Fraunhofer IKS

Fraunhofer Institute for Cognitive Systems IKS

2022-10-21

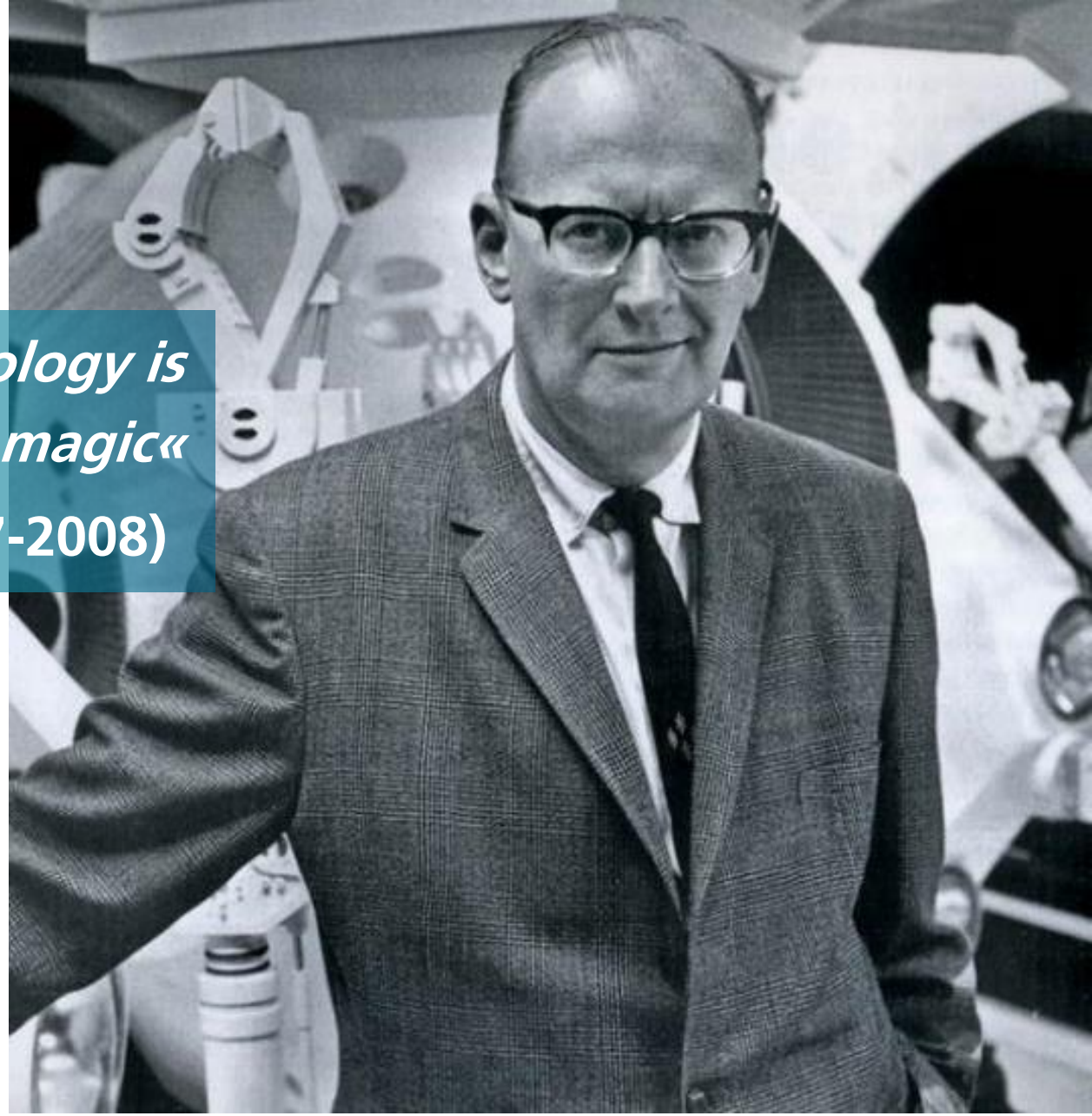
Deutsche Sektion der Internationalen Juristen-Kommission e.V.
Prof. Dr. Simon Burton

Trustworthy AI-based autonomous systems:

What is artificial intelligence?
Definitions and applications fields,
What are the limits of autonomous systems?

»Any sufficiently advanced technology is indistinguishable from magic«

Arthur C. Clarke (1917-2008)



Artificial Intelligence: capability to acquire, process, create and apply knowledge, held in the form of a **model**, to conduct one or more given tasks

http://hitchhikers.movies.go.com/downloads/h2g21024x768b.jpg

General AI: "AI that addresses a broad range of tasks with a satisfactory level of performance"
Science fiction (still?!)

Narrow AI: "AI that is focused on defined tasks to address a specific problem"
Reality (now!)

Artificial Intelligence: capability to acquire, process, create and apply knowledge, held in the form of a **model**, to conduct one or more given tasks

Machine Learning: process of optimizing model parameters through computational techniques, such that the model's behaviour reflects the **data** or experience

Supervised Machine Learning: machine learning that makes use of **labelled data** during training

