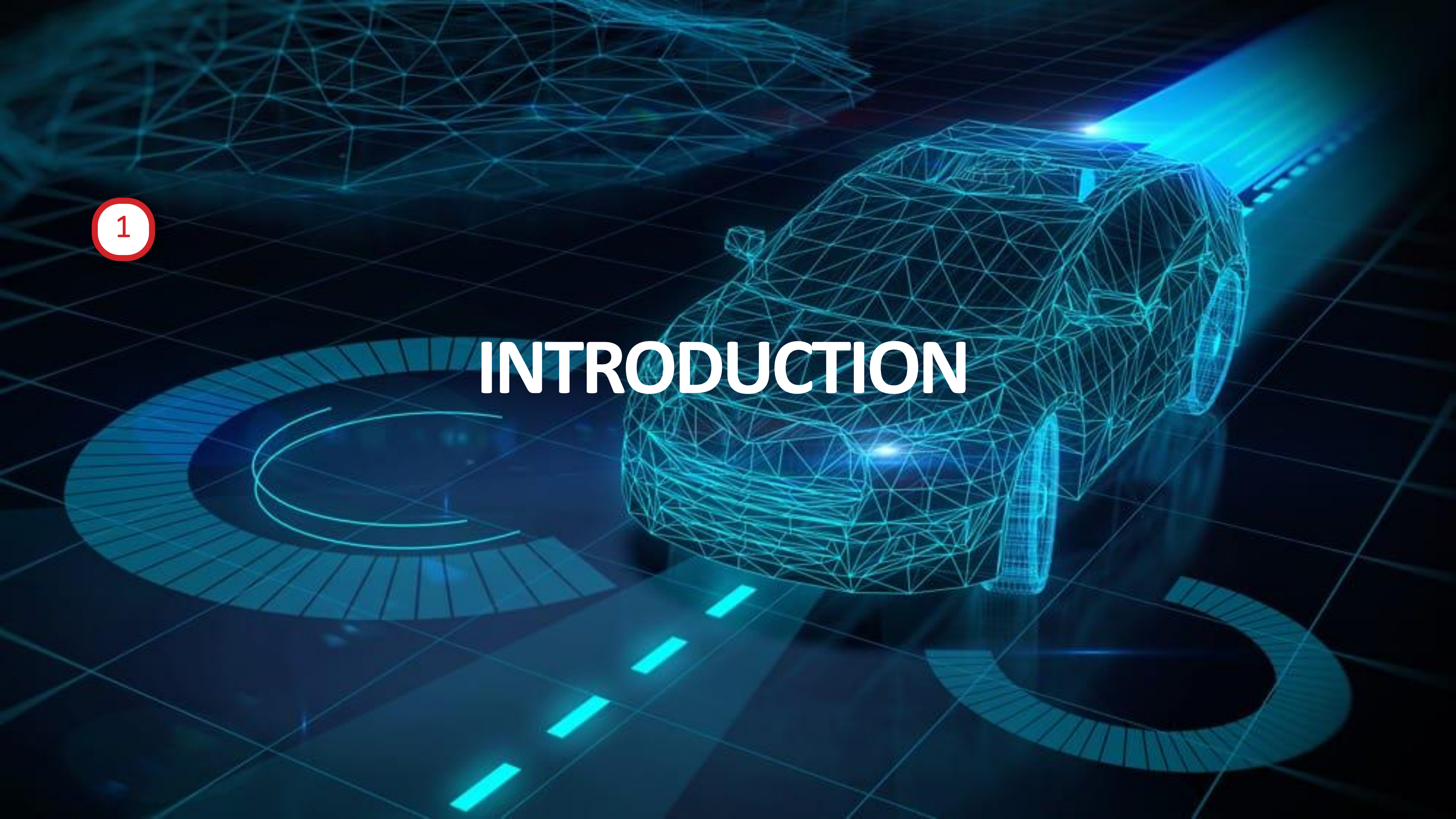
Assuring safety of AI-based Automated Driving Systems

