Safe-by-design development method for AI-based systems

DATAIA days on Safety & AI 2019

* presented at SEKE 2019 conference, July 2019

Gabriel Pedroza, Morayo Adedjouma

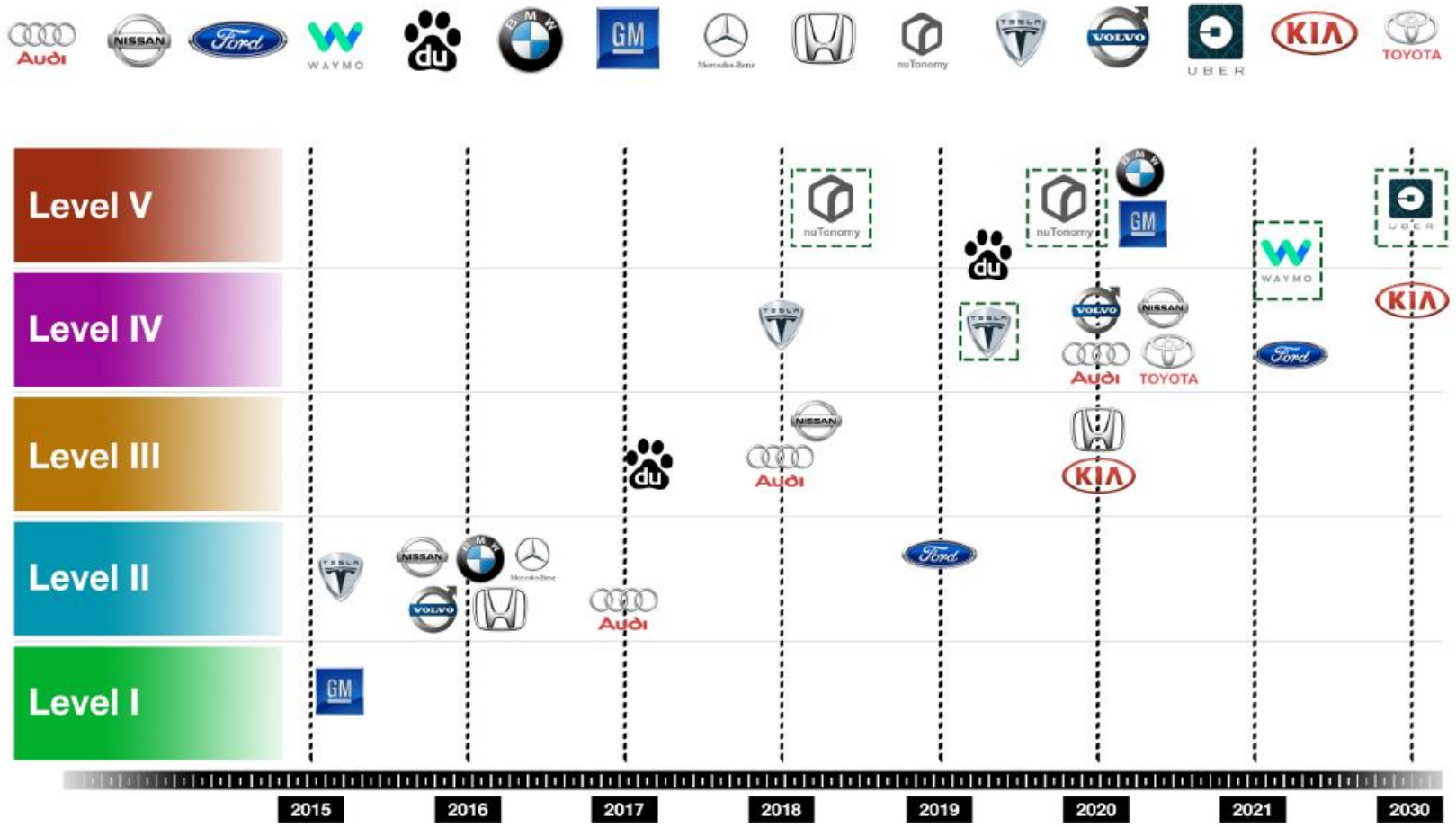
Institut CEA LIST
Département Ingénierie Logiciels et Systèmes
Laboratoire d'ingénierie d'Exigences et Conformité des Systèmes (LECS)

OVERVIEW

- **Context and problem/stakes**
- **Proposed approach**
  - AI-based reference architecture
  - Development method for AI-based systems
  - Integration of safety concerns
- **Evaluation on case study & findings**
- **Conclusions and perspectives**

OVERVIEW OF THE AV MARKET OVER TIME

- **Autonomy levels I to V as defined in SAE J3016**



Favarò FM, Nader N, Eurich SO, Tripp M, Varadaraju N (2017) « *Examining accident reports involving autonomous vehicles in California* ». PLOS ONE 12(9): e0184952.