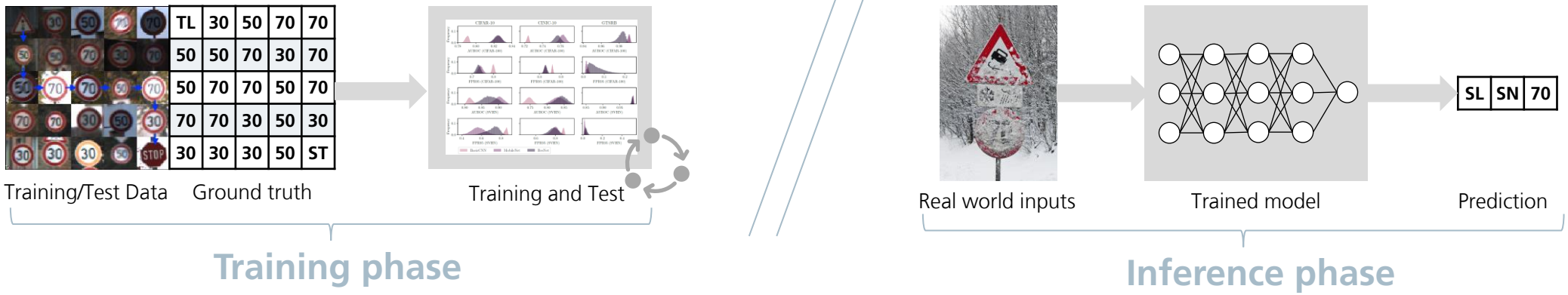
Deep Learning: approach to creating **rich hierarchical representations** through the training **of neural networks** with many hidden layers

Trustworthy AI

Machine Learning in a nutshell

Machine Learning

Optimizes model parameters* through computational techniques, such that the models behaviour reflects the data or experience

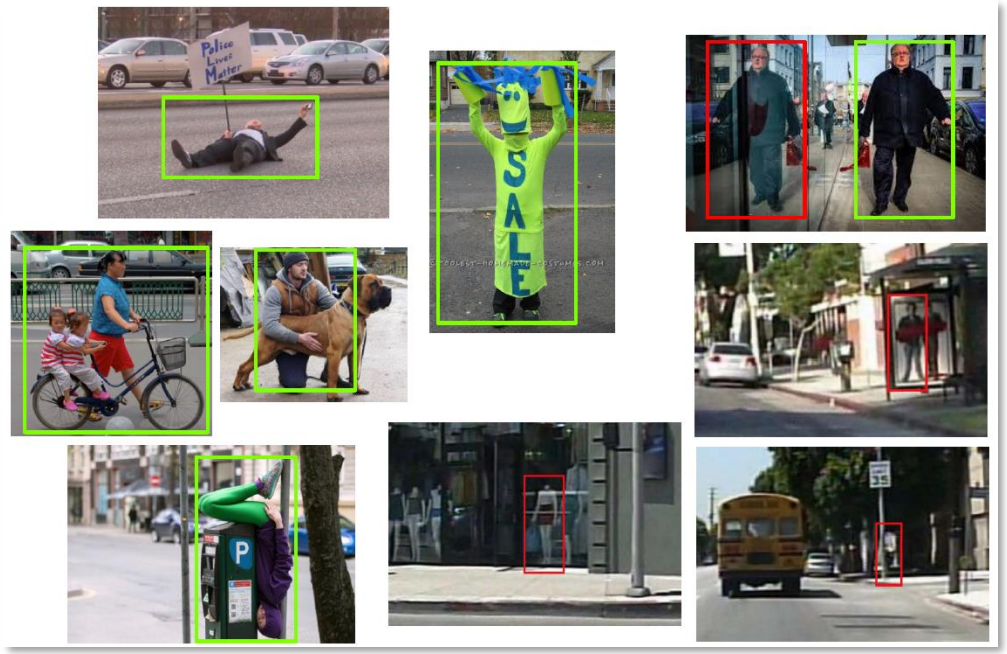


*Can be hundreds of thousands or millions of parameters

Trustworthy AI

How difficult can it be?

Source: Twitter @teenybiscuit:
https://twitter.com/teenybiscuit



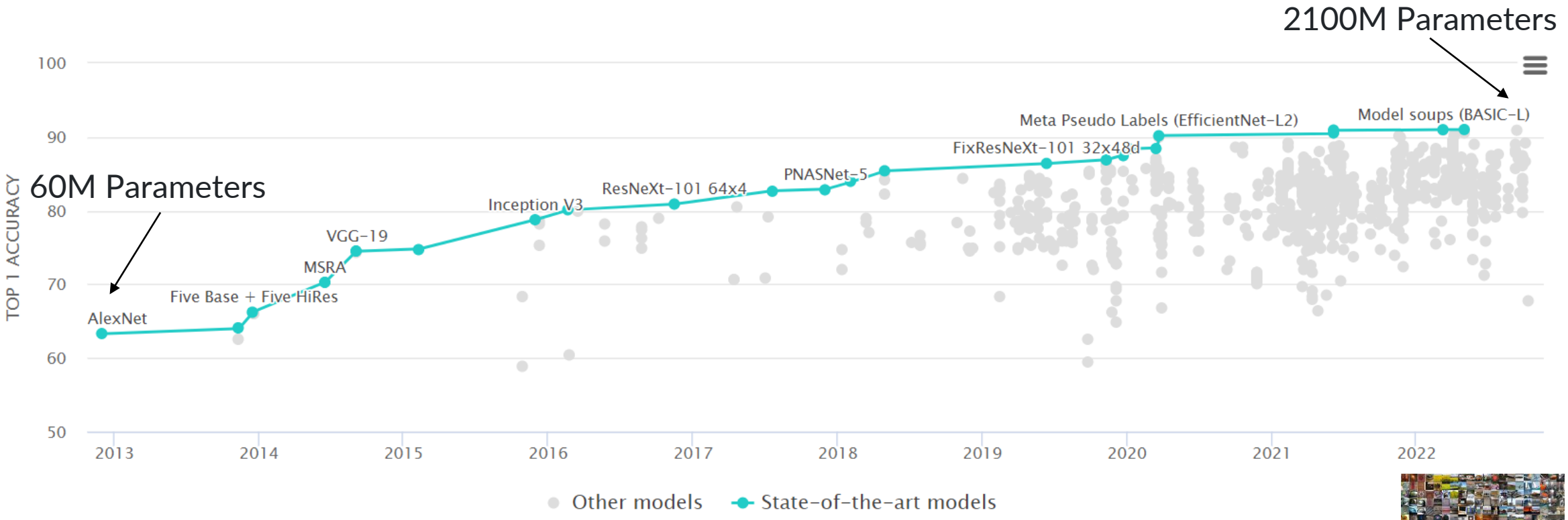
Krzysztof Czarnecki and Rick Salay: "Towards a Framework to Manage Perceptual Uncertainty for Safe Automated Driving."