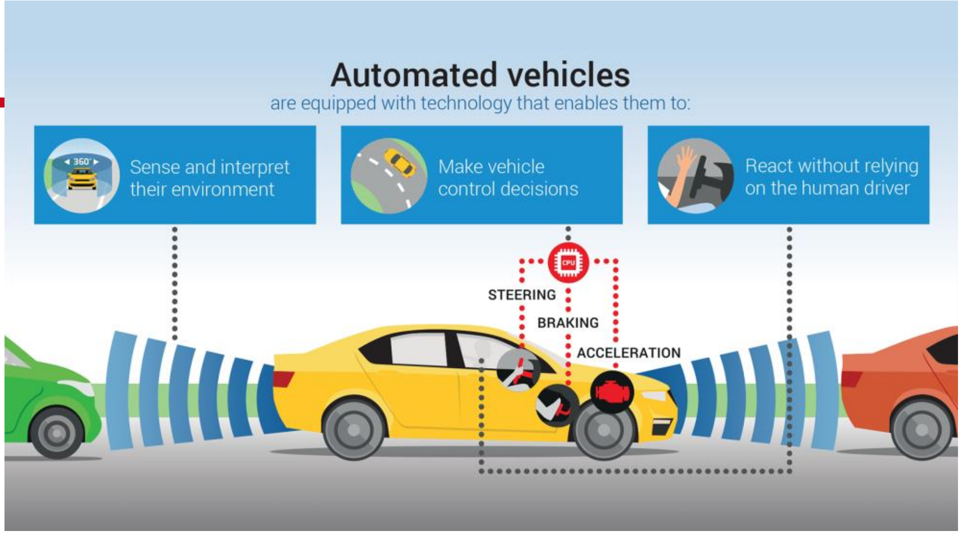
Vamsi Madasu & Kevin Anderson
SYSTRA ANZ

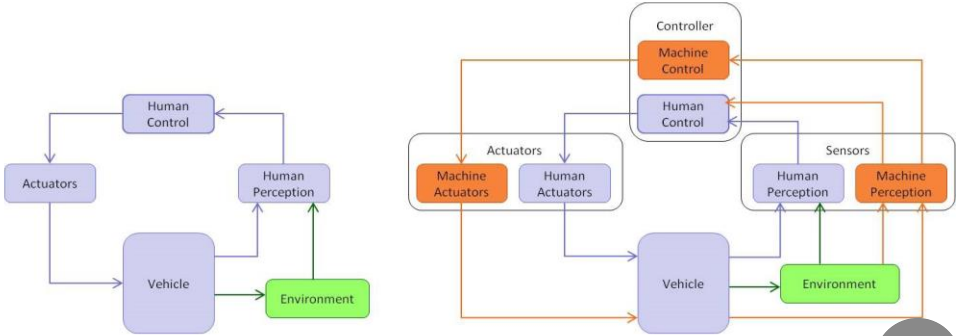
Agenda

- Introduction
- Safety Assurance
- Risk-based safety assurance
- Safety Assurance and Artificial Intelligence (AI)
- Three major concepts of AI Safety:
  - Robustness
  - Assurance
  - Specification
- Conclusion


