TRUSWORTHY EMBEDDED AI

RISK ANALYSIS AND CERTIFICATION FRAMEWORKS FOR CRITICAL TRUSTED AI APPLICATIONS

Morayo Adedjouma

Lab for Design of Embedded & Autonomous Systems (CEA LIST/ DILS/LSEA)

Agenda

1. **Foundations evolution in the light of AI**
   - Critical applications using AI: what, how, example
   - The problems they poses for risk assessment and qualification
   - New foundations and their challenges

… The future is already (nearly) there

Critical AI systems are everywhere, more and more involved in our daily lives.

Les cobots légers se multiplient

PATRICE DESMEDT | INDUSTRIE 4.0, ROBOTIQUE | PUBLIÉ LE 22 AVRIL 2015 À 16H42

TWITTER FACEBOOK LINKEDIN GOOGLE + EMAIL

Sur la Hannover Messe, ABB, Kuka et Universal Robots montraient leurs petits robots collaboratifs. Ces cobots légers travaillent en bonne intelligence avec l'homme et s'adaptent facilement à de nombreuses tâches.

The Economist explains

How swarm drones are mimicking nature

Oct 1st 2015, 23:00 BY K.B. | Timekeeper

THE notion of autonomous drones can conjure up dreams of easy, efficient parcel delivery—and nightmares of algorithms taking the place of human judgment in warfare.

Collaborating with a cobot

Program them to do multiple tasks and deploy them throughout the plant when they're needed.

December 24, 2014
by PLANT Staff

Topics
General
Technology
Industries
Automotive
Manufacturing
Tags

A LBR iiwa, a robotic assistant, working collaboratively with humans to increase productivity in manufacturing and materials handling. PHOTO: KUKA SYSTEMS.

LBR iiwa wants to get up close and personal. Its name, a pairing of German and English abbreviations, stands for "lightweight robot, intelligent industrial work assistant." One of a new class of small collaborative robots – called cobots – it's designed primarily to serve as a robotic assistant for a human worker and has the potential to change how work is structured.

… The future is already (nearly) there (cont'd)

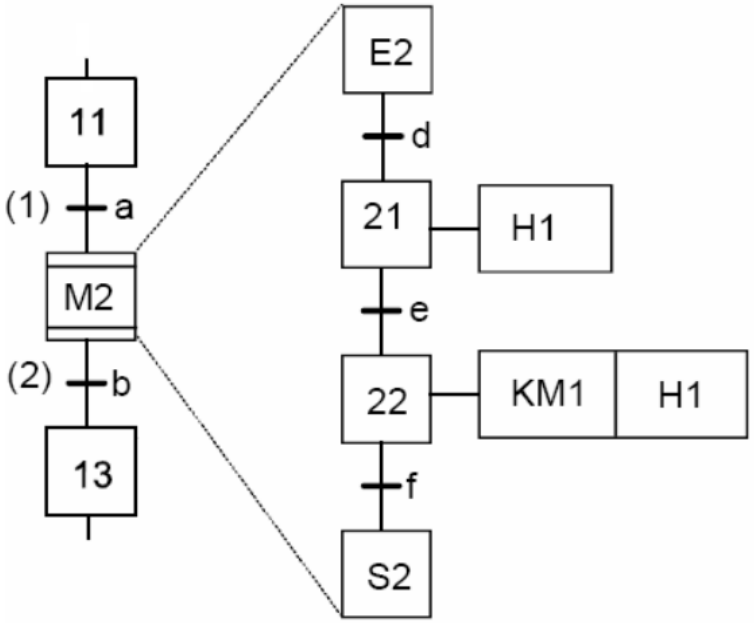
- **Autonomous CPS**
  - Autonomy refers to the degree of freedom a system has regarding potential activities.

- **Autonomy of decision**: degree of freedom allocated to the system when deciding.
  - For example, it can be associated with a set of constraints on a search space. The reduction of this degree by choosing one possibility constitutes the act of decision, using optimization tools.