  Public information

Fraunhofer IKS

Trustworthy AI

How difficult can it be?



https://paperswithcode.com/sota/image-classification-on-imagenet

The assumption in AI has generally been that if it works often enough to be useful, then that's good enough, but that casual attitude is not appropriate when the stakes are high.«

Gary Marcus,
Rebooting AI

Cognitive Safety-Critical Cyber Physical Systems

Real-world applications of AI

Cognitive systems are software-intensive technical systems that imitate cognitive capabilities such as **perception**, **learning**, and **reasoning**.

Automated driving

Industrial Robotics

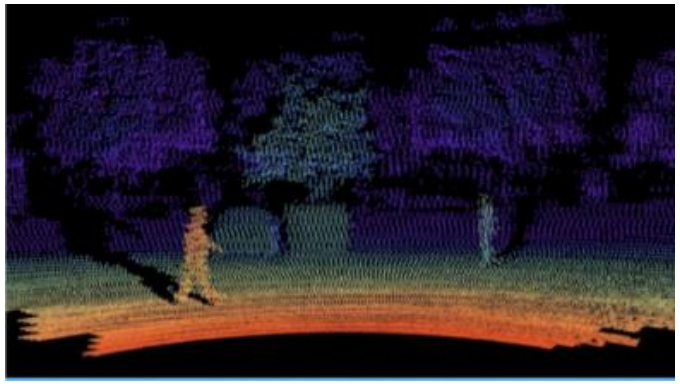
Driverless trains

Medical devices

Cognitive Safety-Critical Cyber Physical Systems

Why are these real-world problems difficult to solve?

Source: https://www.bbc.com/news/world-asia-india-38155635



Source: https://velodynelidar.com

Source https://www.cityscapes-dataset.com/examples

Scope & unpredictability of the environment and critical events

Inaccuracies & noise in environmental sensors and signal processing

Machine learning for perception, learning, and reasoning

Safety is becoming less about what happens when individual technical components break and more about managing the emergent risk associated with increasing complexity

System complexity

Emergent properties of complex systems

A **complex system** exhibits behaviours that are **emergent properties** of the interactions between the parts of the system, where the behaviours would **not be predicted** based on knowledge of the parts and their interactions alone.

Caused by **lack of knowledge** regarding:

- Semi-permeable boundaries
- Non-linearity, mode transitions, tipping points
- Self-organization and ad-hoc systems

Burton, Simon, et al. "Safety, Complexity, and Automated Driving: Holistic Perspectives on Safety Assurance." Computer 54.8 (2021): 22-32.

Consequences of system complexity

The semantic gap

Semantic Gap* – discrepancy between the intended and specified functionality, caused by:

- Complexity and unpredictability of the operational domain
- Complexity and unpredictability of the system itself
- Increasing transfer of decision function to the system

…leads to moral responsibility and legal gaps!

*Burton, Simon, et al. "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective." Artificial Intelligence 279 (2020): 103201.

Uncertainty:

Any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system*

*W. E. Walker et al. “Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support”. In: Integrated Assessment 4.1 (Mar. 2003), pp. 5–17. ISSN: 1389-5176. DOI: 10.1076/iaij.4.1.5.16466.

Consequences of system complexity

Impact of AI on uncertainty

Specification uncertainty:

- Data as the specification: No explicit definition of “safe” behaviour
- Changing environment over time

Technical uncertainty:

- Robustness: outputs sensitive to small changes in the inputs

Assurance uncertainty:

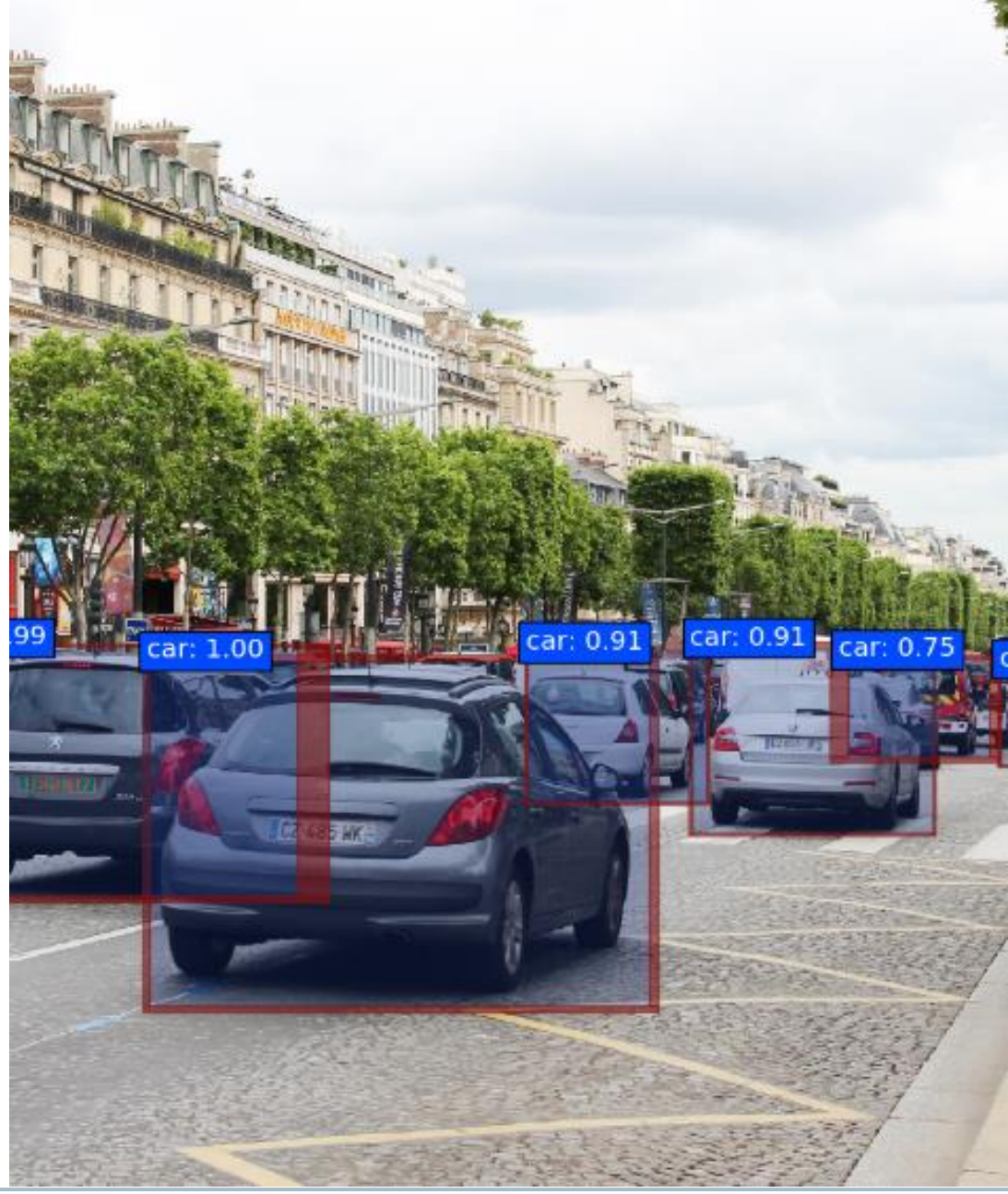
- Explainability: Learnt concepts are not understandable by humans

