CYBERSECURITY CHALLENGES IN THE UPTAKE OF ARTIFICIAL INTELLIGENCE IN AUTONOMOUS DRIVING

A collaboration between ENISA and the Joint Research Centre, European Commission

Ronan Hamon

C-ITS Webinar
27/04/2021

REPORT

CYBERSECURITY CHALLENGES IN THE UPTAKE OF ARTIFICIAL INTELLIGENCE IN AUTONOMOUS DRIVING

#EUcybersecurity

enisa
EUROPEAN UNION AGENCY FOR CYBERSECURITY

European Commission

https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/cybersecurity-challenges-uptake-artificial-intelligence-autonomous-driving

The Joint Research Centre (JRC)

As the science and knowledge service of the European Commission our mission is to support EU policies with independent evidence throughout the whole policy cycle.

The JRC
Facts and figures

JRC

Policy neutral: has no policy agenda of its own, but works with over 20 EC policy departments

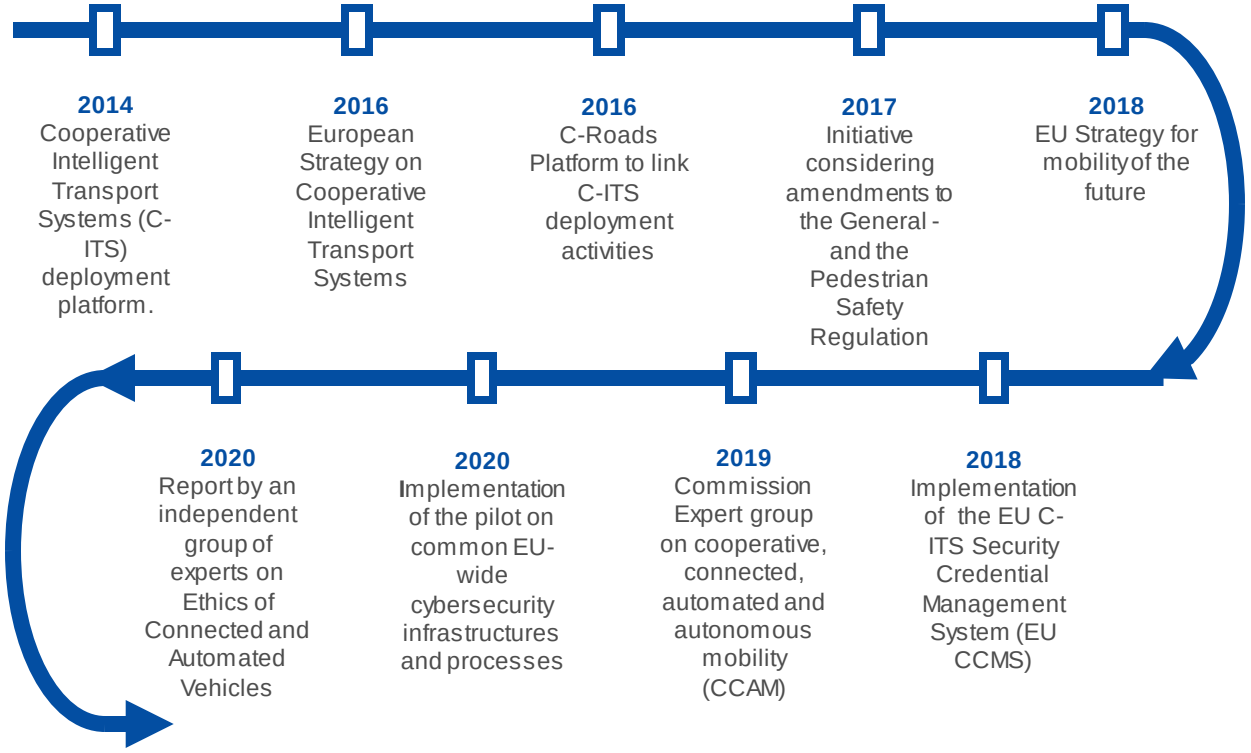
More than **50 large scale research facilities**
More than **110 online databases**

About **2 800 staff,** nearly **70 %** of whom are **scientific/technical staff**

83 % of core **research staff with PhDs**

Over **1 400 scientific publications per year**

Policy context – Autonomous Driving European level



European Commission

2016
Network and Information Security directive

2018
General Data Protection Regulation

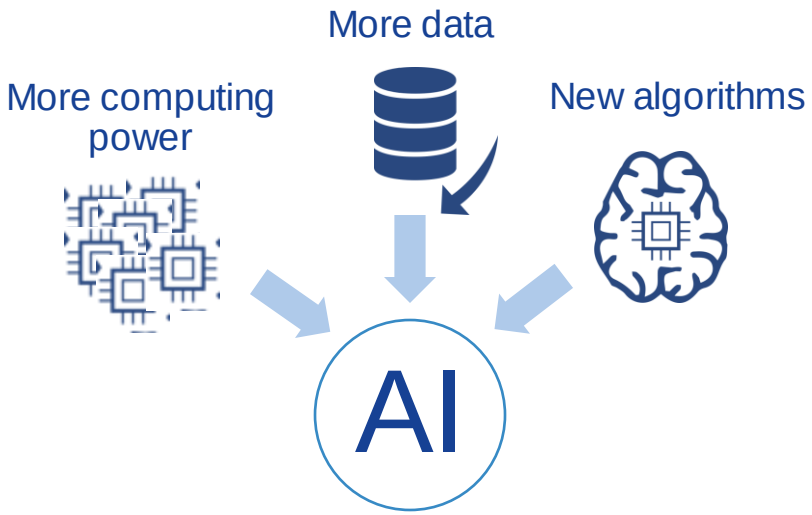
European Commission

AI in autonomous driving - Definition of artificial intelligence

HLEG definition of AI

"Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, **act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take** (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions."