- **Autonomy of action**: concerns the ability to act
  - For example on the real world, through actuators or the digital world through the sending of decisions to apply by others.

… The future is already (nearly) there (cont'd)

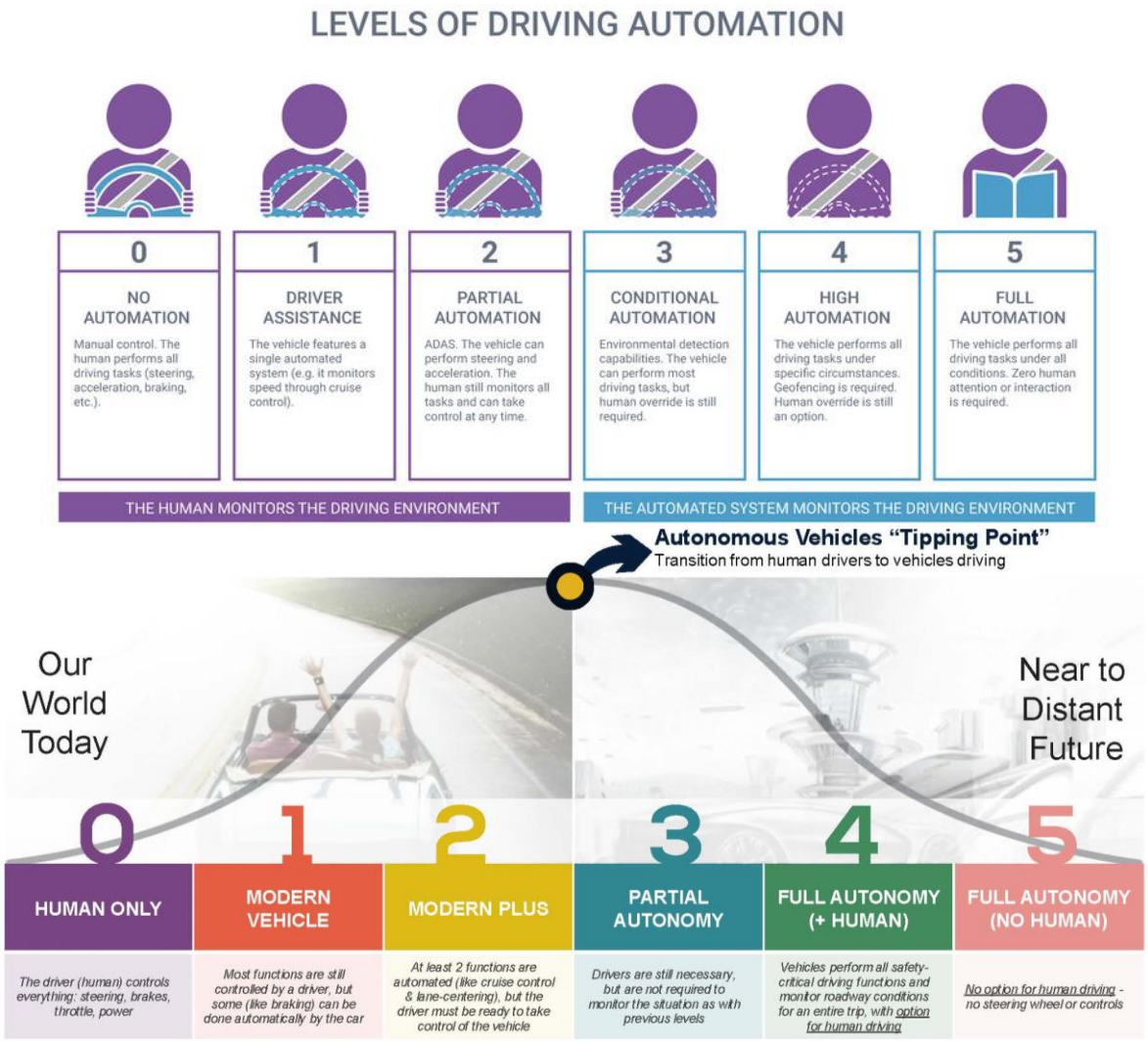
- **Intelligent/learning CPS**
  - Assuming at least a certain level of decisional autonomy opens the possibility for a CPS to learn and adapt its decision with time and with its experience and history
  - A **non learning system** will always generate the same outputs for the same set of inputs, whatever the moment
  - A **learning system** is a system that may generate different but improved outputs for the same inputs at different moments