Ethical uncertainty:

- How good is good enough? Positive risk balance vs. absence of unreasonable risk

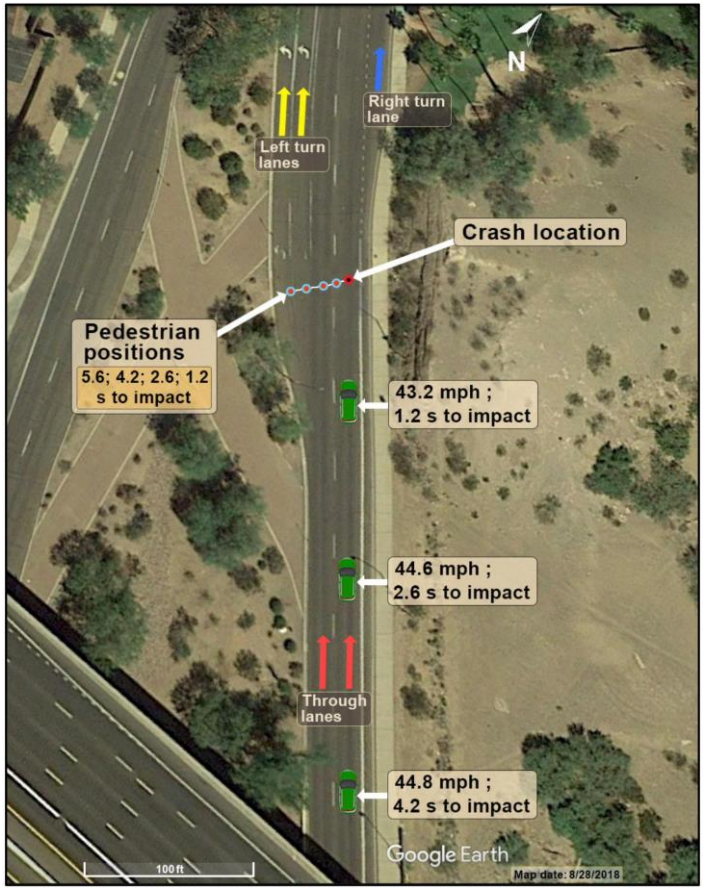
Legal uncertainty:

- No consensus on standards (yet), unclear legal framework, who is responsible for accidents involving autonomous systems



Case Study – Uber Tempe incident

Interacting layers of complexity and uncertainty



Time to Impact (seconds)	Speed (mph)	Classification and Path Prediction[a]	Vehicle and System Actions[b]
-9.9	35.1	--	Vehicle begins to accelerate from 35 mph in response to increased speed limit.
-5.8	44.1	--	Vehicle reaches 44 mph.
-5.6	44.3	Classification: *Vehicle*—by radar Path prediction: *None*; not on path of SUV	Radar makes first detection of pedestrian (classified as vehicle) and estimates speed.
-5.2	44.6	Classification: *Other*—by lidar Path prediction: *Static*; not on path of SUV	Lidar detects unknown object. Object is considered new, tracking history is unavailable, and velocity cannot be determined. ADS predicts object's path as static.
-4.2	44.8	Classification: *Vehicle*—by lidar Path prediction: *Static*; not on path of SUV	Lidar classifies detected object as *vehicle*; this is a changed classification of object and without a tracking history. ADS predicts object's path as static.
-3.9[c]	44.8	Classification: *Vehicle*—by lidar Path prediction: Left through lane (next to SUV); not on path of SUV	Lidar retains classification *vehicle*. Based on tracking history and assigned goal, ADS predicts object's path as traveling in left through lane.
-3.8 to -2.7	44.7	Classification: alternates between *vehicle* and *other*—by lidar Path prediction: alternates between *static* and left through lane; neither considered on path of SUV	Object's classification alternates several times between *vehicle* and *other*. At each change, tracking history is unavailable; ADS predicts object's path as static. When detected object's classification remains same, ADS predicts path as traveling in left through lane.
-2.6	44.6	Classification: *Bicycle*—by lidar Path prediction: *Static*; not on path of SUV	Lidar classifies detected object as *bicycle*; this is a changed classification of object and object is without a tracking history. ADS predicts bicycle's path as static.
-2.5	44.6	Classification: *Bicycle*—by lidar Path prediction: Left through lane (next to SUV); not on path of SUV	Lidar retains *bicycle* classification; based on tracking history and assigned goal, ADS predicts bicycle's path as traveling in left through lane.

