86th IFIP 10.4 Meeting – Gold Coast, Australia

Session 1 summary

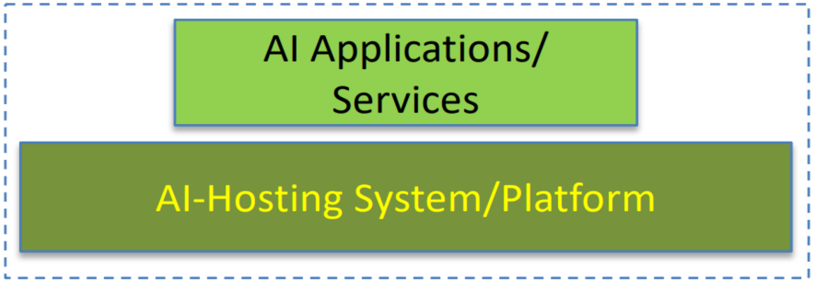
Presented by Ilir Gashi

Overview

- Title of session: **Security, Safety and Fault Tolerance of AI systems**
- Two talks:
- On Fault Tolerance of AI Systems
  - **Long Wang**, Tsinghua University, China
- Safe and Secure AI/ML-driven Autonomous Vehicles? Not anywhere near yet …
  - **Paulo Esteves-Veríssimo,** RC3 (Resilient Computing and Cybersecurity Centre), CEMSE, KAUST, Saudi Arabia

On Fault Tolerance of AI Systems

- The outline of the talk from Long
  - Fault Tolerance (FT) in Classical Computing
  - FT of AI Systems
    - FT of AI Applications
    - FT of AI-Hosting Systems
  - Case Study: FT of AIGC Applications

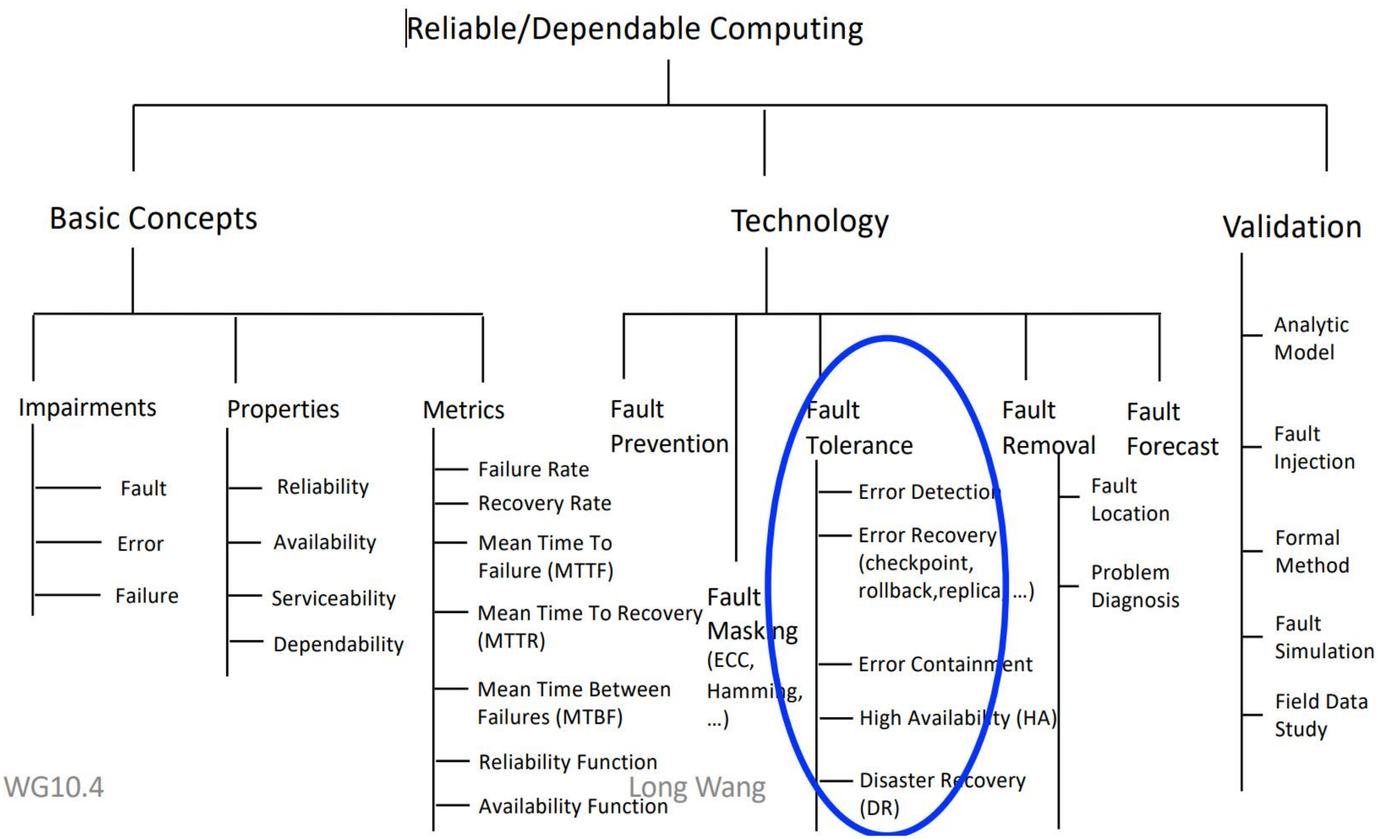


AI Systems

- The main aim of the talk was to compare the fault tolerance strategies, fault and failure models for classical computing systems with those of AI applications and AI hosting systems
- The focus of Long's talk was on fault tolerance against non-maliciously induced faults and failures