Adapted from (Mitchell, 1997)

… The future is already (nearly) there (cont'd)

- Intelligent/learning CPS level of autonomy



Example #1: the cooperating train (CPS)

SurferLab — Distributed Intelligence for Transportation Systems Laboratory

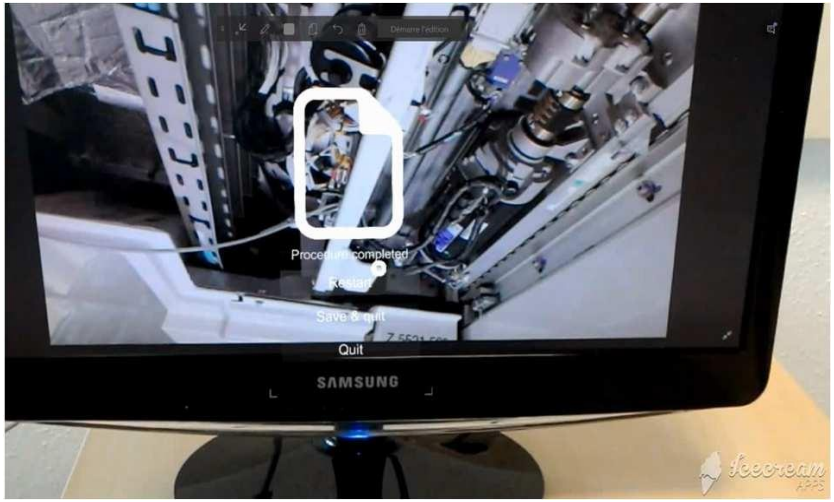
BOMBARDIER

Prosyst — Master of system automation

Université Polytechnique HAUTS-DE-FRANCE

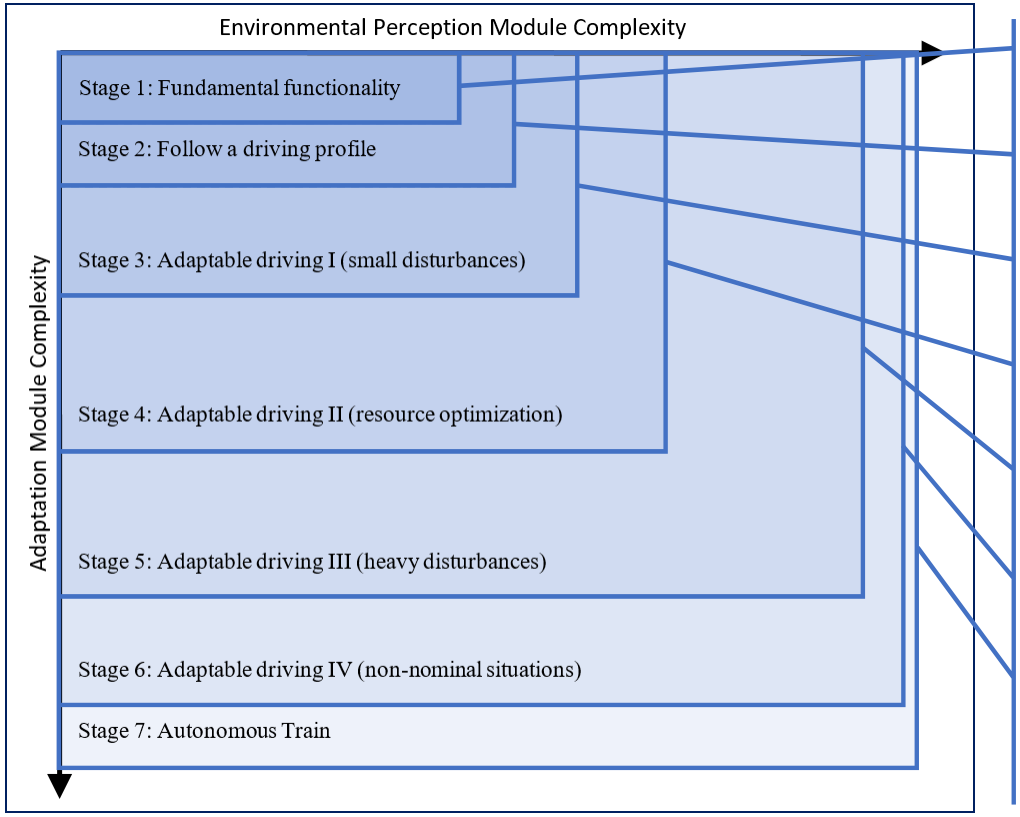
Concept of « peer to peer » cooperation

An "intelligent" train detects the maintenance operator and warns him, using embedded behavioral models of a risk about its health status. This interaction is done using augmented virtual reality systems (hololens, tablets) applied here to the opening time cycle of a door.



l'Europe s'engage en Nord Pas de Calais - Picardie avec le FEDER — UNION EUROPÉENNE — cnrs — LAMIH UMR CNRS 8201 — I-Trans — INSTITUT CARNOT ARTS

Example #2: the autonomous train (CPS)



- Start, maintain constant speed, stop.

-Acceleration and braking modes. (precise positioning and speed)

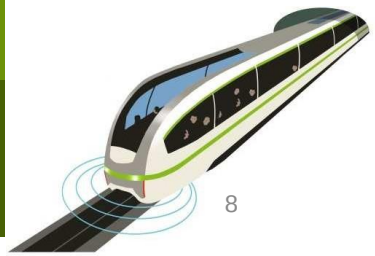
-Timing-related disturbances (diving profile modifications, LTV,

- Itinerary alterations, performance optimizations (energy consumption, passenger

-Disturbances caused by the dynamic environment (movable

-Semi-autonomous operation (simulation of unlikely situations,

-Fully autonomous driverless operation



But a frigthening future for researchers and engineers

AI Systems failure may result in death or serious injury to people, or damage to equipment or environmental harm.

« A French researcher is being sued for murderer: a robot killed an operator after having learnt from his learning algorithm ! »

The French researcher


What We Know About the Bomb Robot Used to Kill the Suspected Dallas Shooter [UPDATE]

Darren Orf

Northrop Grumman Andros bomb disposal bot in Dallas in 2015. Image: Stewart F. House/Getty Images