ACCIDENTS ANALYSIS

- **Summary of accidents and comparison between AV and conventional vehicle performance**

Table 2. Google's fleet breakdown and accident frequencies.

Type of Vehicle	Total Number of Vehicles	Percentage of Fleet	Percentage of Total Reported Accidents	Total Miles Travelled	Accident Frequency	Miles per Accident
Google Prototype	37	61.7%	46%	403,226	2.4e-5	40,322
Retrofitted Lexus	23	38.3%	54%	649,841	1.8e-5	54,153

https://doi.org/10.1371/journal.pone.0184952.t002

No more than 3 people killed in accidents involving full AV(1)

Table 3. Accident frequencies by reporters/make

Type of Vehicle	Total num		nt Frequency	Miles per Accident
Nissan (Nissan and GM Cruise)			2.8e-4	3,576
Delphi/Audi			5e-5	19,787
Chevrolet (GM Cruise)	1	8,447	1.2e-4	8,447
Google Prototype	10	403,226	2.4e-5	40,322
Retrofitted Lexus	12	649,841	1.8e-5	54,153

https://doi.org/10.1371/journal.pone.0184952.t003

Table 4. **Comparison of estimated accident frequencies for AV vs. conventional vehicles.** Estimate for conventional vehicles is based on [19, 20] which provide updated data until the end of 2015. Data for 2016 and 2017 is still being process by FHWA and NHTSA.

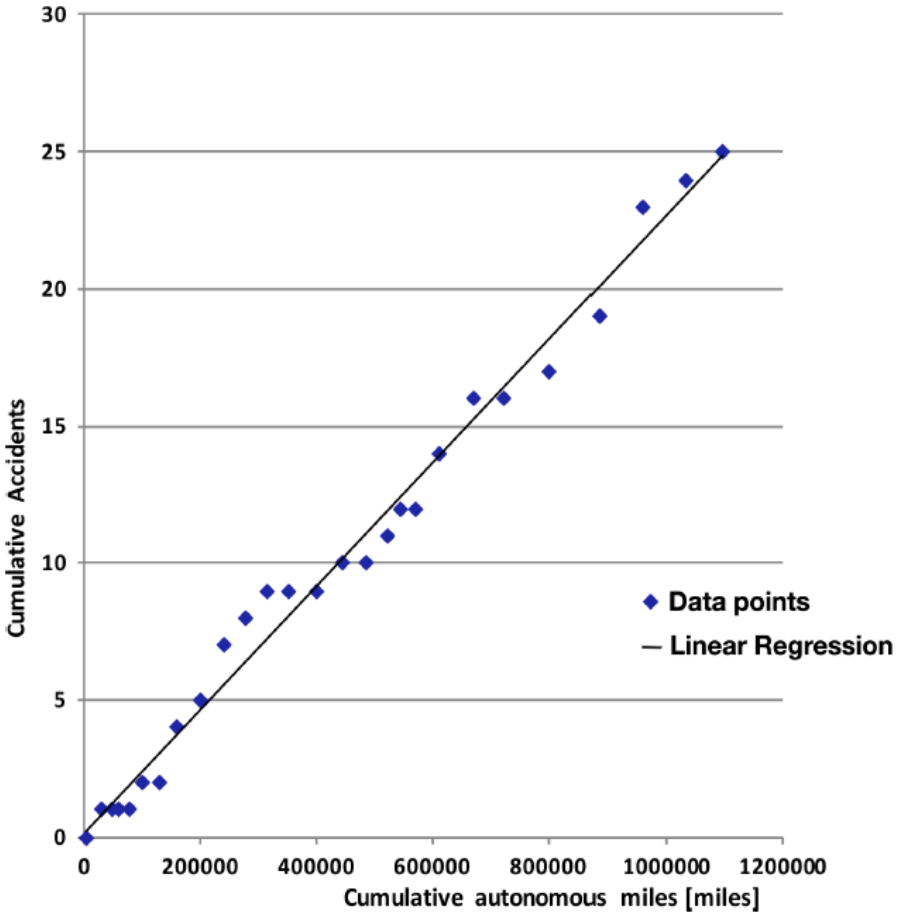
Type of Vehicle	Total number of Accidents	Total Miles Travelled	Accident Frequency	Miles per Accident
AV	26	1,088,453	2.38e-5	42,017
Conventional	6,296,000	3.148 trillions	2.0e-6	500,000