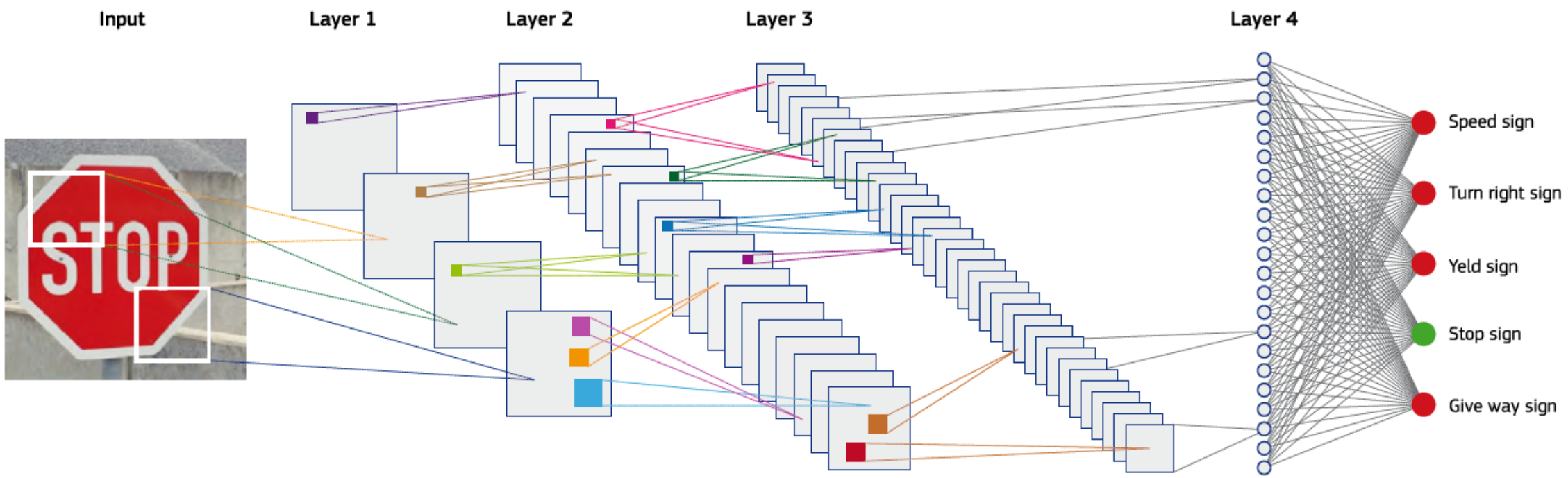
European Commission's High-level expert group on Artificial Intelligence, from "*A definition of AI: main capabilities and scientific disciplines*", 18 December 2018



When using machine learning, human developers no longer program an algorithm to ***tell*** **the computer how to solve** a given problem but instead they **program it to teach the computer to *learn how* to solve** the problem.