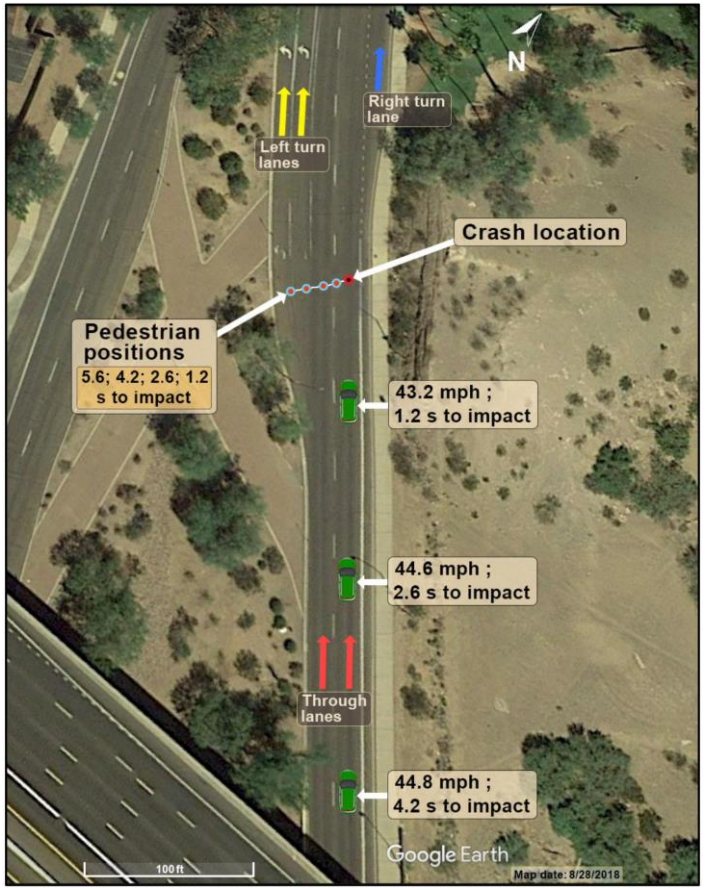
Source: National Transportation Safety Board. Collision between vehicle controlled by developmental automated driving system and pedestrian Tempe, Arizona march 18, 2018. 2019.

Technical

Failure of system to correctly detect pedestrian and avoid collision

Case Study – Uber Tempe incident

Interacting layers of complexity and uncertainty



Time to Impact (seconds)	Speed (mph)	Classification and Path Prediction[a]	Vehicle and System Actions[b]
-9.9	35.1	--	Vehicle begins to accelerate from 35 mph in response to increased speed limit.
-5.8	44.1	--	Vehicle reaches 44 mph.
-5.6	44.3	Classification: *Vehicle*—by radar Path prediction: *None*; not on path of SUV	Radar makes first detection of pedestrian (classified as vehicle) and estimates speed.
-5.2	44.6	Classification: *Other*—by lidar Path prediction: *Static*; not on path of SUV	Lidar detects unknown object. Object is considered new, tracking history is unavailable, and velocity cannot be determined. ADS predicts object's path as static.
-4.2	44.8	Classification: *Vehicle*—by lidar Path prediction: *Static*; not on path of SUV	Lidar classifies detected object as *vehicle*; this is a changed classification of object and without a tracking history. ADS predicts object's path as static.
-3.9[c]	44.8	Classification: *Vehicle*—by lidar Path prediction: Left through lane (next to SUV); not on path of SUV	Lidar retains classification *vehicle*. Based on tracking history and assigned goal, ADS predicts object's path as traveling in left through lane.
-3.8 to -2.7	44.7	Classification: alternates between *vehicle* and *other*—by lidar Path prediction: alternates between *static* and left through lane; neither considered on path of SUV	Object's classification alternates several times between *vehicle* and *other*. At each change, tracking history is unavailable; ADS predicts object's path as static. When detected object's classification remains same, ADS predicts path as traveling in left through lane.
-2.6	44.6	Classification: *Bicycle*—by lidar Path prediction: *Static*; not on path of SUV	Lidar classifies detected object as *bicycle*; this is a changed classification of object and object is without a tracking history. ADS predicts bicycle's path as static.
-2.5	44.6	Classification: *Bicycle*—by lidar Path prediction: Left through lane (next to SUV); not on path of SUV	Lidar retains *bicycle* classification; based on tracking history and assigned goal, ADS predicts bicycle's path as traveling in left through lane.

Source: National Transportation Safety Board. Collision between vehicle controlled by developmental automated driving system and pedestrian Tempe, Arizona march 18, 2018. 2019.