https://doi.org/10.1371/journal.pone.0184952.t004

(1) https://www.quora.com/How-many-people-have-died-in-self-driving-cars

MAIN STAKES OF AI-BASED TECHNOLOGY

- **Cumulative accidents vs. cumulative miles →**
  - Need to measure AV performance vs. conventional vehicle performance
  - Need to evaluate vehicle safety:
    - ASIL levels defined in ISO 26262 do not suffice anymore:
      - Severity
      - Likelihood
      - Controllability
    - Autonomy relies upon AI and DL modules:
      - Evaluation of malfunctioning likelihoods
      - Increasing smartness of self-control w.r.t. AI/DL limits



Favarò FM, Nader N, Eurich SO, Tripp M, Varadaraju N (2017) *« Examining accident reports involving autonomous vehicles in California »*. PLOS ONE 12(9): e0184952.

MAIN STAKES OF AI-BASED TECHNOLOGY

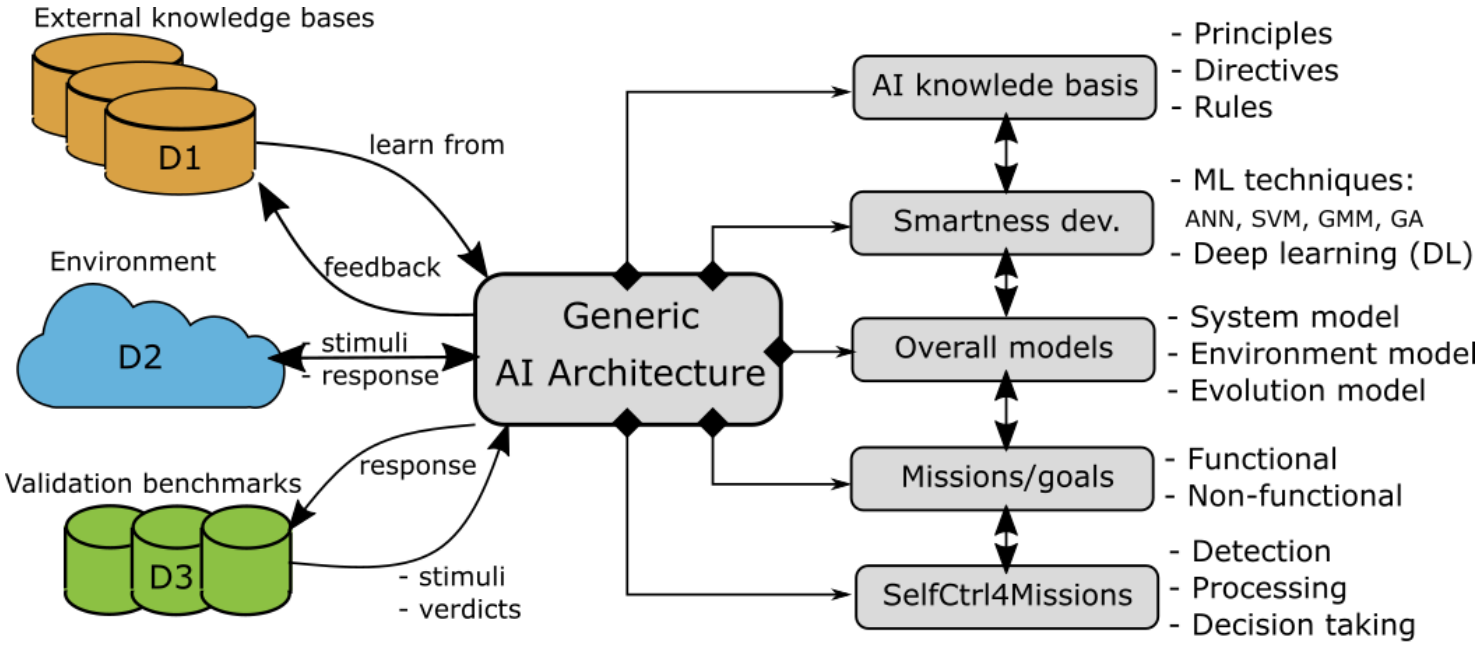
- **To increase AI-based systems safety, one must consider:**
  - Limits of AI-based systems:
    - detection capabilities (<90% in average),
    - algorithms to face unforeseen situational scenarios
  - Ensure negligible likelihoods:
    - critical hazards
    - malfunctioning
  - Conventional development methods at stake:
    - Phases, sequencing are almost static
    - Status development methods for AI-based systems: experimental phase
    - Engineering phases and their order may vary:
      - dependency engineering process ←→AI technology:
        - **Knowledge bases maturity**
        - **Knowledge bases representativeness: data sets, events, phenomena**