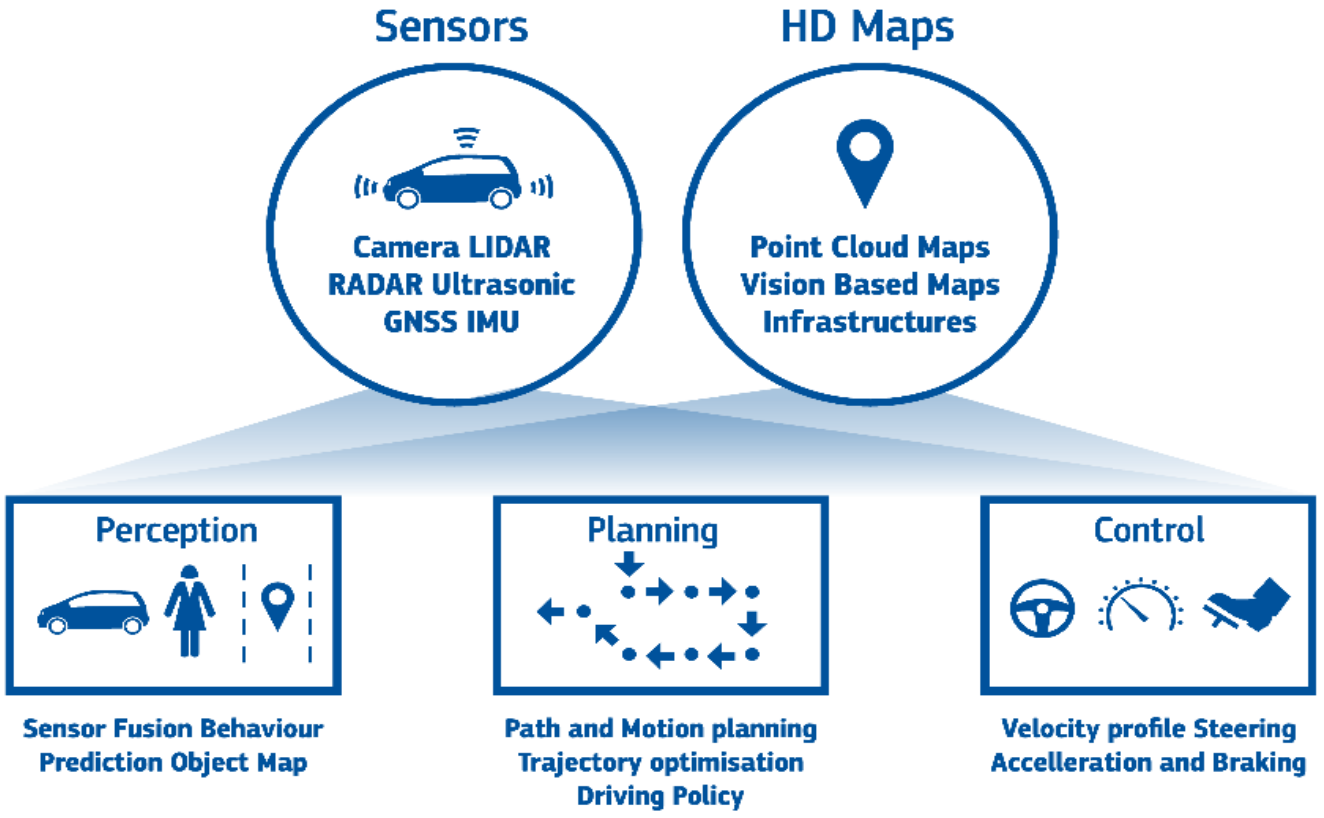
European Commission

AI in autonomous driving systems – Complexity of deep learning

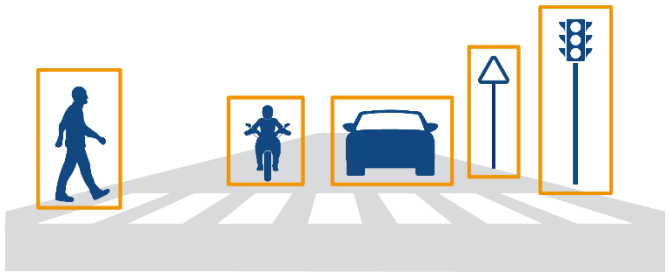


European Commission

AI in autonomous driving – Components

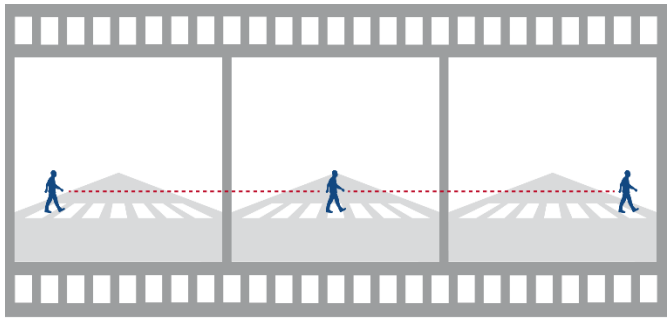


AI in autonomous driving – Perception



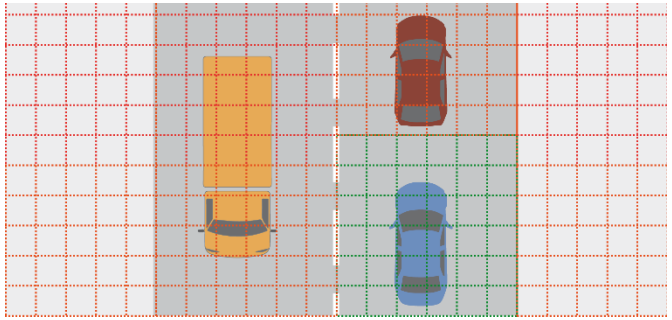
Scene Understanding

- Identification of roads and lanes
- Detection of moving agents and objects
- Traffic signs and markings recognition
- Sound event classification



Scene Flow Estimation

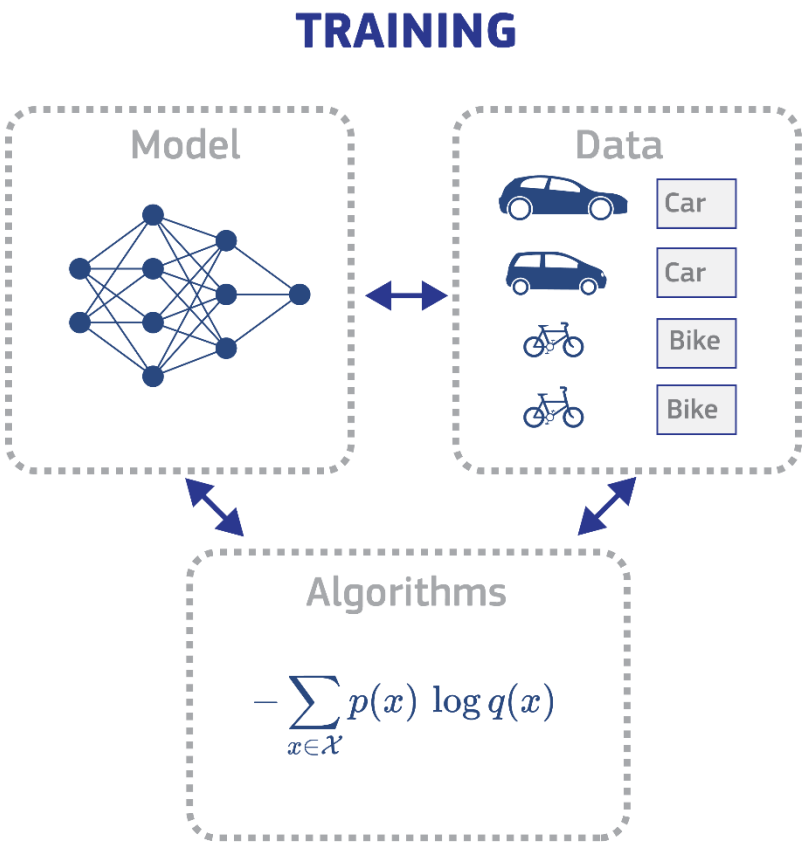
- Tracking and prediction of objects, moving agents and obstacles