Terminator redux? Robot kills a man at Haryana's Manesar factory

Rao Jaswant Singh & Sanjay Yadav | TNN | Aug 13, 2015, 04.39 AM IST

A- A+

GURGAON: This one's straight out of a Terminator film. Sharp welding sticks jutting out of the robotic arm of a machine pierced a worker killing him at a factory here on Wednesday. The worker had apparently moved too close to the robot while

Photo: Bloomberg/Getty Images

On 7 May, a Tesla Model S was involved in a fatal accident in Florida. At the time of the accident, the vehicle was driving itself, using its Autopilot system. The system didn't stop for a tractor-trailer attempting to turn across a divided highway, and the Tesla collided with the trailer. In a statement, Tesla Motors said this is the "first known fatality in just over 130 million miles [210 million km] where Autopilot was activated" and suggested that this ratio makes the Autopilot safer than an average vehicle. Early this year, Tesla CEO Elon Musk told reporters that the Autopilot system in the Model S was "probably better than a person right now."

Example : the autonomous car (CPS)

- Researchers and industrialists develop autonomous cars able to be safer than humans
  - Google car, BMW, Audi, PSA…
  - It is estimated that in the USA 94% of the car crashes are due to driver errors *(Jenkins, 2016)*

- Indeed….

https://www.youtube.com/watch?v=LfmAG4dk-rU

Example : the autonomous car (CPS)

National Transportation Safety Board

Report on Tesla first accident

Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida May 7, 2016
Accident Report, NTSB/HAR-17/02, PB2017-102600

Level 2 of autonomous vehicle:
the "driver is disengaged from physically operating the vehicle by having his or her hands off the steering wheel AND foot off pedal at the same time, " according to the SAE. The driver must still always be ready to take control of the vehicle, however.