Tsinghua University

Overview of Classical Reliable Computing

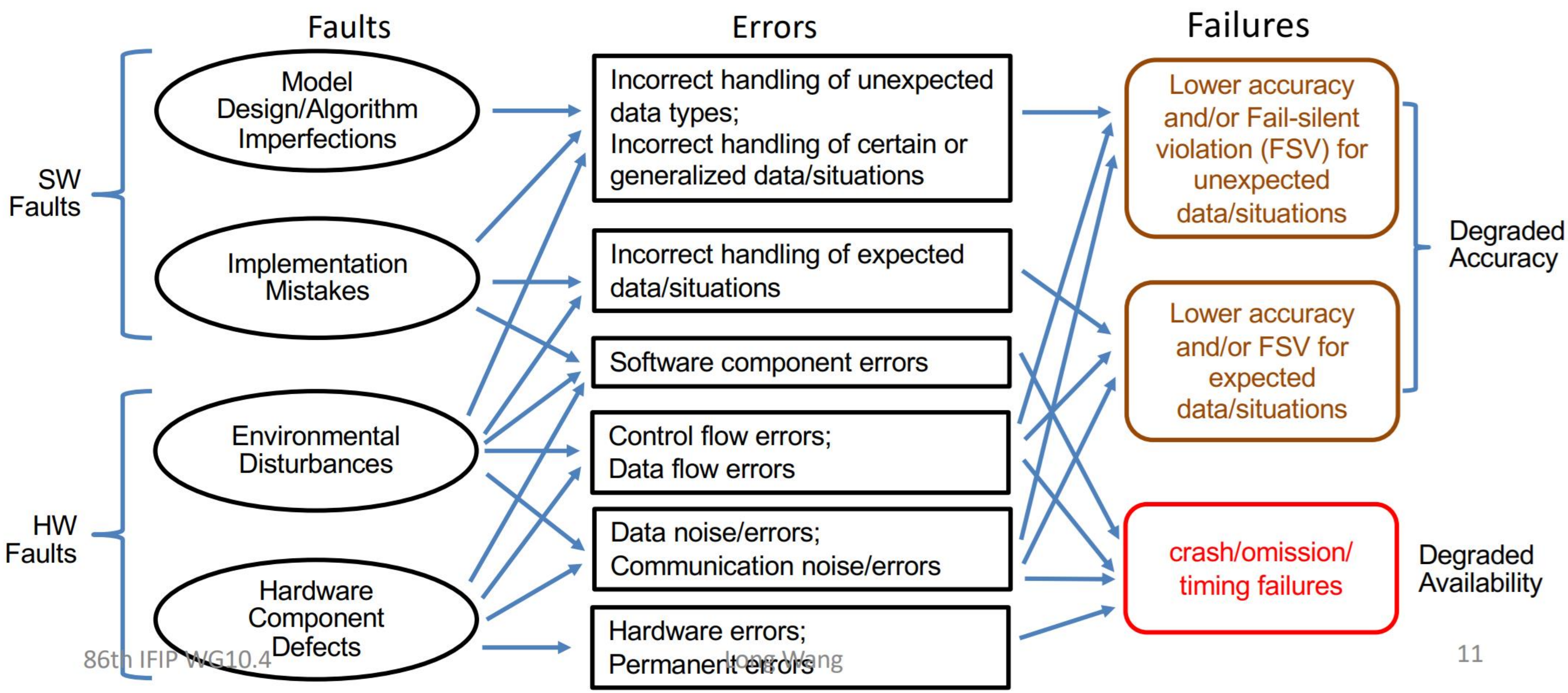


86th IFIP WG10.4 Long Wang 5

Fault Tolerance in Classical Computing (cont.)

- Error Detection
  - Watchdog timers, Heartbeats
  - Consistency and capability checking
  - Exception handling
  - Control-flow checking
  - Data audits, data flow checking
- Error recovery
  - Restart
  - Checkpoint and rollback
  - Rollforward
  - Replicas/replication with failover support
- Fault tolerance
  - Hardware Redundancy
    - Triple Module Redundancy, m-out-of-n structure, active-active, active-passive
    - Voting
  - Software Fault Tolerance
    - Robust data structures
    - Recovery blocks
    - N-version programming
    - Process pair
    - Voting or Acceptance Test
  - Combining specific error detection and error recovery techniques