1
INTRODUCTION

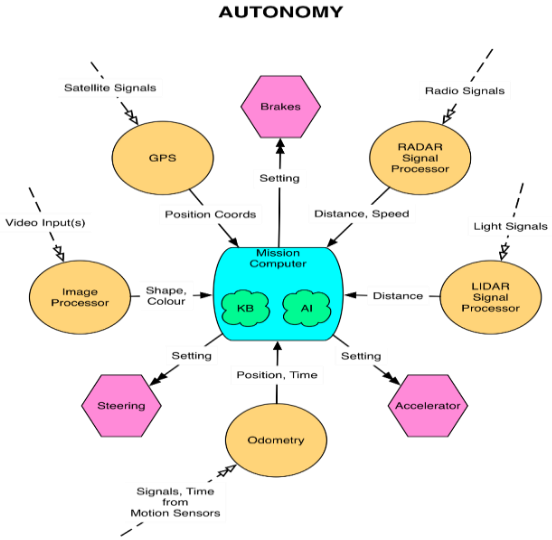
Automated Vehicles

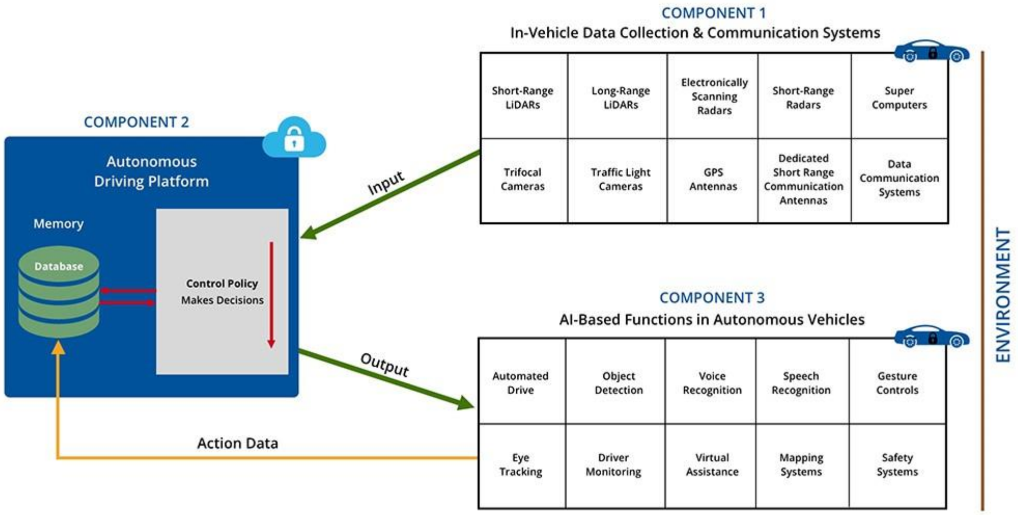




Autonomy

- Automated Vehicles employ a number of sensor-based systems that rely on Artificial Intelligence for decision making in Automated Driving Systems (ADS)
- Full autonomy already exists in the realm of 'extant technologies'





AI techniques employed in ADS

Main AI Technique	Topic	AI Techniques
Artificial Neural Network	Conceptual Model and Framework	HoughTransforms, HoughLines, LocalMaximaFinder, Kalman filters and Convolutional Neural Network (CNN)
	Fault Prevention	KNN, SVM Regression (SMO), ANN
	Navigation and Control	CBR, ANN, fuzzy logic, Nearest-Neighbor Retrieval Algorithm, Basic AI Path Planning algorithms such as A* and D*
		ANN combined to Genetic Algorithm - Neuroevolution of Augmenting Topologies (NEAT)
	Sensors and Perception	ANNs, AdaBoost, SVM, Hidden Markov Models (HMMs), CRFs
		Clustering algorithm k- mean, ANN
		HOG, SVM, PCA, ANN
Hidden Markov Based Models	Navigation and Control	GMM, Continuous Hidden Markov Model (CHMM), Discrete Hidden Markov Model (DHMM)
	Sensors and Perception	HMM, Viterbi algorithm, Adaboost trained Haar-like feature detector
Hough Transformation	Navigation and Control	Haar Feature Based Cascade Classifier, Canny edge detection and Hough line transformation
Novel Image Recognition Technique	Sensors and Perception	Combination of mathematical techniques
Regression Based Models	Navigation and Control	(DRF) and Linear Regression (LR)
Support Vector Machine (SVM)	Sensors and Perception	Haar, HOG, LBP, Chanel features, SVM
		k-Nearest Neighbours (kNN), Naïve Bayes classifier (NBC), SVM
		Principal component analysis network (PCANet), SVM
		SVM, HOG