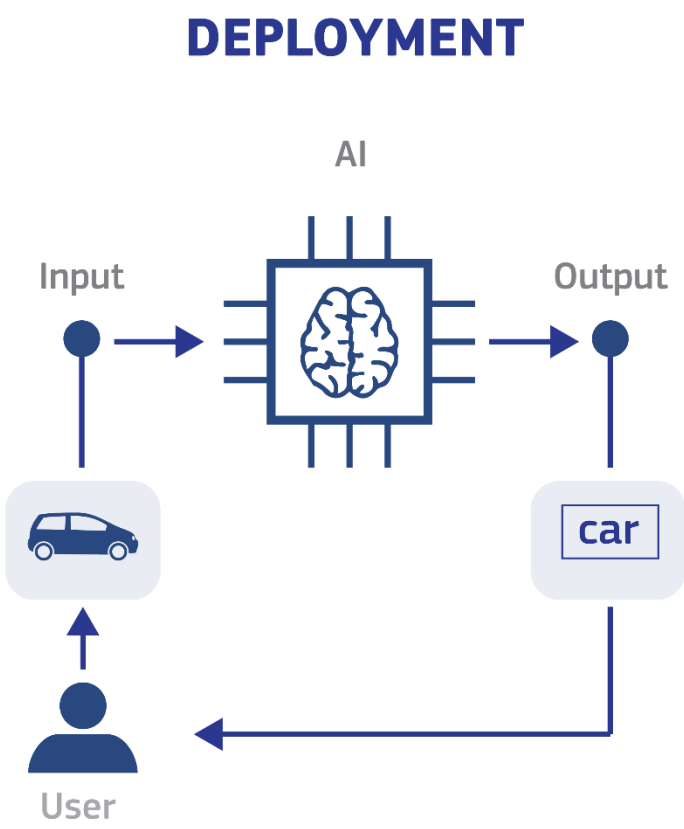


Scene Representation

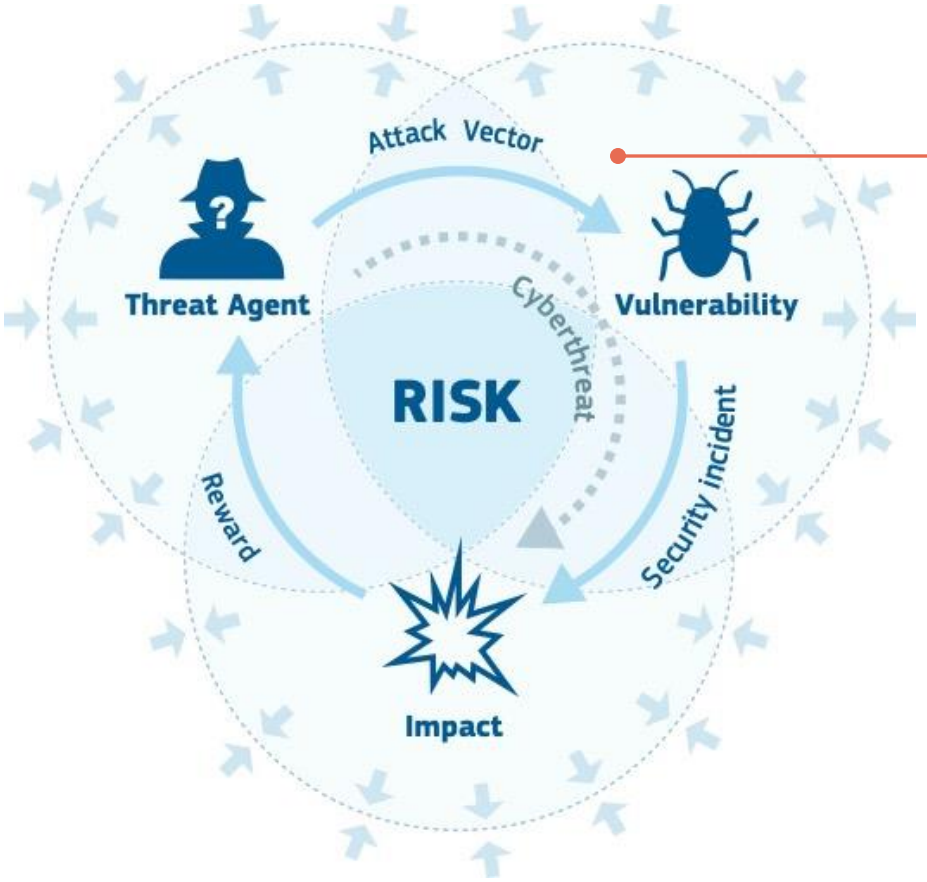
- Localization
- Occupancy Maps and Grids