FT of AI Applications – Fault Model and Error Manifestation



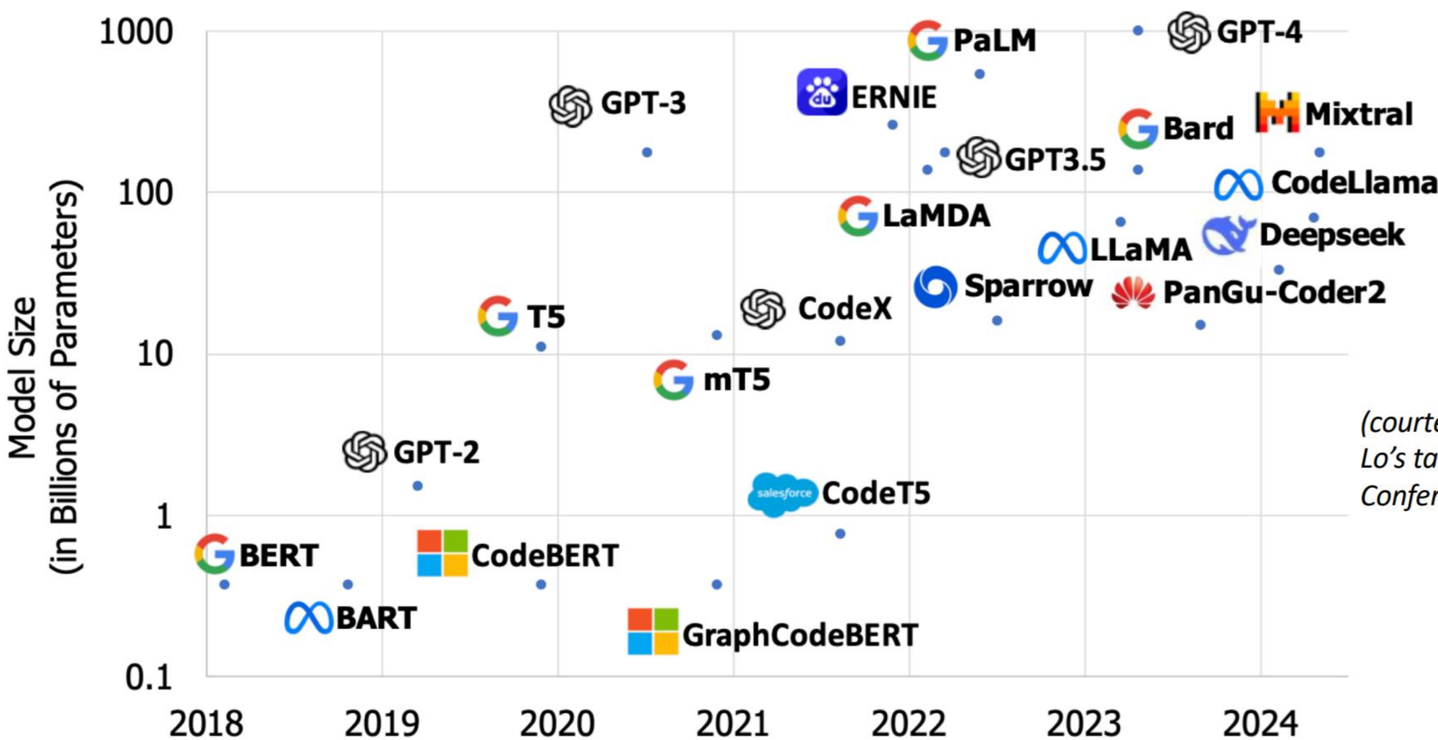
Fault Tolerance of AI Applications

Failure Category	Degraded Accuracy for Inference Tasks	Degraded Accuracy for Training Tasks	Degraded Availability
Error/Failure	• Incorrect task results (e.g. misclassification)	• Incorrect model states (e.g. bad model weights) • Longer convergence	• Crash, omission, timing failures • Control flow errors • Data flow errors

Error Detection	• The failure (incorrect result) itself
	  ◦ E.g. user feedback
	• Acceptance check
	  ◦ Rule based
	  ◦ AI model based
	• Accuracy metric monitoring
		• Incorrect inference results on test data sets during training
		  ◦ Lower accuracy
		• Range checking of model weights or intermediate values
		• Task-specific metrics for detection
			• Classical error detection
			  ◦ Crash/hang detection, heartbeat, consistency checking, multi-replica vote, control/data flow check, data audits

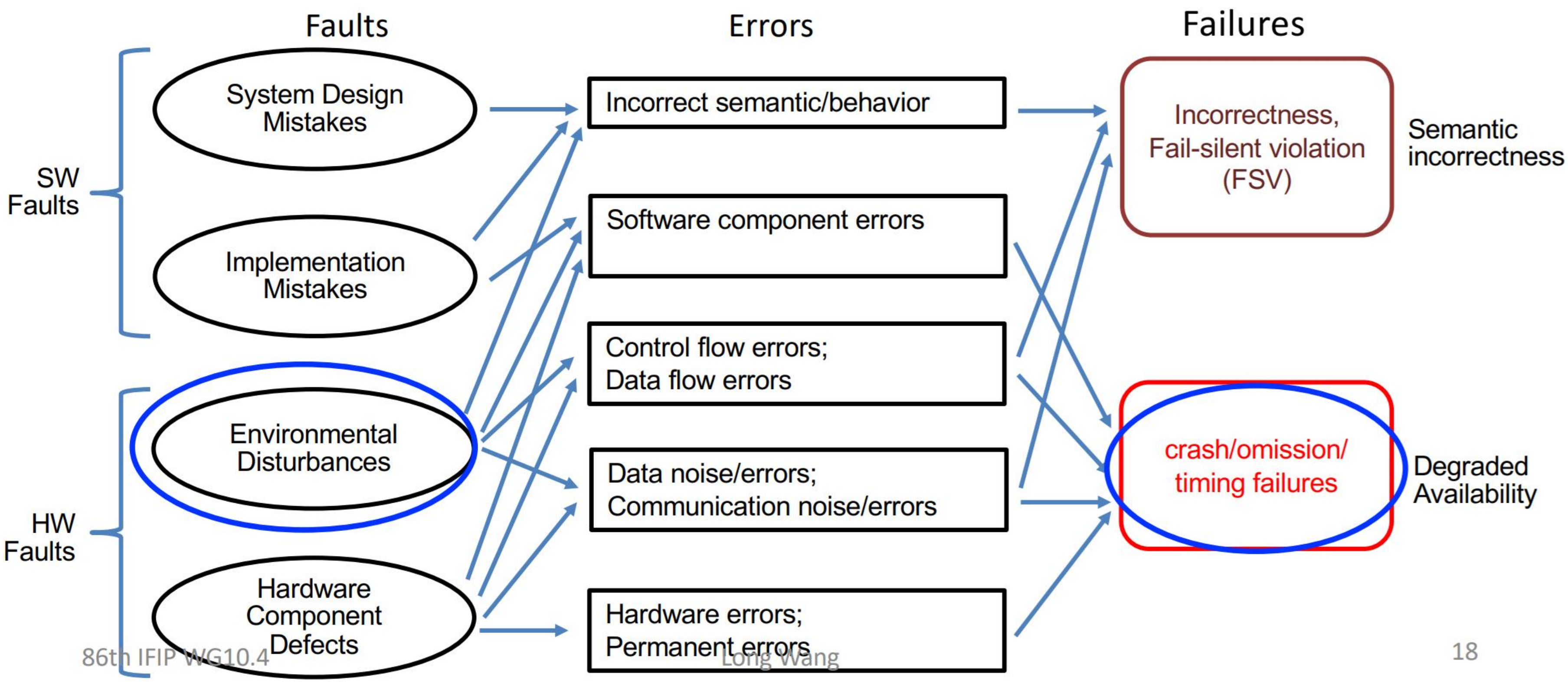
Error Recovery/ Tolerance	• Re-execute, restart
	• Rollforward
	• Multi-replica vote (e.g. TMR)
	  ◦ May be in partitioned module level
	  ◦ Multi-version or diversified models
	• Error analysis and root-causing
	• Improving with fine-tuning
	• Re-training
		• Checkpoint/backup and rollback
		  ◦ Saving model state periodically for recovery without losing progress
		• Multi-replica vote
		  ◦ TMR-like
		  ◦ Ensemble Learning
		  ◦ May be in partitioned module level
		• Error analysis and root-causing
		• Improving with fine-tuning
			• For training tasks
			  ◦ Checkpoint and rollback, multi-replica vote
			• For inference tasks
			  ◦ Re-execute, restart, rollforward, multi-replica vote, replicas/replication with failover