Human factors

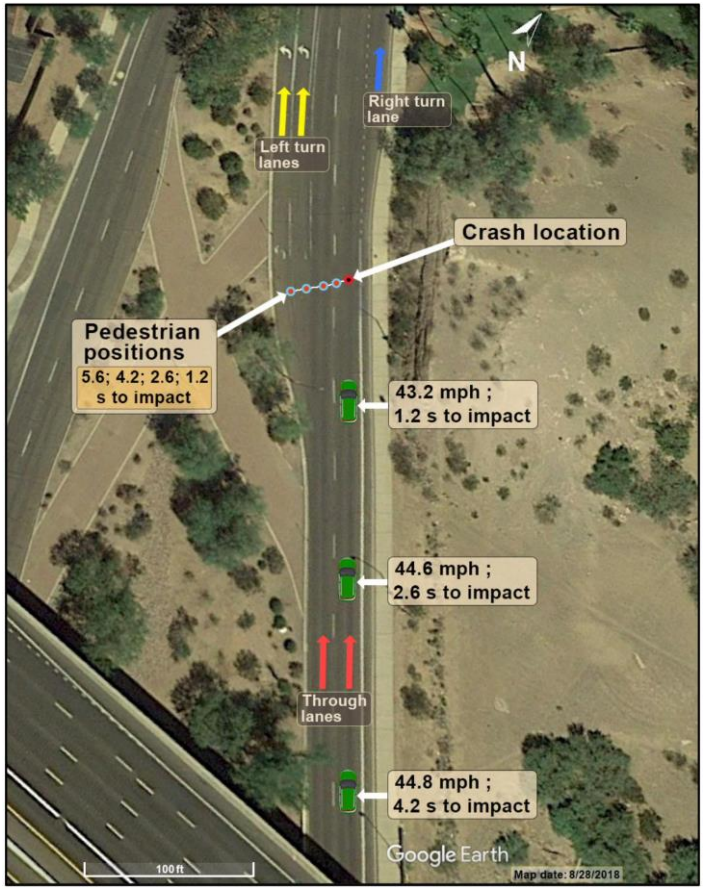
Failure of safety driver to detect that system was not operating correctly

Technical

Failure of system to correctly detect pedestrian and avoid collision

Case Study – Uber Tempe incident

Interacting layers of complexity and uncertainty



Time to Impact (seconds)	Speed (mph)	Classification and Path Prediction[a]	Vehicle and System Actions[b]
-9.9	35.1	--	Vehicle begins to accelerate from 35 mph in response to increased speed limit.
-5.8	44.1	--	Vehicle reaches 44 mph.
-5.6	44.3	Classification: *Vehicle*—by radar Path prediction: *None*; not on path of SUV	Radar makes first detection of pedestrian (classified as vehicle) and estimates speed.
-5.2	44.6	Classification: *Other*—by lidar Path prediction: *Static*; not on path of SUV	Lidar detects unknown object. Object is considered new, tracking history is unavailable, and velocity cannot be determined. ADS predicts object's path as static.
-4.2	44.8	Classification: *Vehicle*—by lidar Path prediction: *Static*; not on path of SUV	Lidar classifies detected object as *vehicle*; this is a changed classification of object and without a tracking history. ADS predicts object's path as static.
-3.9[c]	44.8	Classification: *Vehicle*—by lidar Path prediction: Left through lane (next to SUV); not on path of SUV	Lidar retains classification *vehicle*. Based on tracking history and assigned goal, ADS predicts object's path as traveling in left through lane.
-3.8 to -2.7	44.7	Classification: alternates between *vehicle* and *other*—by lidar Path prediction: alternates between *static* and left through lane; neither considered on path of SUV	Object's classification alternates several times between *vehicle* and *other*. At each change, tracking history is unavailable; ADS predicts object's path as static. When detected object's classification remains same, ADS predicts path as traveling in left through lane.
-2.6	44.6	Classification: *Bicycle*—by lidar Path prediction: *Static*; not on path of SUV	Lidar classifies detected object as *bicycle*; this is a changed classification of object and object is without a tracking history. ADS predicts bicycle's path as static.
-2.5	44.6	Classification: *Bicycle*—by lidar Path prediction: Left through lane (next to SUV); not on path of SUV	Lidar retains *bicycle* classification; based on tracking history and assigned goal, ADS predicts bicycle's path as traveling in left through lane.

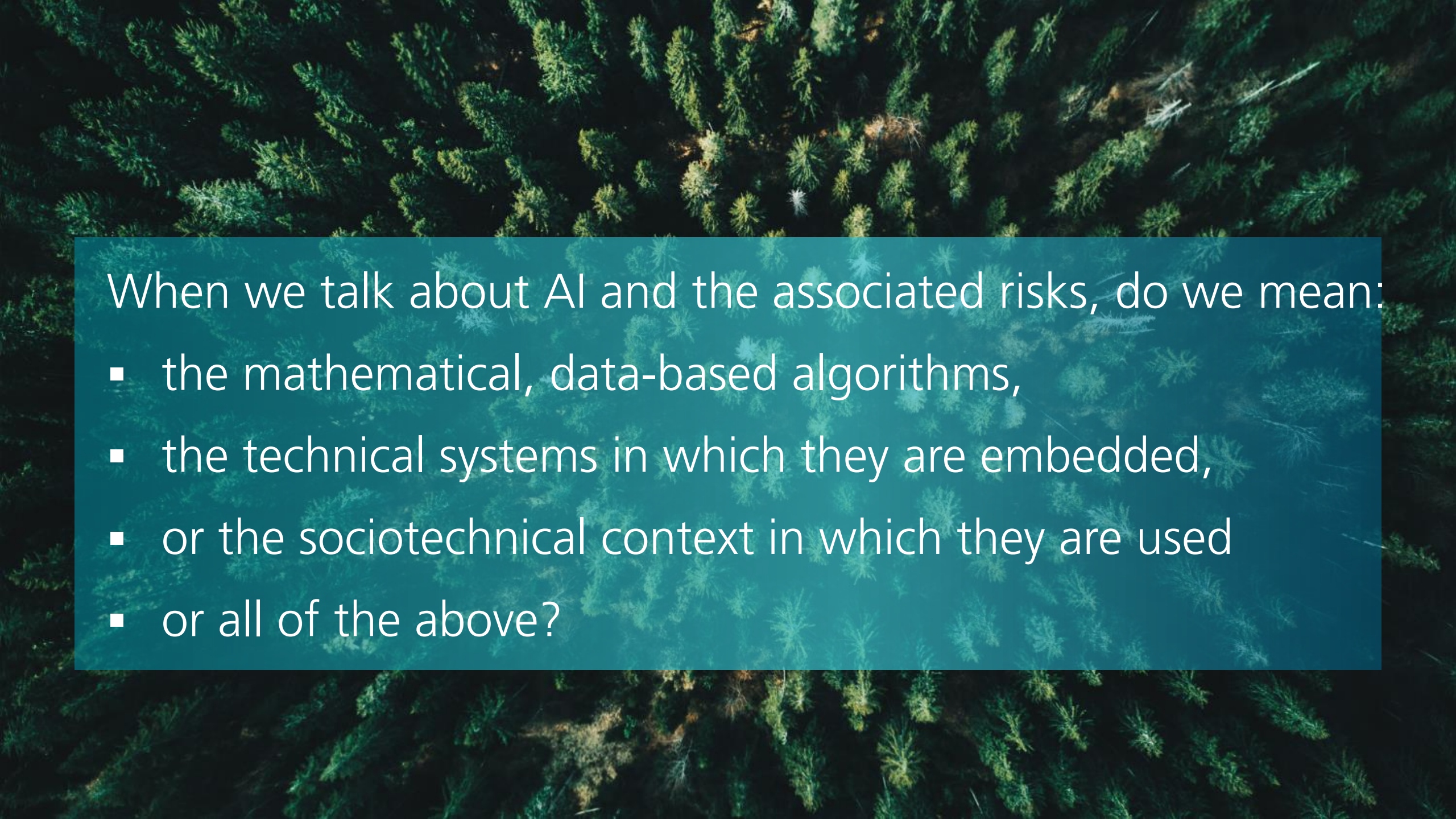
Source: National Transportation Safety Board. Collision between vehicle controlled by developmental automated driving system and pedestrian Tempe, Arizona march 18, 2018. 2019.