Cybersecurity of AI in Autonomous Driving – Normal use





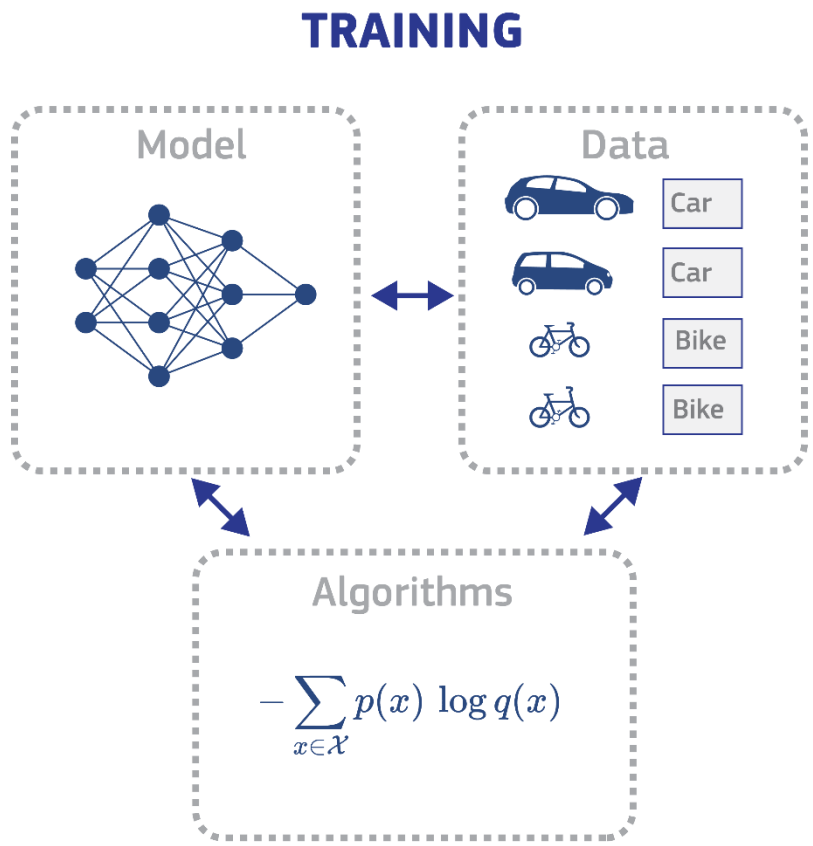
Cybersecurity of AI in Autonomous Driving – A risk based approach to AI

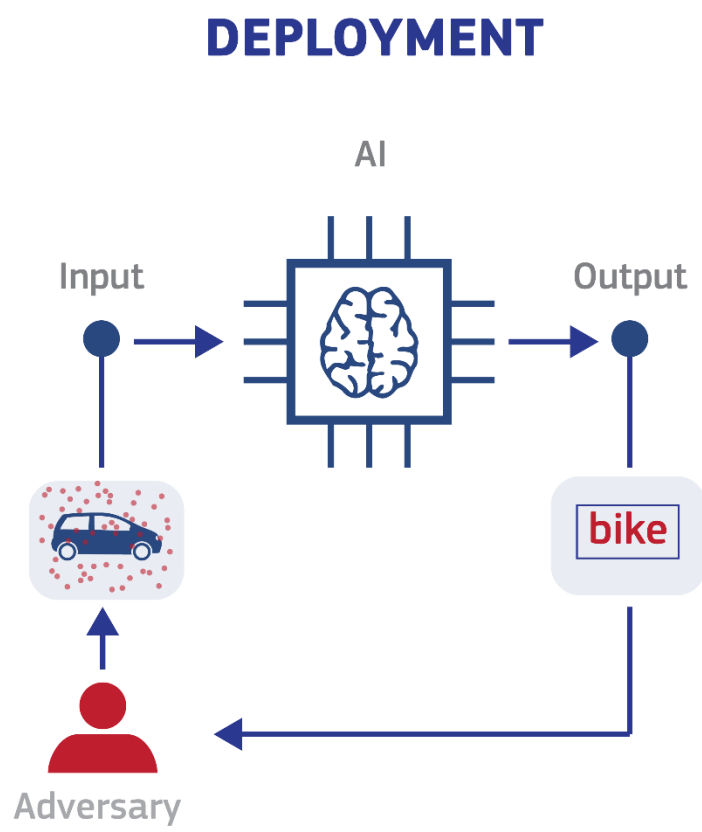


Conceptual model depicting the logical links between the different components of the cybersecurity risk in the context of the influence of AI and Digital Transformation