LLM Models Grow to Huge Sizes



(courtesy of Prof. David Lo's talk in Huawei STW Conference 2024)

FT of AI-Hosting Systems – Fault Model and Error Manifestation

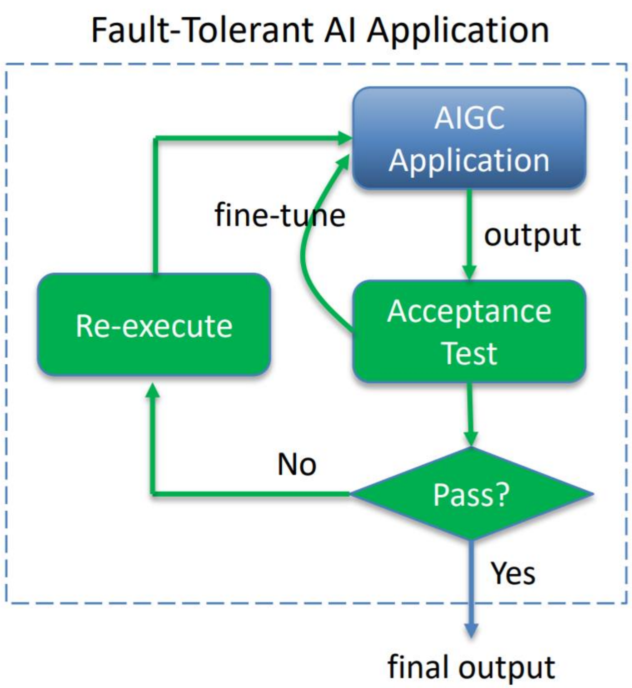


Fault Tolerance of AI-Hosting Systems

- Failure Category
  - Degraded availability
  - Semantic incorrectness
- Error Detection
  - For degraded availability
    - Classical error detection (process crash, system exception, log information, error-detecting code like CRC checksum)
  - For semantic incorrectness that result in degraded accuracy, and for other semantic incorrectness
    - Error detection of AI application outputs against degraded accuracy
    - May need specific error detections (error-detecting code like CRC checksum, rule check, control/data flow check, customized check)
- Error Recovery and Tolerance
  - Similar to error recovery and tolerance of cloud systems or data centers
    - E.g. fail-forward for inference jobs, checkpoint/rollback for training jobs
  - As AI-Hosting systems are just such infrastructure as cloud systems, data centers, or simpler-structure computer systems

Case Study: FT of AIGC Application using Acceptance Test

- Combining error detection and error recovery for providing FT of AI applications
  - AIGC application: AI-generated content
  - Error detection: acceptance test
  - Error recovery: re-execute
- Acceptance Test
  - Rule based
    - Depending on scenarios, there may be rules that can be implemented to check if the output of is correct
  - AI model based
    - AI models as discriminators to check if the output is acceptable or not
- If the acceptance test fails, re-execute the AIGC application with different initial input
- The final output has much higher accuracy than the original one
- The acceptance test can also help fine-tune the AIGC application/model



Summary

- FT technologies in classical computing mostly still applies to AI applications/ systems (with adaptations if needed)
  - Error detection, error recovery, and a combination of them
  - E.g. acceptance test largely improves the AI application accuracy
- Failure models of AI applications/systems mainly fall into two categories
  - Degraded accuracy and degraded availability
- Semantic analysis based rule checking helps detect degraded accuracy of AI applications
- We can learn a lot from experiences of FT in cloud and supercomputer systems for FT of AI applications/systems, because
  - AI applications share a lot of similarities with supercomputing applications or cloud services
  - AI-hosting systems share a lot of similarities with cloud systems and supercomputer systems