Governance — Failure to regulate accountability for safety of automated driving

Management — Inadequate engineering and operating processes, lack of oversight of safety driver

Human factors — Failure of safety driver to detect that system was not operating correctly

Technical — Failure of system to correctly detect pedestrian and avoid collision



When we talk about AI and the associated risks, do we mean:

- the mathematical, data-based algorithms,
- the technical systems in which they are embedded,
- or the sociotechnical context in which they are used
- or all of the above?

Emergent complexity and **uncertainty** in the AI system and its environment leads to unpredictable behavior and difficulties in defining conditions for trustworthiness, moral responsibility and legal liability

Closing the gap between expectations and reality requires a holistic view on the system and a continuous re-evaluation of the risk from technical, human factors, operational and regulatory perspectives


Closing the gaps between expectations and reality…

Closing the gap between expectations and reality

EU Artificial Intelligence Act

Objectives:

- Ensure that AI systems are safe and respect existing law on fundamental rights and union values
- Enhance governance of existing law on fundamental rights and safety applicable to AI systems
- Ensure legal certainty to facilitate investment and innovation in AI
- Facilitate the development of a single market for lawful, safe and trustworthy market

EUROPEAN
COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

EU Artificial Intelligence Act:
https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

Closing the gap between expectations and reality

Regulation and assurance gaps

Does the system fulfill all the technical criteria required to be considered trustworthy?

What impact will the system have on overall risk for a given operational domain?

*Source: Ethics Guidelines for trustworthy AI, Independent high-level expert group on Artificial Intelligence, EU Commission, 2019

Closing the gap between expectations and reality

The role of Standards and Norms

- Product liability laws require manufacturers to ensure safety in accordance with the state-of-the-art
- Published international standards and norms are considered to define the lower bound of state-of-the-art
- Assessment and certifications are a fundamental part of safety regulations
- Assessments are typically performed against criteria defined by the standards

Closing the gap between expectations and reality

Realistic expectations on standards

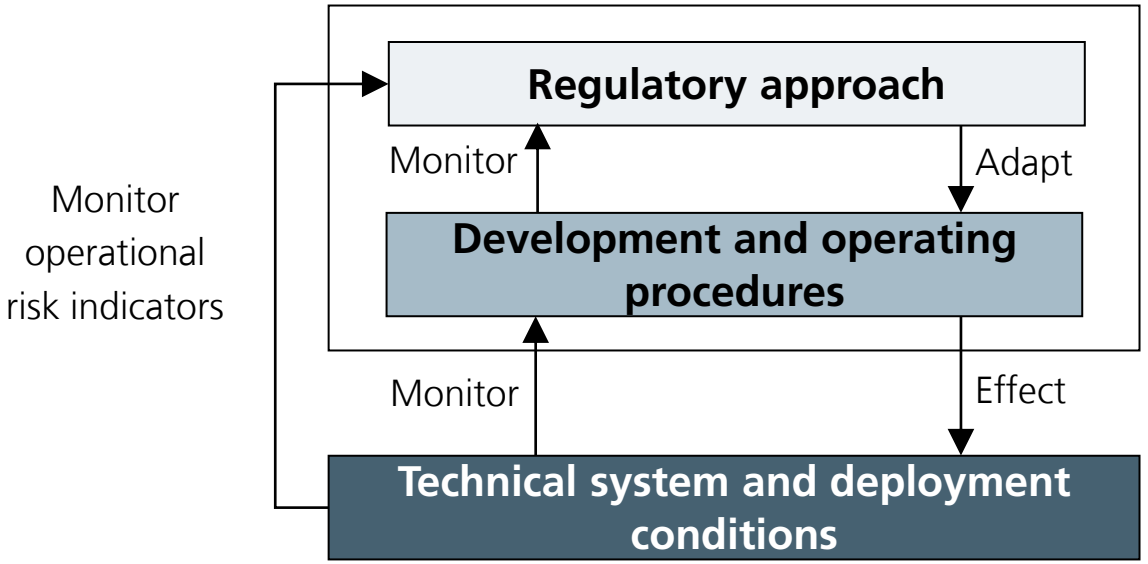
- International standards can only document a consensus of state-of-the-art and best practice
- It takes time for state-of-the-art to be established and standards to be written
- There is still much uncertainty regarding whether and how AI/ML can be made “safe”
- The field of AI and machine learning is developing very quickly, stable requirements should therefore not be expected in the near term
- This has a consequence on the safety assessment and certification of AI-based systems