2

SAFETY ASSURANCE

Principles of Safety Assurance

- Risk based principles are endemic to safety assurance and due diligence:
  - ‘Not less safe’ – a difficult progression at Level 3 “Conditional Automation” - requiring the driver to remain available (sober!) to take over
  - ‘Compliance with standards’ – ISO 26262, UL4600 and IEC 61508
  - ‘Good practice’ - gradual leaching of top-end luxury systems into the mass market
  - ‘SFARP’ - monitoring of technology developments as to what is reasonably practicable should accelerate decreasing TLOS in line with road toll reductions
  - ‘Continuous improvement’ - A ‘risk timeline’ is implied both by the setting of Levels of Automation and the setting of Automotive Safety Integrity Levels (ASILs) in ISO 26262. This presages setting the Driverless Car TLOS as a fraction of the road toll.
- These principles may be utilised in parallel or subsume one another.

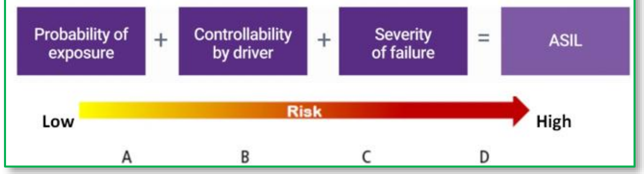
Safety Assurance of Automated Vehicles

- Automated Vehicles (AVs) are generally tested by the companies that design/manufacture them.
- AV companies must comply with relevant national regulations and Motor Vehicle Safety Standards; and certify that their vehicles are free of safety risks.
- Many AV companies are testing vehicles with higher levels of automation to ensure that they operate safely as intended, **but**
- More work remains to be done by AV manufacturers and national regulators to ensure the safe operation of automated vehicles before they are available for consumers to purchase.

Safety Assurance Techniques

- Functional Safety
  - Automotive Safety Integrity Levels (ASILs)
  - ISO 26262 (derived from IEC 61508)
- Failure probabilities
  - Autonomy fails, and
  - Vehicle fails to detect or execute action
- Failures rates
  - Number of miles driven before failure
- Risk-based Safety

Automotive Safety Integrity Level (ASIL)



$$P_{Loss(i)} = P_{Failure(i)} * ((1-P_{Detection(i)}) + (1-P_{Mitigation(i)})) + P_{HumanMistake}$$

$$Risk = Sum(P_{Loss(i)} * Severity_{(i)})$$

Statistical Question: How many miles (years[a]) would autonomous vehicles have to be driven...	Benchmark Failure Rate: (A) 1.09 fatalities per 100 million miles?	(B) 77 reported injuries per 100 million miles?	(C) 190 reported crashes per 100 million miles?
(1) without failure to demonstrate with 95% confidence that their failure rate is at most...	275 million miles (12.5 years)	3.9 million miles (2 months)	1.6 million miles (1 month)
(2) to demonstrate with 95% confidence their failure rate to within 20% of the true rate of...	8.8 billion miles (400 years)	125 million miles (5.7 years)	51 million miles (2.3 years)
(3) to demonstrate with 95% confidence and 80% power that their failure rate is 20% better than the human driver failure rate of...	11 billion miles (500 years)	161 million miles (7.3 years)	65 million miles (3 years)

[a] We assess the time it would take to compete the requisite miles with a fleet of 100 autonomous vehicles (larger than any known existing fleet) driving 24 hours a day, 365 days a year, at an average speed of 25 miles per hour.

3

RISK-BASED SAFETY ASSURANCE

Risk-based Target Levels of Safety (TLOS)

- Our view of socially acceptable TLOS (one chance per million years) is based on quantified fatality risk targets (Madasu and Anderson, ASSC 2017).
- Table 1 provides some statistics related to motor vehicle deaths in Australia by year.
- Extrapolation to 2039 suggests a target of 10% of the road toll of less than one chance of individual fatality per million years.
- This shows just how much the 'continuous improvement' principle holds sway over 'not less safe'.