Q&A on Long’s talk

- Good Q&A on the talk from Long
- Colleagues encouraged Long to expand the analysis to look at not only non-malicious but malicious faults and failure also.
- The Fault-Error-Failure (FEF) model can still be used, and the extension done in the MAFTIA project for example can be applied (the Attack-Vulnerability-Intrusion Fault-Error-Failure model):
  - ➢ D21.pdf (ncl.ac.uk)

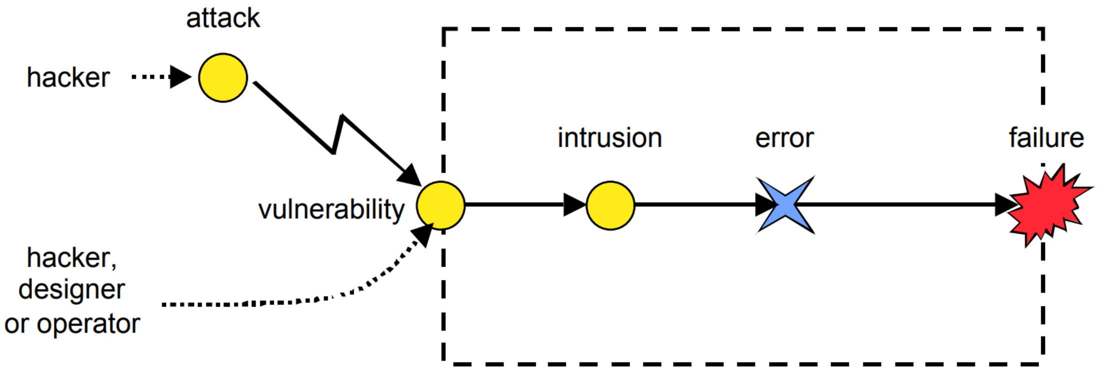


Figure 8 — Intrusion as a composite fault

Safe and Secure AI/ML-driven Autonomous Vehicles? Not anywhere near yet …

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Resilient Computing and Cybersecurity Center

Resilient Computing and Cybersecurity Center

Brief Analysis of the Cyberspace *today*

- distributed infrastructure:
  - *Pervasive CPS and IoT*; seamless integration with Internet/Cloud/Web.
- highly exposed to threats:
  - Huge *pressure to go "digital"*: Govs; BigTechs; Social nets.
- steadily increasing software vulnerabilities:
  - Common SW yearly *rate increased* 2-3-fold; *CPS/IoT* in great increase
- degradation of the threat surface:
  - *Even more* powerful adversary actors and sophisticated exploit tools

So, what's wrong about the current autonomous vehicles ecosystem?

- ***To start with, the very notion that there is an ecosystem is inexistent***

- ***An analysis of the ecosystem as a critical infrastructure is missing***

Safety-security gap in vehicle ecosystems

Faults in a well designed car ecosystem lead to an **infinitesimal and acceptable** probability of catastrophic failure;

Faults in a well designed car may imply a **non-negligible** probability of catastrophic failure

Vulnerabilities in a car ecosystem **will** lead, rather sooner than later, to catastrophic failures;

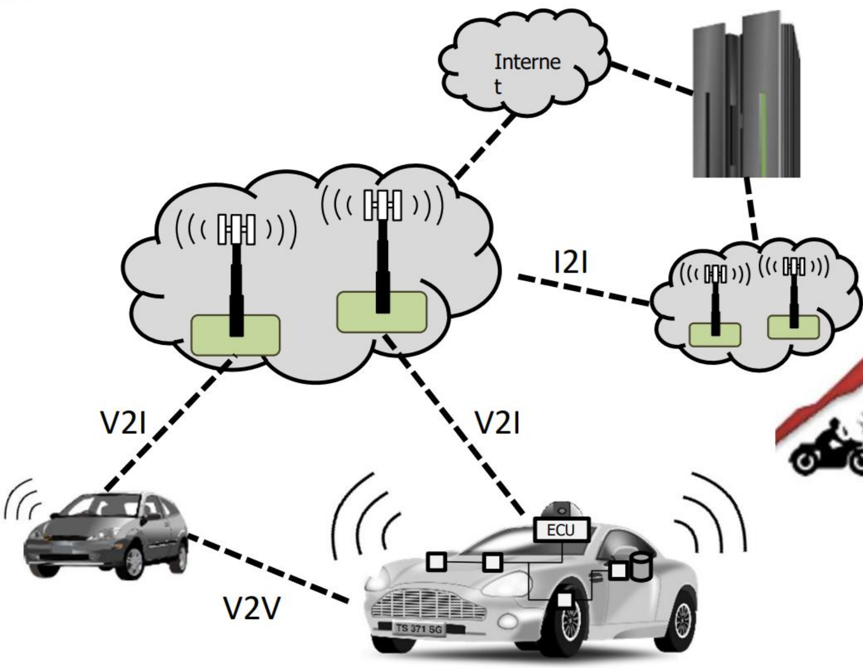
Towards Safe and Secure Autonomous and Cooperative Vehicle Ecosystems. Lima, A; Rocha, F; Volp, M; Verissimo, P. in Proc's 2nd ACM Workshop on Cyber-Physical Systems Security and Privacy (2016, October) @CCS, Vienna-Austria