- Adversarial attacks
- Data poisoning
- Data leakage
- Model theft
- Backdoors

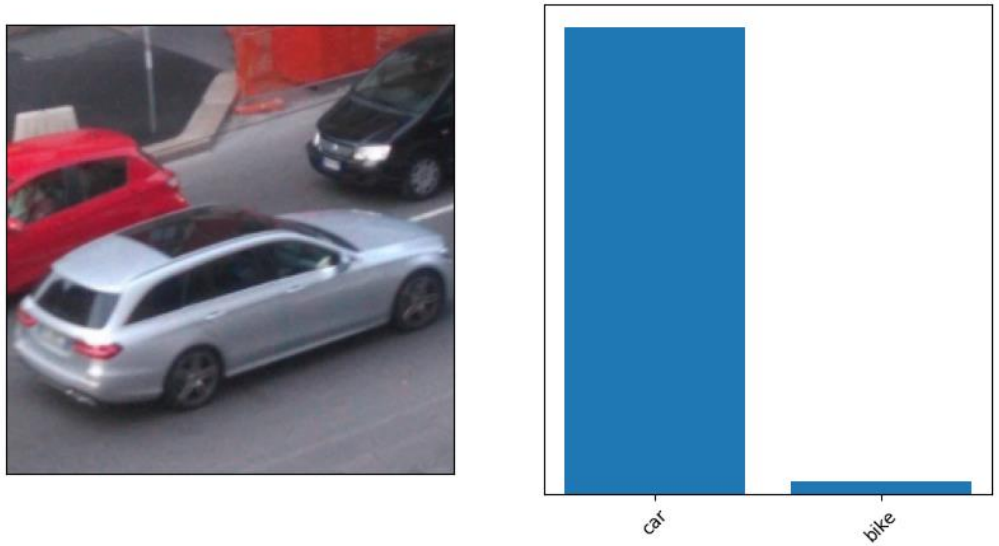
Cybersecurity of AI in Autonomous Driving – Adversarial Attack