MAIN STAKES OF AI-BASED TECHNOLOGY

- **To increase AI-based systems safety, one must consider**
  - New standards for certification of AI-based systems:
    - "**ISO/IEC WD 23053**: Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)" →
      - In progress
    - "**ISO/PAS 21448:2019**: Road vehicles -- Safety of the intended functionality" →
      - Limited to certain levels of autonomy: I and II
      - Oriented to one application domain: automotive

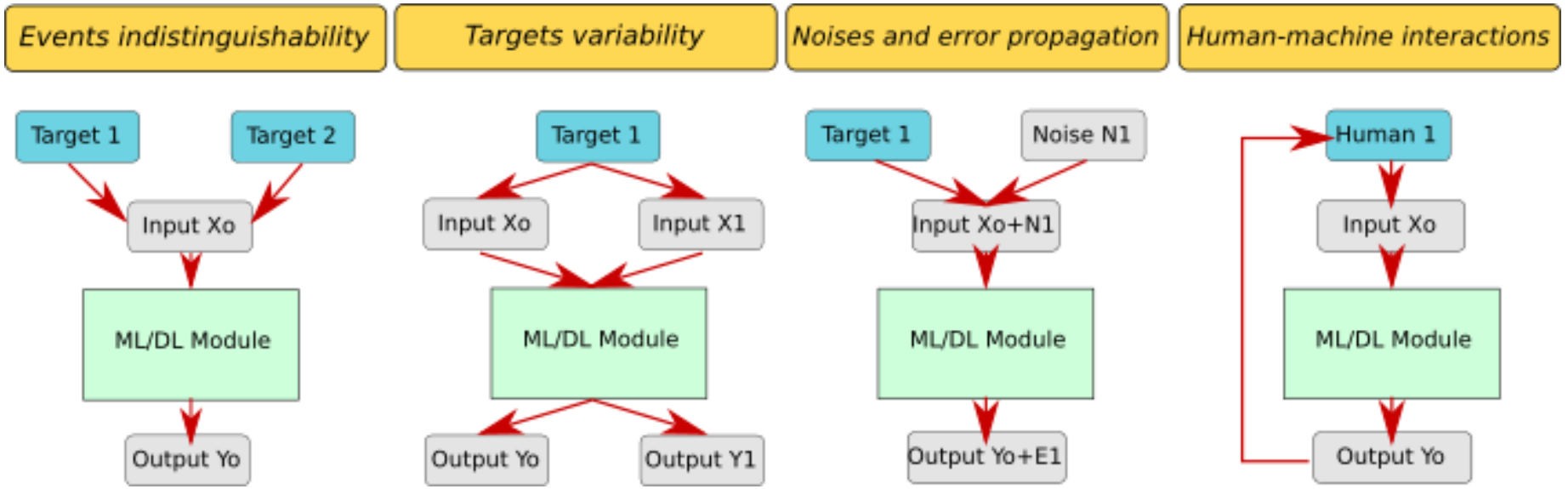
REFERENCE AI-BASED ARCHITECTURE

- **Key points:**
  - Engineering process dependent on AI technology
  - Engineering process dependent on knowledge bases
  - Knowledge bases maturity - completeness, representativeness, etc.



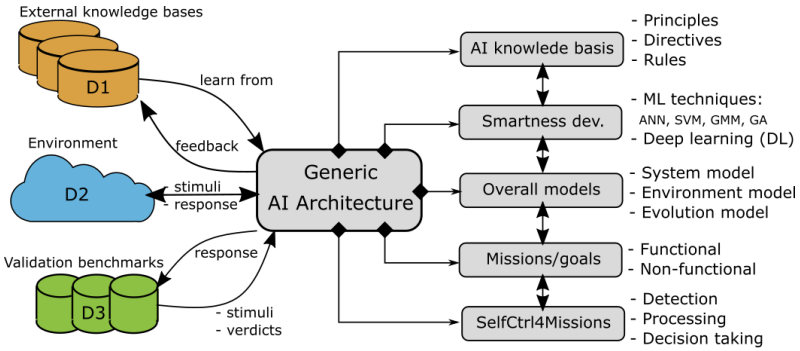
SAFETY CONCERNS (TO BE ADDRESSED)