Findings

...

3. The Tesla's automated vehicle control system was not designed to, and did not, identify the truck crossing the car's path or recognize the impending crash...

...

5. If automated vehicle control systems do not automatically restrict their own operation to those conditions for which they were designed and are appropriate, the risk of driver misuse remains.

...

Recommendation ***Incorporate system safeguards that limit the use of automated vehicle control systems to those conditions for which they were designed. (H-17-41)***

Example : the autonomous car (CPS)

Laws!
« No law, no pb »

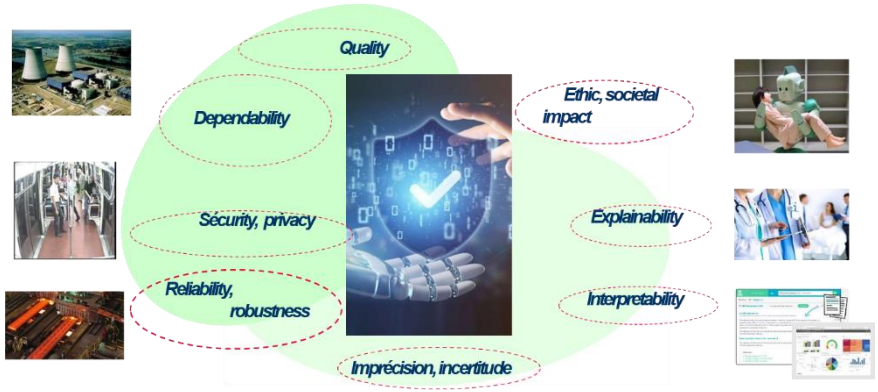
Ugo Pagallo
From Automation to Autonomous Systems: A Legal Phenomenology with Problems of Accountability

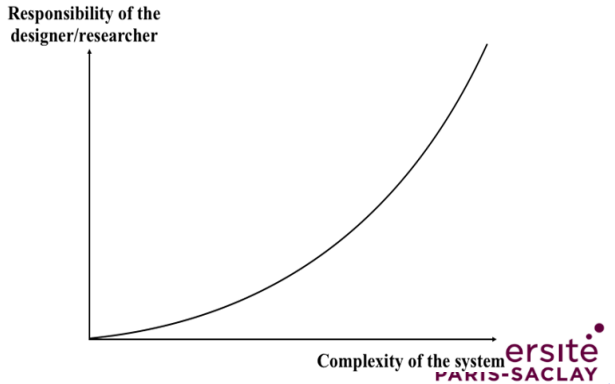
Ostrich temptation

- AI would become legal responsible entities
  (The provider, developer would not be responsible of the consequences of their possible failures)

- AI are just assistant, human will « remain » in the loop
  - E.g « we target only Level 4 autonomous vehicles »

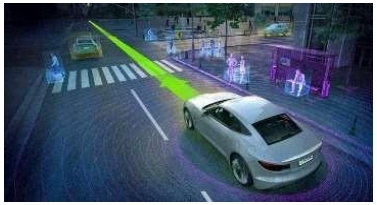
For the moment …trusted AI: a set of issues

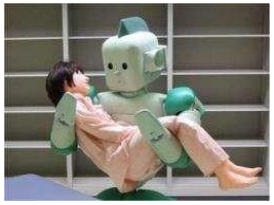




… So, should we worry about TRUSTWORTHINESS and QUALIFICATION of AI technologies?