Autonomous Vehicle Ecosystem

King Abdullah University of Science and Technology

جامعة الملك عبدالله للعلوم والتقنية

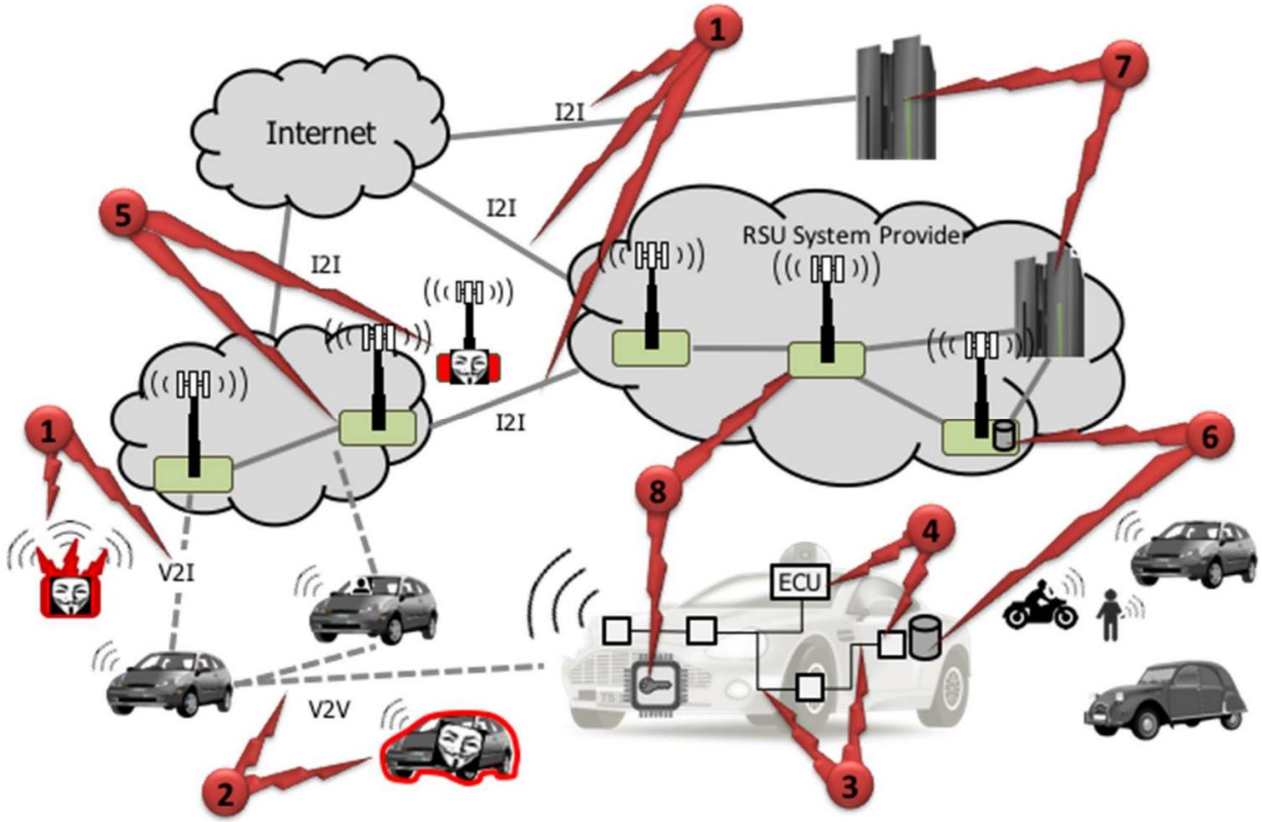
2016



Towards Safe and Secure Autonomous and Cooperative Vehicle Ecosystems. Lima, A; Rocha, F; Volp, M; Verissimo, P. in Proc's 2nd ACM Workshop on Cyber-Physical Systems Security and Privacy (2016, October) @CCS, Vienna-Austria