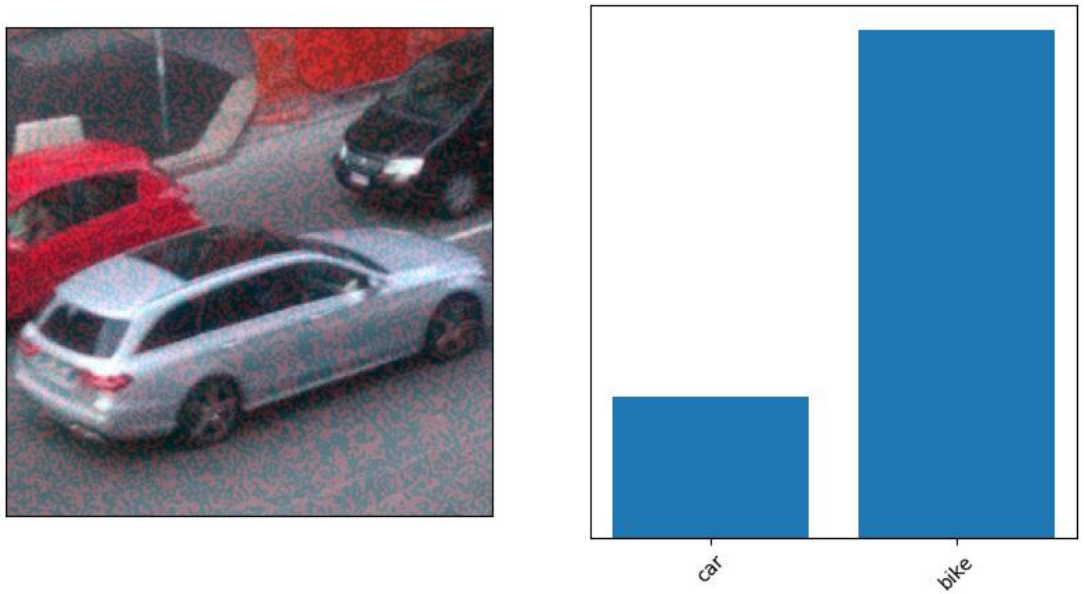


Cybersecurity of AI in Autonomous Driving – Adversarial Attacks

Original image

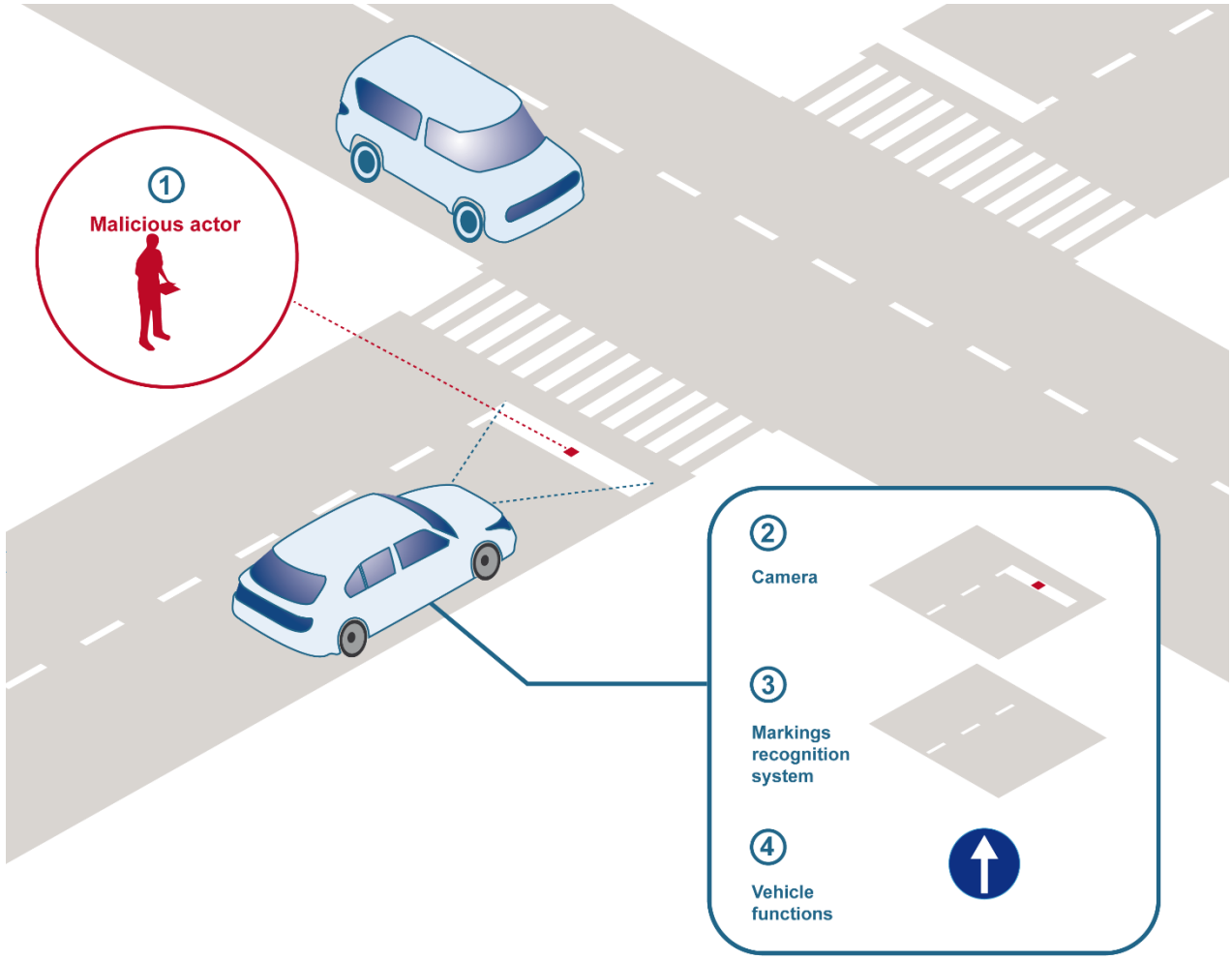


Adversarial image



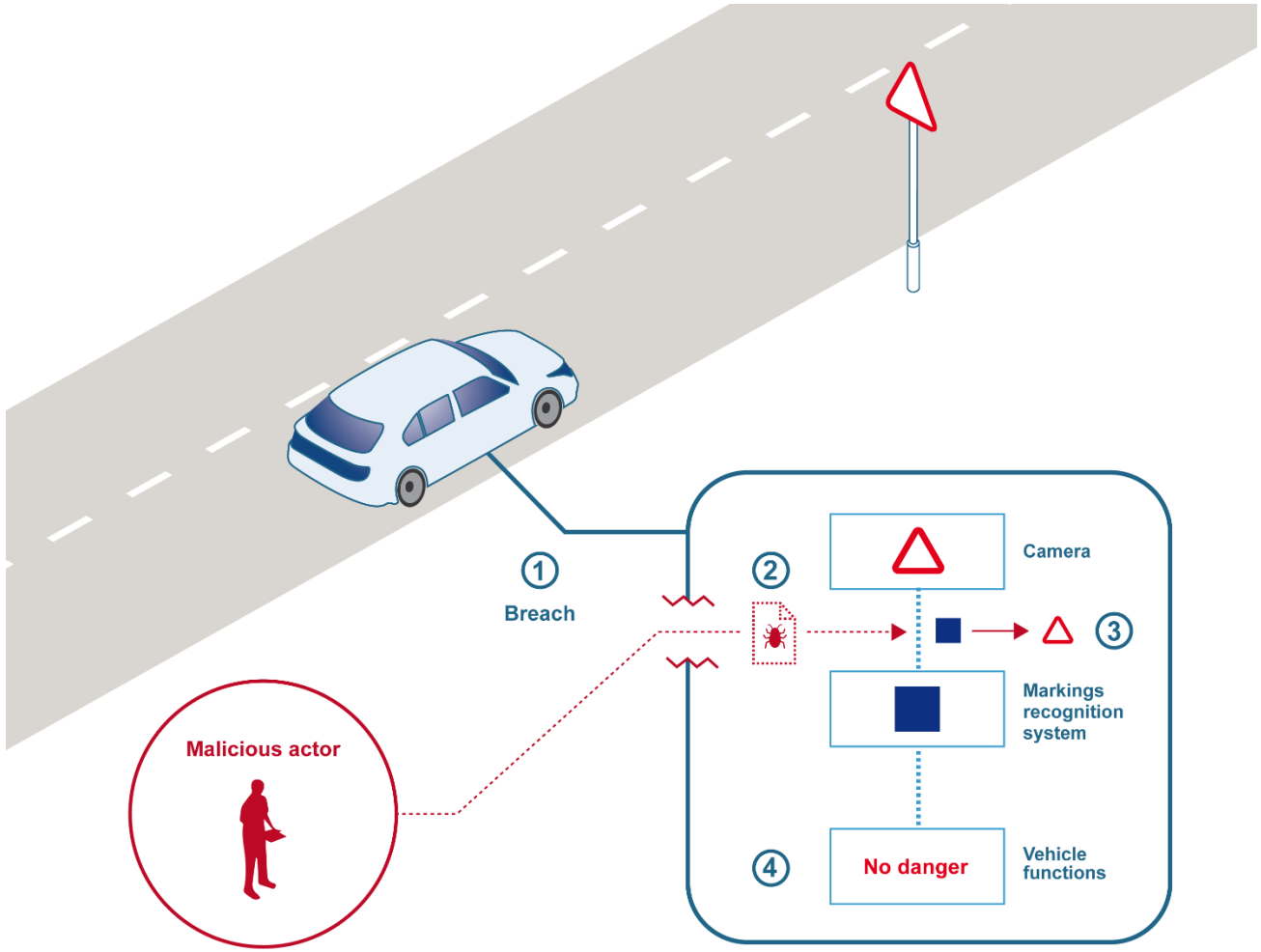
Experiment done using the Resnet-50 model pretrained on ImageNet dataset. 'car' corresponds to label 'sport car', 'bike' to label 'mountain bike'. The adversarial perturbation is constrained to be in the red channel, with high intensity dots.