CONFIANCE*: le besoin est là…





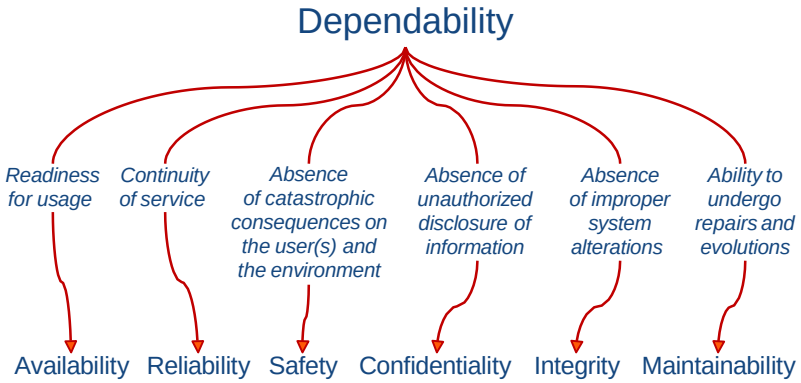
https://www.youtube.com/watch?v=OY8A-cCwL18&feature=emb_logo

*Confiance = Thrustworthiness

Not Afraid! Let's try one definition!

Trustworthiness: the ability to behave so that the others trust the information coming from the CPS and are confident about the ability of the CPS to engage actions to reach a clear, readable, public objective

→ Similar notion to Dependability



Concept	Dependability	Trustworthiness
Goal	1)ability to deliver service that can justifiably be trusted 2)ability of a system to avoid service failures that are unacceptably frequent or severe	assurance that a system will perform as expected
Threats present	1)development faults (e.g., software flaws, hardware errata, malicious logic) 2)physical faults (e.g., production defects, physical deterioration) 3)interaction faults (e.g., physical interference, input mistakes, attacks, including viruses, worms, intrusions)	1)hostile attacks (from hackers or insiders) 2)environmental disruptions (accidental disruptions, either man- made or natural) 3)human and operator errors (e.g., software flaws, mistakes by human operators)

Not Afraid! Let's try one definition!

From Trustworthy computing ..... to Trustworthy AI

Computing = Hardawre + Software+ people

AI=data+ML model+task

Trustworthy =

+Reliability

+Safety

+Security

+Privacy

+Availability

+Usability

+ **Accuracy:** How well does the AI system do on new (unseen) data compared to data on which it was trained and tested?

+**Robustness:** How sensitive is the system's outcome to a change in the input?

+**Fairness:** Are the system outcomes unbiased?

+**Accountability:** Who or what is responsible for the system's outcome?

+**Transparency:** Is it clear to an external observer how the system's outcome was produced?

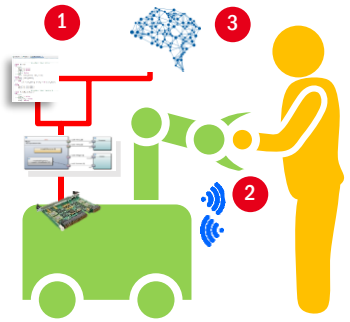
+**Interpretability/Explainability:** Can the system's outcome be justified with an explanation that a human can understand and/or that is meaningful to the end user?

+**Ethical:** Was the data collected in an ethical manner? Will the system's outcome be used in an ethical manner?

+... others, yet to be identified

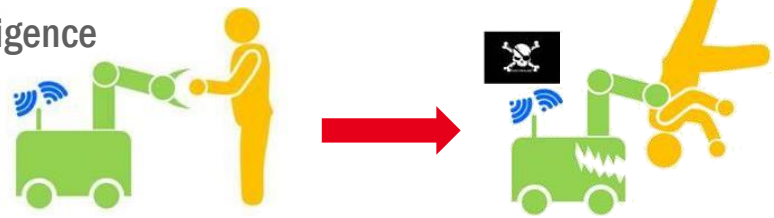
...The safety challenge

Safety of open and complex systems engineering is a true challenge



❶ Critical software functions
❷ Communications in an open world
❸ Embedded Artificial Intelligence

e.g. collaborative robots



How to be convinced that none of its behaviors could be dangerous?