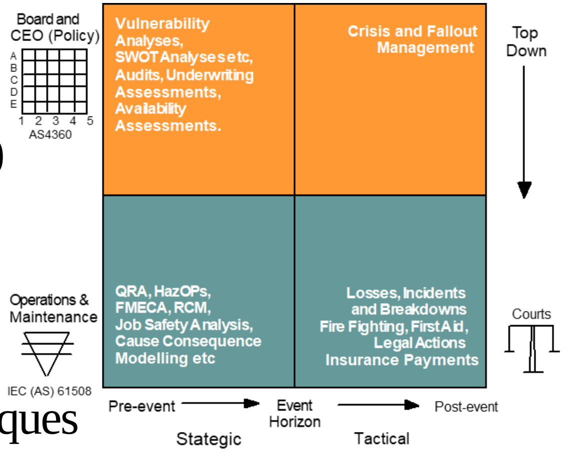
	AV Commercial Drivers (Workers)	AV Commuters / Pedestrians & Other road users (General Public)
Individual risk	6E-09 fatalities per hour	6E-10 fatalities per hour
Collective risk	5E-05 fatalities per annum	5E-06 fatalities per annum

Year	Road Deaths	Population (m)	Rate / mil.yr
1939	1,426	7.00	203.7
1959	2,264	10.16	222.8
1979	3,508	14.60	240.3
1999	1,764	18.92	93.2
2019	1,194	25.37	47.1
2039	You do	The maths	<1or 2

Note:
Multiple fatalities can occur but the exposure in terms of hours per year is dominated by single person occupancy events.

Assurance techniques for TLOS

- TLOS can be achieved using a combination of Fault Tree Analysis (FTA) and Event Tree Analysis (ETA), *aka*, Cause-consequence models
- These fall under the category of ‘Failure analysis’ of Inputs, Processes and Outputs which include:
  - Markov Models
  - Failure Mode Effects and Criticality Analysis (FMECA)
  - Reliability Block Diagrams (RBD)
  - Monte-Carlo simulation
- Also relevant are AI techniques, such as:
  - Neural networks, Bayesian approaches, Learning techniques
  - Data analytics



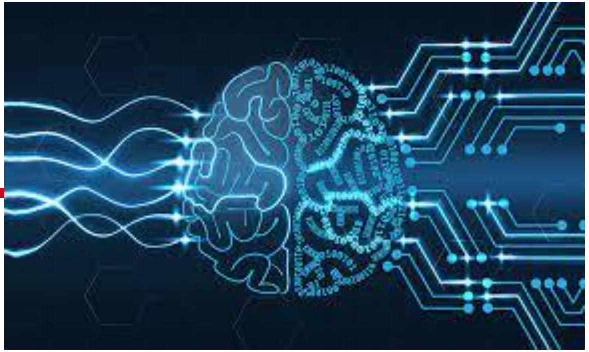
Cause –consequence modelling

- Reliability of inputs/outputs can be calculated using Cause-Consequence modelling
- Herein, we separate the causal FTA likelihood from the resultant ETA consequence.
- In IEC 61508, the transition point is called a ‘Hazardous Situation’. We call it ‘Loss of Control’ (LOC) as, when an incipient hazard actually happens AND the concomitant Control System also fails, the result is a balance of probability:

Input	HAZARD					HARM (fatality)	
GPS, Radar, Lidar,	per hour				1%		
Odometry, Image processor	H	L				1.00E-09	per hr
Output	1.00E-04	1.00E-05				1000	hr per yr
Steer, Brake, accelerate			LOSS OF CONTROL			1.00E-06	per yr
Mission Computer			1.00E-07	per hour			
Knowledge Base	CONTROL		Tradeoff (not HH)			NULL (accident)	
Artificial Intelligence	% risk reduction fail				99%		
	L	H					
	1.00E-03	1.00E-02					