Scenario 1: Adversarial attacks on street markings



1. A malicious actor applies a sticker with physical perturbations onto a stop marking.
2. The camera of the AV sees the stop markings and the sticker.
3. The markings recognition system is deceived into perceiving the stop marking.
4. The planning and control systems handle the situation as if there was no stop.

Scenario 2: Man-in-the-Middle Attack on the AI modules



1. A malicious actor gains access to the vehicle's ICT system by exploiting a vulnerability.

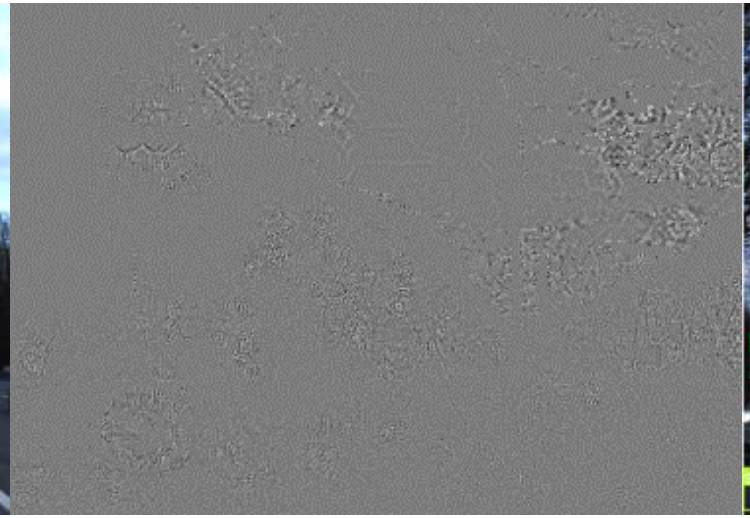
2. This allows the actor to move freely within the system. They can also tamper with the AI functions.

3. In this case a small perturbation is introduced into the pipeline from the camera to the recognition systems, designed to deceive into not recognizing danger signs.

4. The planning and control systems handle the situation as if there was no danger.

Examples of adversarial attacks – Overflow attack

YOLOv5 object detector:
Original image



Adversarial Perturbation

YOLOv5 object detector:
Adversarial image

European Commission

Examples of adversarial attacks – Spoofing attack

YOLOv5 object detector:
Original image

YOLOv5 object detector:
Adversarial image

Cybersecurity of AI in Autonomous Driving – Some real world attack example cases

2019 the *Harman International, Automotive Security Business Unit* demonstrated a complete pipeline to produce physical adversarial stickers to fool commercial AV perception systems under real conditions.

N. Morgulis, A. Kreines, S. Mendelowitz, and Y. Weisglass, 'Fooling a Real Car with Adversarial Traffic Signs', preprint arxiv: 1907.00374, 2019.

2019 Tencent researchers tested hacking remotely a Tesla car and attacking the AI-based autopilot systems only with access to the outputs of various neural network models e.g. attacking the lane detection assistance system to turn the car into the reverse lane.

Tencent Keen Security Lab, Experimental Security Research of Tesla Autopilot, 2019

2019-2020 McAffee demonstrating fooling the Tesla autopilot (level 2 car) with a simple attack consisting of an elongated tape on a speed limit sign.

S. Povolny and S. Trivedi, 'Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles', McAfee Blogs, 2020.

2019-2020 McAffee researchers successful attack the ADAS Mobileye camera by injecting spoofed traffic signs using a drone with a projector.

D. Nassi, R. Ben-Netanel, Y. Elovici, and B. Nassi, 'MobilBye: Attacking ADAS with Camera Spoofing', preprint arxiv: 1906.09765, 2019.

Recommendations

1. Systematic security validation of AI models and data
2. Supply chain challenges related to AI cybersecurity
3. End-to-end holistic approach for integrating AI cybersecurity with traditional cybersecurity principles
4. Incident handling and vulnerability discovery related to AI and lessons learned
5. Limited capacity and expertise on AI cybersecurity in the automotive industry

Thank you for your attention

EU Science Hub: ec.europa.eu/jrc

European Commission

Keep in touch

EU Science Hub: ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub – Joint Research Centre

EU Science, Research and Innovation

EU Science Hub