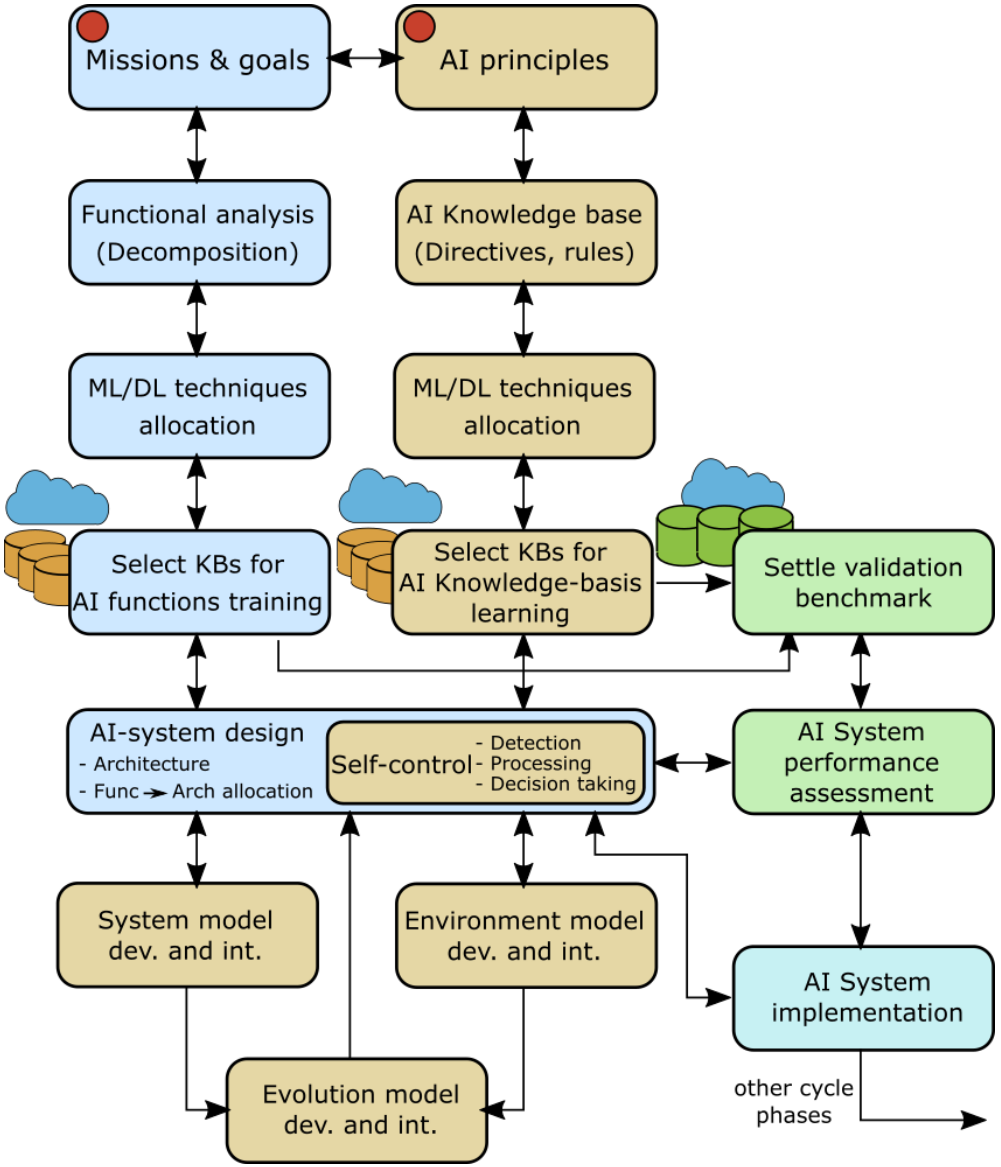
- **Mostly related to AI technology limits**
  - Indistinguishability of events
  - Variability of targets to be detected
  - Background noises and error propagations
  - Human machine interactions: driver take over machine



METHOD FOR AI SYSTEMS DESIGN

- **Main features:**
  - Traditional cycle (blue)
  - AI-layers development (brown)
  - AI-modules validation (green)
  - Help to develop and detail the generic architecture