Closing the gap between expectations and reality

Pragmatic next steps

Deliberate and planned bootstrapping approaches should be taken to increasing operational context and functional scope, whilst monitoring impact of complexity and uncertainty in AI systems

- Requires a calibrated level of tolerable residual risk
- Observation points must be defined to act as early warning indicators for increased risk/uncertainty
- Should be considered with the phased introduction of regulation and standards, e.g. for automated driving (ALKS, Highway Chauffer, Delivery Drones,…)
- Ensure an agile approach to regulation, e.g. making use of regulatory sandboxes* to learn from pilot projects



*https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf

Safe Intelligence

The path to trustworthy AI-based systems

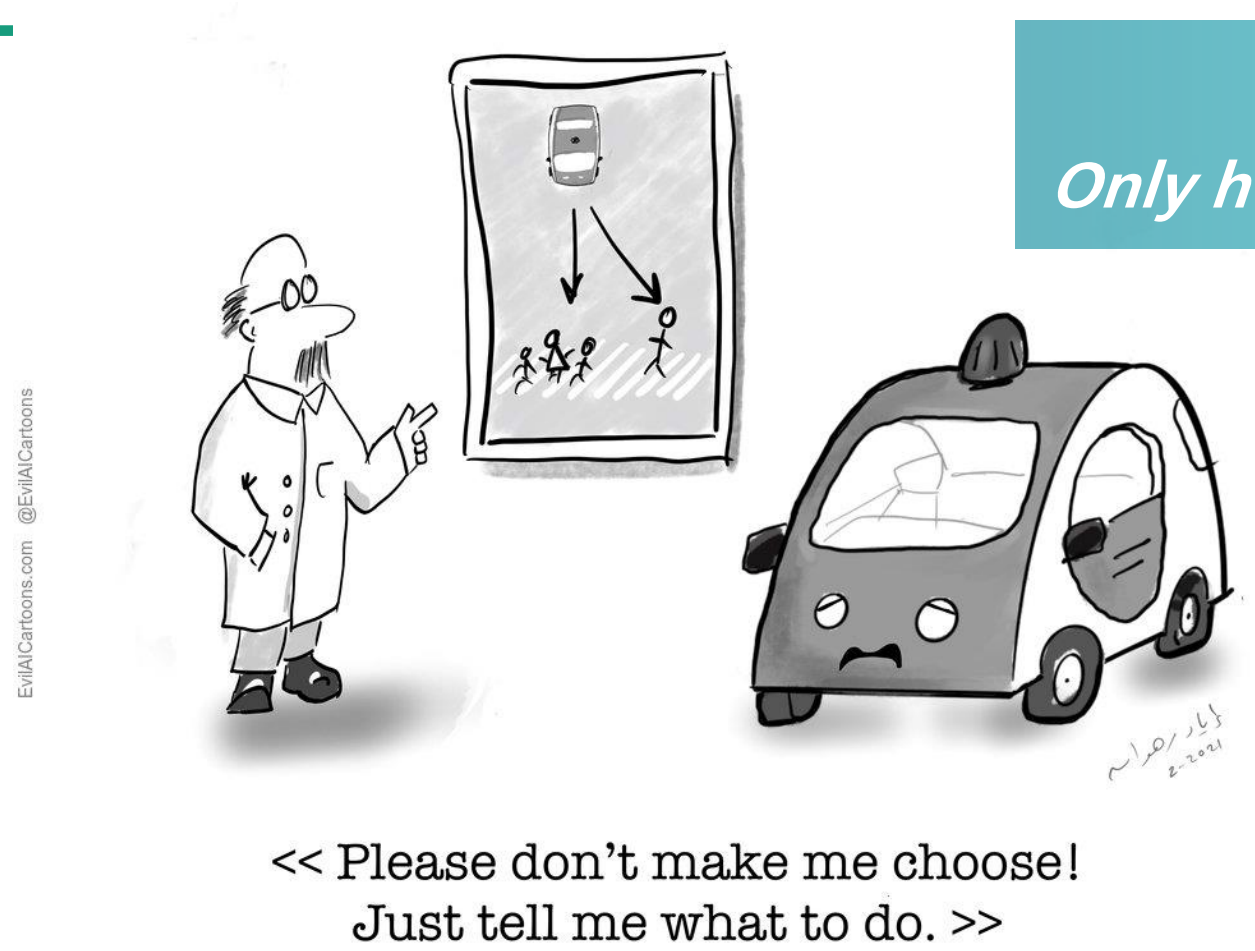
More **research** into **safe and trustworthy AI** that is reliable enough for critical applications (avoid a theoretical AI research bubble!)

Holistic consideration of technical as well as **ethical aspects** of trustworthiness in the design of AI systems (**how safe is safe enough?)**

Standardisation needed to provide consistent set of (application-specific) technical criteria for evaluating the trustworthiness of AI-based systems

Agile regulatory approaches (e.g. Regulatory sandboxes) required in order to develop effective technical, legal and organizational conditions for effective control of risk

…and lastly



AI is just mathematics -

Only humans can solve ethical dilemmas

https://www.evilaicartoons.com/archive/only-humans-can-solve-ethical-dilemmas

Contact

Prof. Simon Burton
Research Division Director, Safety Assurance
Tel. +49 89 547088-341
simon.burton@iks.fraunhofer.de

Fraunhofer IKS
Hansastrasse 32
80686 München
www.iks.fraunhofer.de