4
AI & SAFETY ASSURANCE

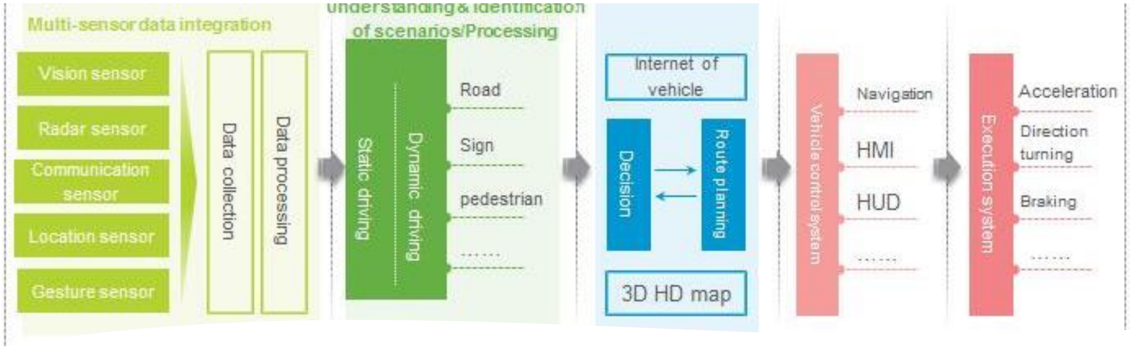
AI for assuring safety



- AI itself is 'Not Recommended' (NR) in IEC 61508 and thereby, ISO 26262.
- The use of AI approaches may have some value in calculating trends and deriving rules for diverse channels based on specification checking, subject to no common faults in such specifications.
- UL 4600 defines AI techniques as computational algorithms and other <u>techniques</u> that include inductive learning, intentionally non-deterministic behaviour, rule-based systems, computer vision, and heuristic searches.
- This encompasses software that is not generally amenable to pre-AI software safety approaches, <u>whether or not</u> actual "intelligence' is actually involved.

Assuring AI Safety

- An AI-based Automated Driver System (ADS) must operate safely under a wide range of road conditions.
- Artificial Intelligence/Machine Learning (AI/ML) systems follow a pre-specified algorithm to learn from data, enabling them to achieve a specific goal.
- It is necessary to specify what constitutes a safe behavior of the AI function, including under which conditions the component will provide which service.
- Additionally, it must be assured that the behavior implemented by the AI function is safe under all conditions.



5

ASSURING SAFETY OF AI-BASED ADS

Specification

- Specification of AI systems refers to defining a system's goal in a way that ensures its behaviour aligns with the human operator's intentions.
- Poor specification of a machine learning system's goal can lead to:
  - AV will not operate as intended; and
  - Safety hazards as AVs operate in a high-stakes environment
- Designers must take special care to specify an objective that will lead to the desired behaviour. If the goal set by the system designer is a poor proxy for the intended behaviour, the system will learn the wrong behaviour and be considered *mis-specified*.

Specification ensures that an AI system's behaviour aligns with the operator's true intentions.

Robustness

Robustness ensures that an AI system continues to operate within safe limits upon perturbations.

- Challenging inputs for AI systems in ADS can come in many shapes and guises, including situations a system may never have encountered before
- Operating safely in such scenarios means that a system must:
  - Recognize that it has not been trained for such a situation; and,
  - Have a way to act safely
- The ADS should have the ability to quantify whether or not it is confident about a prediction (Predictive uncertainty estimates) -
  - ➢ Reduces the chance of failure in situations which the system is not well-prepared to handle
  - ➢ The system, upon recognizing it is in a setting it was not trained for, could then revert to a safe fallback option or alert a human operator