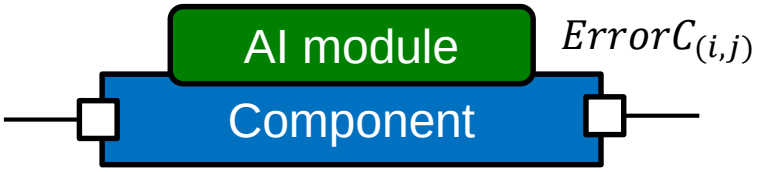




METHOD FOR SAFE-BY-DESIGN AI SYSTEMS

- **Integration of safety:**
  - Situation analysis

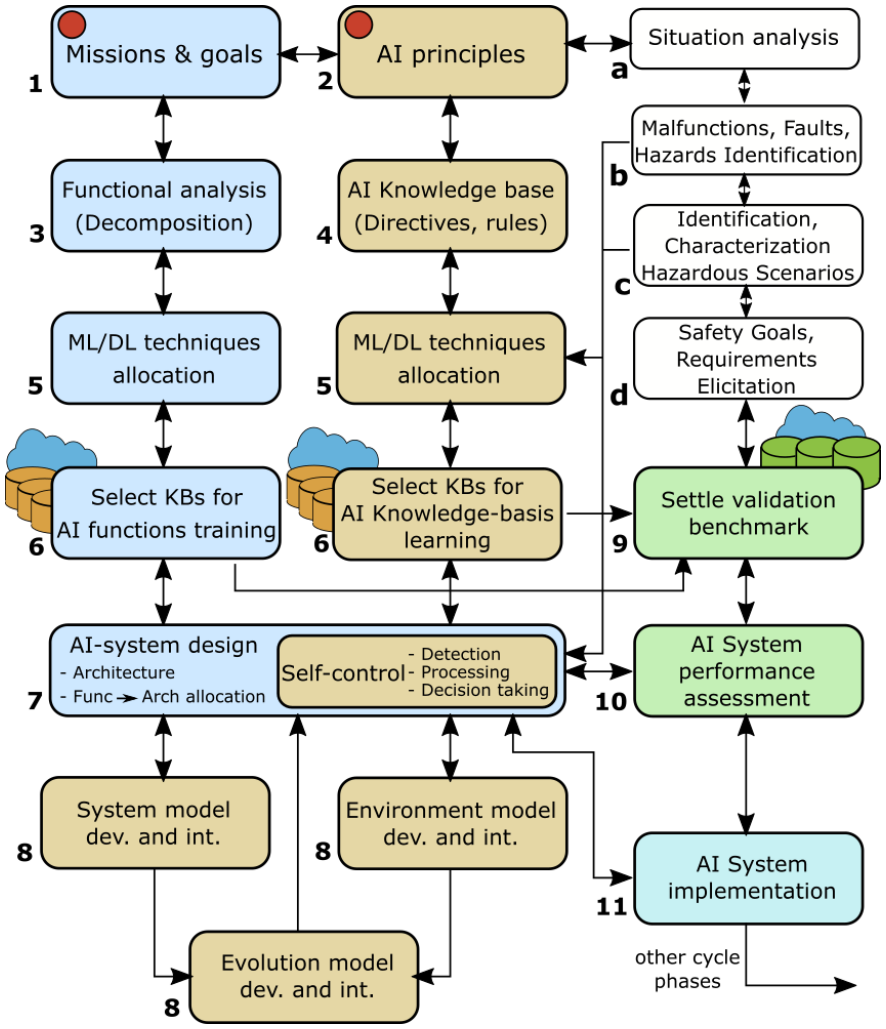
  - Malfunctions, faults, hazards identification



Input	System/Module response	
	Recognized	Not recognized
Legitimate Input	OK	*False negative*
Improper Input	*False positive*	OK

$$P[ErrorC_{(i,j)}] := P[FalsePos_{(i,j)}]+P[FalseNeg_{(i,j)}],$$
$$FalsePos_{(i,j)} := \cup_j\{C_i(Accept, B_j)\},$$
$$FalseNeg_{(i,j)} := \cup_j\{C_i(Reject, T_j)\}.$$