Autonomous vehicle ecosystem
threat surface perhaps wider than many think

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

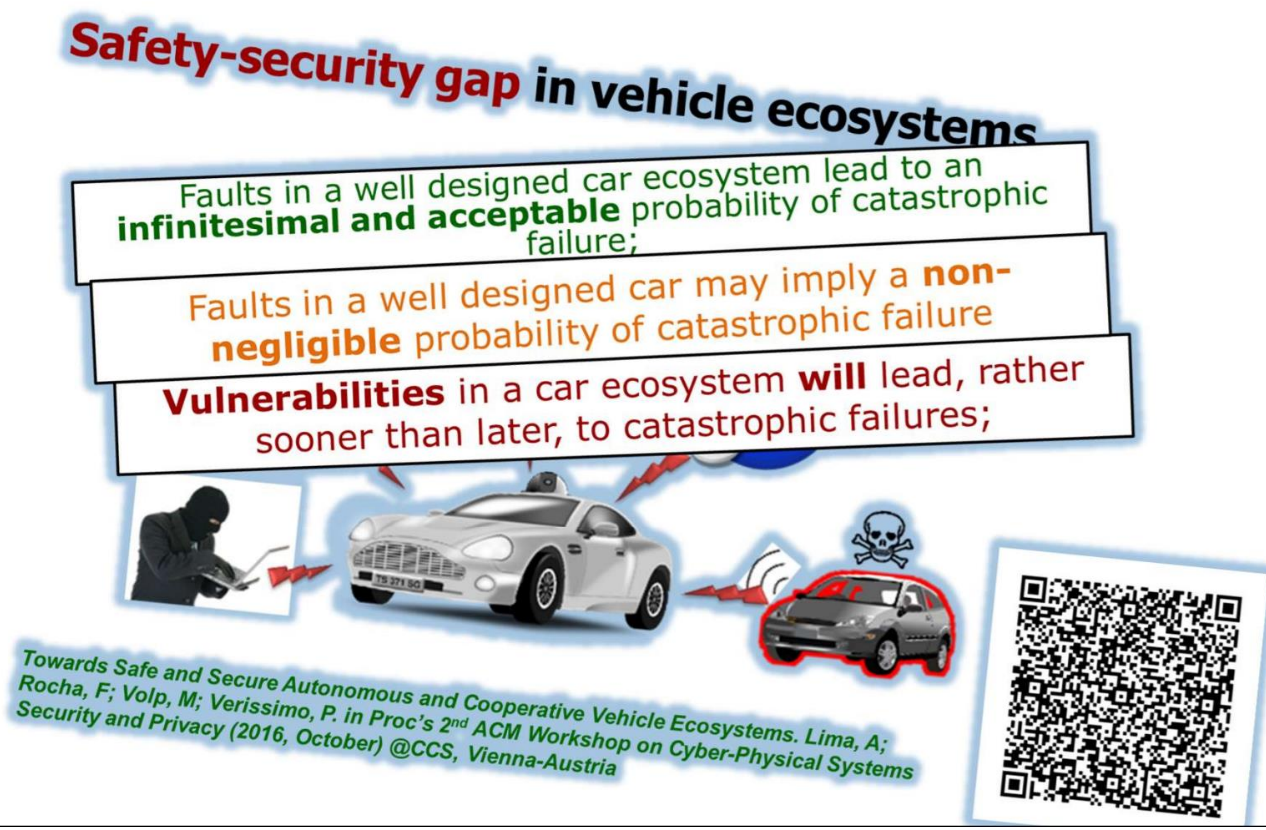


Threat Vectors

CITY
UNIVERSITY OF LONDON
EST 1894

How serious is that?

«IF IT AIN'T SECURE, IT AIN'T SAFE»



Homogeneous ML-based systems cannot give strong assurance and resilience guarantees

- **Status-quo**
  - *Autonomous cars use ML-powered multi-sensor perception (mainly vision) and control, and sometimes redundant modules to which the MLearned module hands over in case of problems.*
- ***Assurance***
  - *LOW- Infeasible to provide reliable figures/conclusions, impossible to certify*
- ***Resilience***
  - *LOW- Fair success in handling unforeseen, emergent or out-of-envelope behaviours; often even blind to those situations*

Philosophical side of the problem:

«Control the physics of event interleaving in autonomous object ecosystems, acting in real time, in open and largely unpredictable environments»

Solutions? ...

COMPONENT-BASED, INDIVIDUALIZED

ATTACK PREVENTION, ACCESS CONTROL, FWALLS, ETC.

VULNERABILITY PREVENTION AND REMOVAL

HUMAN-STEERED AD-HOC MITIGATION

A part of the long journey towards

RESILIENT AUTONOMOUS VEHICLE ECOSYSTEMS

More recently, A. Shoker and R. Yasmin at CybeResil@KAUST, M.Voelp CRITIX@UNILU, V. Rahli @U.BIRMINGHAM, J. Decouchant@U.DELFT

CORTEX Project Info

[2001-04]

INFORMATION SOCIETY TECHNOLOGIES
(IST) PROGRAMME

ist
information society technologies

Project acronym: ***CORTEX***
Project full title:
CO-operating Real-time senTient objects: architecture and EXperimental evaluation

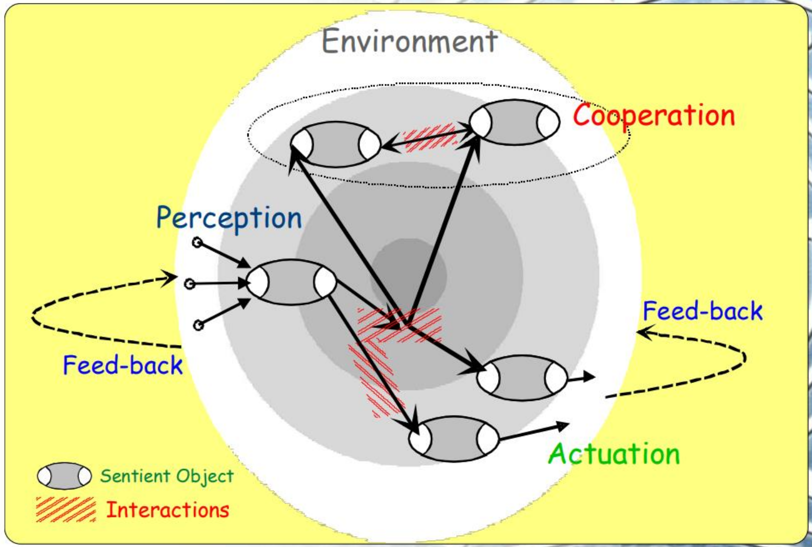
- Members:
  - Univ. Lisboa Fac. Of Sciences (PT) **(proj. coord.)**
  - Trinity College of Dublin (IR)
  - U. of Lancaster (UK)
  - U. of Ulm (DE)
- Duration:
  - 3 years, starting April 2001
- Budget:
  - 2 MEURO

'Sentient objects' interaction model

Abstract safe distributed real-time (DRT) autonomous control of free-running objects

should support the classes of R/T interactions objects need to perform:

- sentience of body and of environment;
- environment-to-object and vice-versa;
- object-to-object



[P. Veríssimo and A. Casimiro. The Timely Computing Base Model and Architecture. IEEE Tacs. on Computers, 2002]