From TRUSTWORTHY SYSTEM ENGINEERING *... to TRUSTWORTHY AI ENGINEERING?*

- New definition of intrinsic safety (and security) properties
- Integrate new AI design techniques: *Explainable AI, Compositional AI, Bayesian/Probabilistic deep learning...*
- *Develop analysis for stability and robustness*
- Questions on what are the other properties required to be validated?

...The verification & validation challenges

Machine learning has become alchemy

Ali Rahimi (Google)

Engineering artifacts have preceded the theoretical understanding

Yann LeCun (facebook)

Formal, traced & rational approach

- Requirements
  - *Functions Sub-functions Action*

Each instruction, each values is deduced and justified

VS

An empirical approach

- Informal requirements « by examples »
  - *trails and errors*

Poor justification, explanation of the result

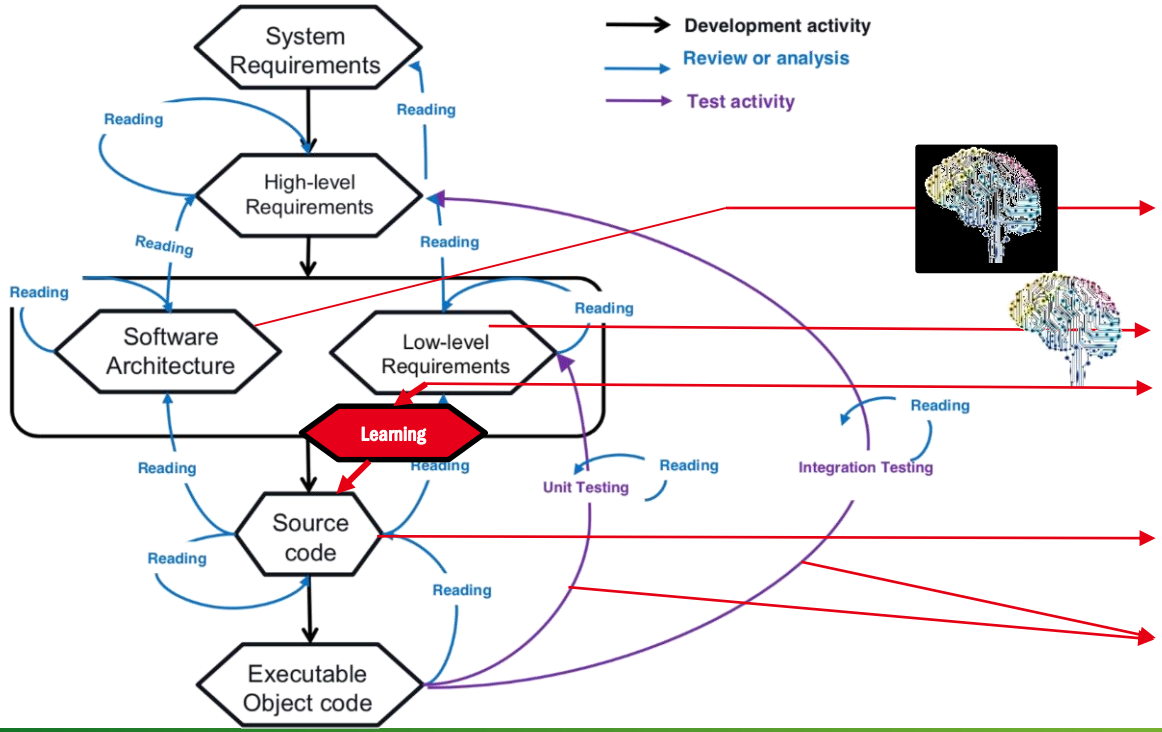
Its true for both knowledge based AI and data based AI
With a very pregnant pressure on ML based AI

AI/ML qualification is still an open issue*

Informal requirement with less / no structuration, dynamic evolution of system definition
Formal methods remains applicable but …
Hard-to-scale-up operating conditions, Verification completion criteria: when are we done with testing?

Breaks all the conformity assessment principles and processes...?

… to a 3rd AI Winter?



If AI = a single component: OK

How to formalize the requirements?

How to qualify data?