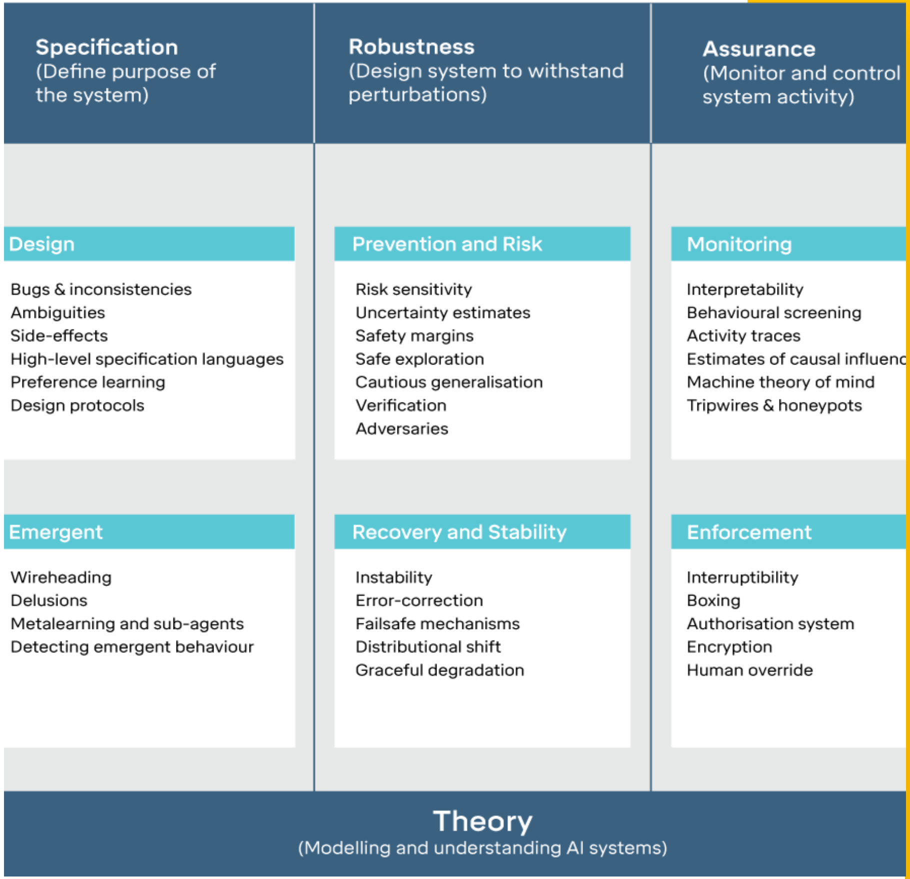
Assurance

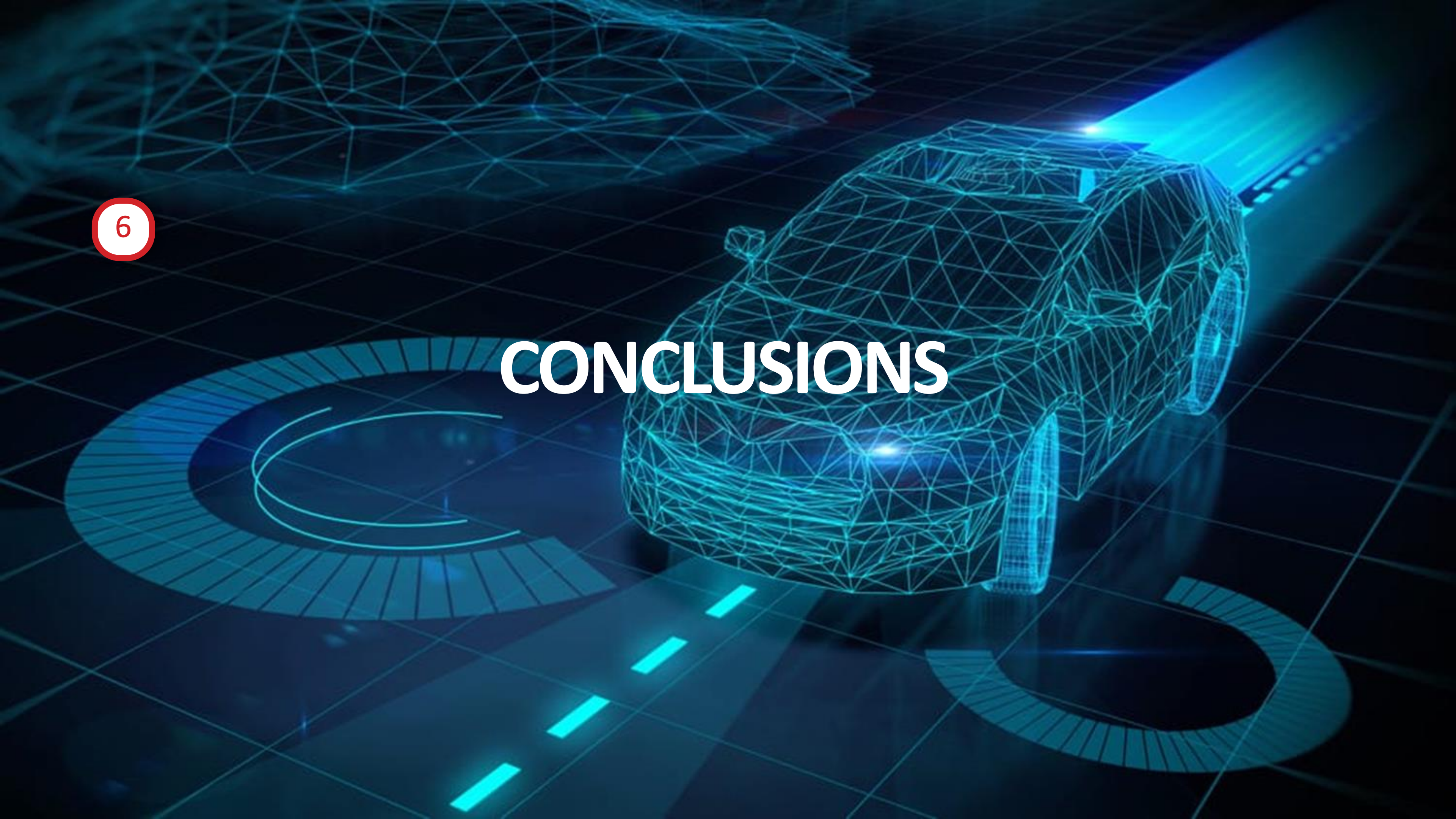
- Extant safety assurance techniques are poorly suited to modern AI and Machine Learning (ML) systems in ADS, such as, Deep Neural Networks.
- Interpretability (also sometimes called explainability) in AI refers to the study of how to understand the decisions of the AI function, and how to design systems whose decisions are easily understood, or interpretable.
- Interpretability will be crucial in giving drivers the confidence to act on predictions obtained from AI/ML based ADS as they will interact with system in real-time.

Assurance ensures that we can understand and control AI systems during operation.

Implementing AI Safety

- **Requirements specification**
  - Safety Goals / Objectives
  - Safety requirements
- **Design Analysis**
  - Safety architecture
  - Risk analysis
- **Sensitivity Analysis**
  - Monitoring
  - Trustworthiness
  - Explainability




6
CONCLUSIONS

Key Points

- Risk-based safety assurance for AVs has a higher-level objective (TLOS)
- Cause-consequence modelling can be employed for showing compliance with the higher-level objective
- In addition to compliance with regulations and standards, AI techniques can also be utilised to assure safety of AV systems
- Three concepts are introduced for assuring safety of AI based ADS:
  - Robustness guarantees that a system continues to operate within safe limits even in unfamiliar settings;
  - Assurance seeks to establish that it can be analyzed and understood easily by human operators; and
  - Specification is concerned with ensuring that its behavior aligns with the system designer's intentions.

THANKS FOR YOUR ATTENTION

SYSTRA