  - Identification, characterization of hazardous scenarios S_k



METHOD FOR SAFE-BY-DESIGN AI SYSTEMS

- **Safety goals elicitation:**
  - Scenario S_k associated to a ***monitoring formula***, e.g. ϕ: safety distance between vehicles:

$$P[\phi < \theta] \leq \delta.$$

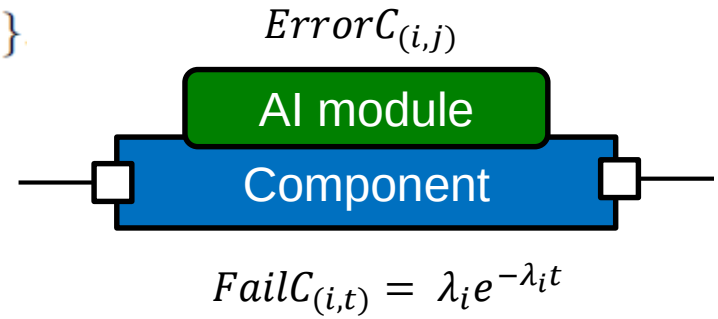
  - The scenario S_k can be validated relying upon a validation test bench. The error is given by:

$$\{P[DisfC_{(i,j,t)}]\}$$

  - The scenario S_k can be simulated. However, for certain scenarios, this can be complex and costly. The error is also given by:

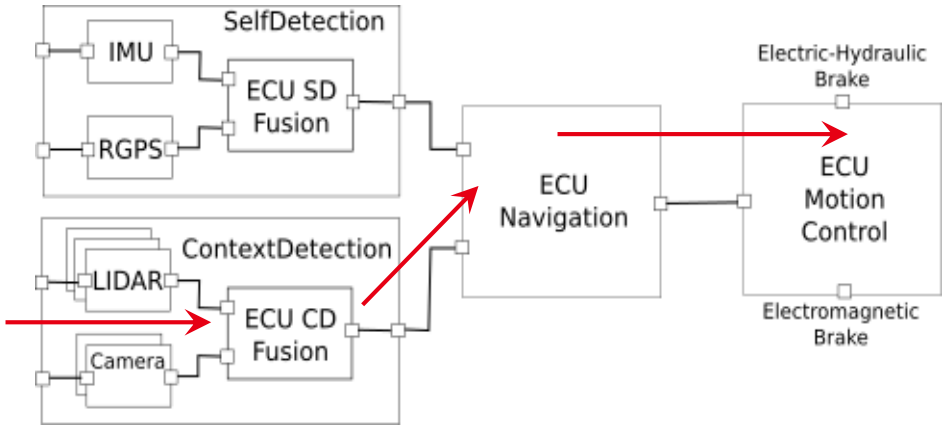
$$\{P[DisfC_{(i,j,t)}]\}$$

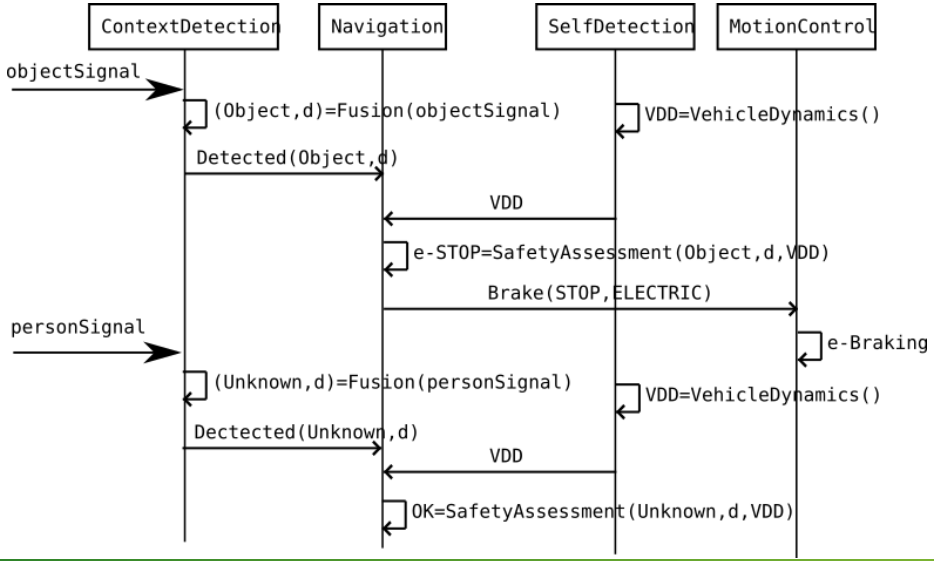
$$P[DisfC_{(i,j,t)}] = \omega_1 P[FailC_{(i,t)}] + \omega_2 P[ErrorC_{(i,j)}],$$
$$\omega_1 + \omega_2 = 1.$$