Code algorithm is « simple »: OK
Data types are « simple »: OK

Test representativity?

...The qualification challenges

- The **frequency of changes** is potentially large.
  - AI-based systems are more influenced by obsolescence of data, sensors, system's operating environment...

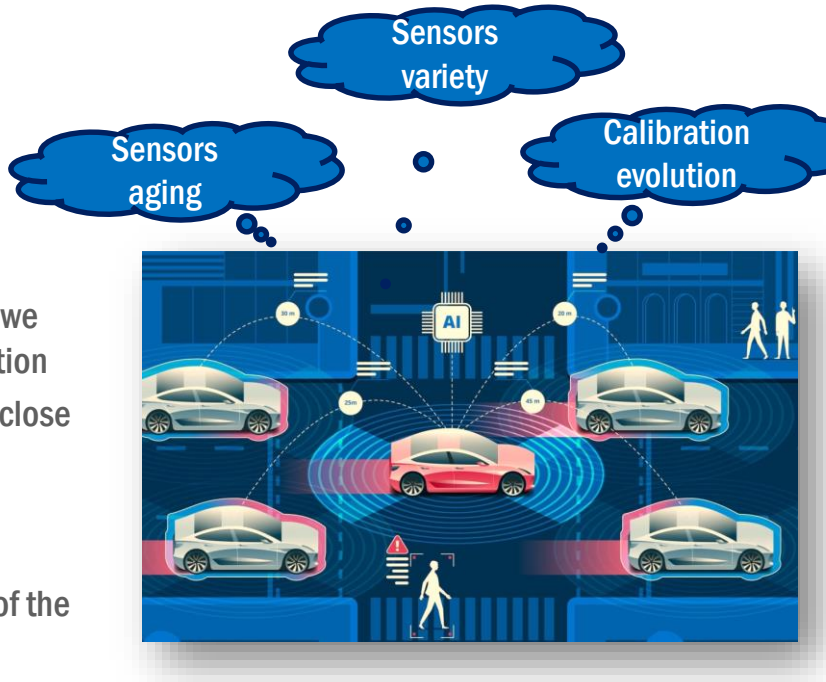
  ➔ ...which leads to need of continuous qualification processes.

- The **complexity of the validation process**
  - ...the costs of revalidation, even for small changes are very high, e.g. we could need re-training the system for slightest modification of a function

    (E.g. deep learning algorithms containing millions of parameters in close interaction)

➔ Evolutionary qualification needs **highly modular AI architectures**

- to ensure that modules and their modifications remain independent of the qualification of the entire system as much as possible.
- to become affordable in terms of re-qualification costs



Re-qualification is easier if the system has been designed with this objective...

But industry poorly equipped to define trusted AI systems

"Current assurance approaches are predicated on the assumption that once the system is deployed, it does not learn and evolve."

Three core Challenges for qualification of AI-based systems

DNN calibration evolution

Sensors aging/variety

Complex/changing operational contexts

New risks introduced by AI
(operational environment, algorithms and data uncertainty, human errors, autonomy level)

How can we assure that learning systems are safe and correct?

Unknown/unsafe unknowns
(out-of-distribution scenarios, data noise, ambiguous scenarios)

How can we manage AI/ML and environment uncertainty?

Complex & costly assurance/qualif.
(prescriptive qualification/certification approaches becomes inadequate for AI systems, obsolescence)

How can we assure that learning systems are safe and correct?

Three core Challenges for qualification of AI-based systems

DNN calibration evolution

Sensors aging/variety

Complex/changing operational contexts

New risks introduced by AI
(operational environment, algorithms and data uncertainty, human errors, autonomy level)

Unknown/unsafe unknowns
(out-of-distribution scenarios, data noise, ambiguous scenarios)

Complex & costly assurance/qualif.
(prescriptive qualification/certification approaches becomes inadequate for AI systems, obsolescence)

Towards an Evolutionary Qualification Approach for AI-based Systems

- Uncertainty-Aware Risk Management
- Runtime Risk Assessment and Learning
- Efficient and Incremental Assurance & Qualification