EVALUATION ON CASE STUDY

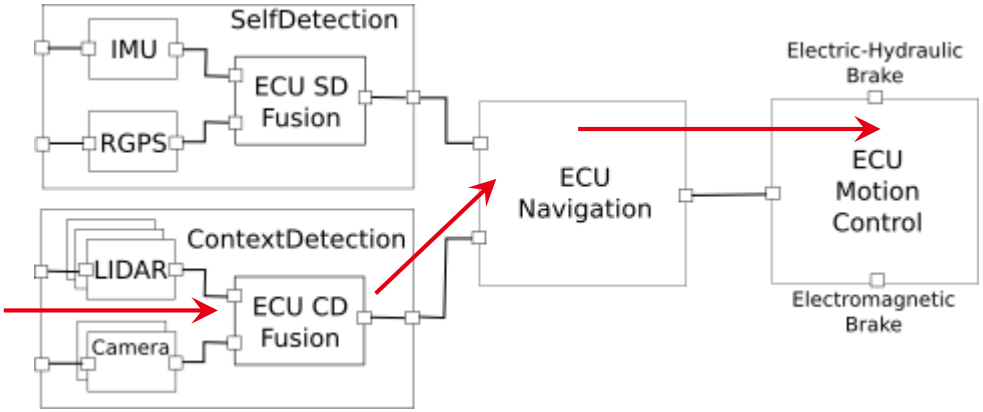
AV for public transportation





EVALUATION ON CASE STUDY

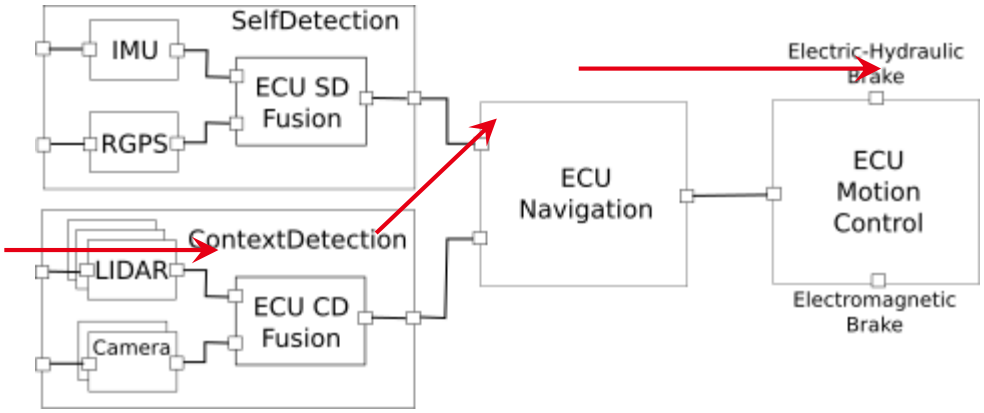
- **Probability of hazardous scenario**



$$P[S_k] := \lambda_{LIDAR}\lambda_{Camera} + P[DisfECU_{CD}] + \lambda_{IMU}\lambda_{RGPS} + P[DisfECU_{SD}] + P[DisfECU_{Navigation}] + \lambda_{MotionControl}.$$

FINDINGS/LIMITATIONS

- **Sources of uncertainty (case study):**
  - Accuracy and maturity of KBs: impact the learning process and performance of ML/DL components
  - Difficulty to apprehend usage-scenarios: infinite possible environmental-operational contexts
  - performance limits of AI-based components
  - Interpretation and decision-taking layers are at stake:
    - contradictory directives in critical scenarios
  - deploy new capabilities in real time



CONCLUSIONS AND PERSPECTIVES

- **Conclusions**
  - Enhancement of typical hazard analysis method to infer safety goals
  - Malfunctioning likelihood of AI-systems = typical failure rate + error probability of ML/DL modules
  - Sources of uncertainty

- **Perspectives**
  - Larger-scale application of the method
  - Applicability of standard-preconceived methods: FMEA and FTA
  - Cover stages of the development cycle, i.e., testing and validation

Gabriel PEDROZA & Morayo Adedjouma, PhD
Research Engineer

Institut CEA LIST
Département Ingénierie Logiciels et Systèmes
Laboratoire d'ingénierie d'Exigences et Conformité des Systèmes (LECS)
Gabriel.PEDROZA@cea.fr, Morayo. Adedjouma@cea.fr
www-list.cea.fr

Commissariat à l'énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

Établissement public à caractère industriel et commercial | RCS Paris B 775 685 019