KARYON PROJECT: Kernel-Based ARchitecture for safetY-critical cONtrol

2011-2014

KARYON

Academia & Research Institutes
SMEs and Industry

FACULDADE DE CIÊNCIAS UNIVERSIDADE DE LISBOA

gmv INNOVATING SOLUTIONS

EMBRAER

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

SP your Science Partner

CHALMERS UNIVERSITY OF TECHNOLOGY

4S Group Technology for Sustainability

Proof-of-concept prototypes
Simulations

Avionics

UAS/Aircraft flight mission

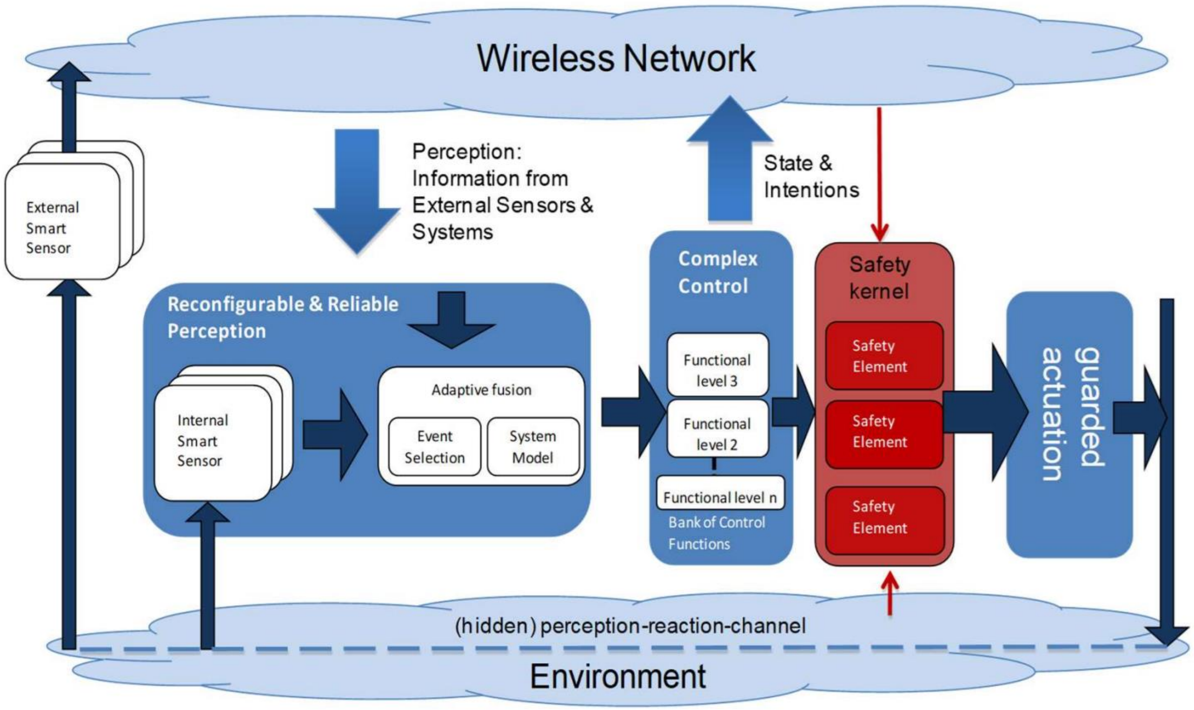
Automotive

Adaptive cruise control
Coordinated lane change
Coordinated intersection crossing

- Provide system solutions for predictable and safe coordination of smart vehicles that autonomously cooperate and interact in an open and inherently uncertain environment

KARYON 93

KARYON architectural view: proof of concept of hybridisation for safety



*A. Casimiro, J. Kaiser, E. Schiller, P. Costa, J. Parizi, R. Johansson, R. Librino, "**The KARYON Project: Predictable and Safe Coordination in Cooperative Vehicular Systems**", in 2nd Workshop on Open Resilient Human-aware CPS (WORCS'13), Jun. 2013.*

Intel Collaborative Research Institute for

Collaborative Autonomous & Resilient Systems *(CARS)*

securityandtrust.lu

CRITIX

https://www.icri-cars.org/

2017-2020

ICRI-CARS » Resilient Autonomy » Mission

ICRI-CARS

- Mission
- Research Topics
- Principal Investigators
- TU Darmstadt
- Aalto University
- Ruhr-University Bochum
- Critix@ University of Luxembourg
- TU Wien
- Collaborations

Intel Collaborative Research Institute for Collaborative Autonomous & Resilient Systems (ICRI-CARS)

About Collaborative Autonomous and Resilient Systems (CARS)

The mission of the ICRI-CARS is the study of security, privacy, and safety of autonomous systems that may collaborate with each other. Examples include drones, self-driving vehicles, or collaborative systems in industrial automation. CARS introduce a new paradigm to computing that is different from conventional systems in a very important way: they must learn, adapt, and evolve with minimal or no supervision. A fundamental question therefore, is what rules and principles should guide the evolution of CARS?

This raises security related questions in multiple research areas:

1. Trustworthy and Controllable Autonomy
2. Fair and Safe Collaboration Tolerating Failures and Attacks
3. Intelligent Security Strategies for Self-Defense and Self-Repair
4. Integration of Safety, Security, and Real-time Guarantees
5. Autonomous Systems, Ecosystem Scenarios, Requirements, Case Studies, and Validation
6. Advanced Platform Security for Long-term Autonomy

Resilience enablers
for autonomous and collaborative vehicles

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Applied safe and secure DRT autonomous control --- general driving

- **Powerful architectures** (e.g. manycores), capable of: high-power computing, enabling security/safety defenses
- **Secure and dependable *real-time* communication**, V2V and V2I, despite accidents and attacks
- **Automatic in-car resilience** mechanisms for safety and security (gateway, ECU, trusted components/enclaves)

Intrusion Resilience System (IRS)

Trustworthy Autonomous Vehicles Architecture (SAVVY)

KAUST In-house Projects

2021----

Towards sustainable security and safety

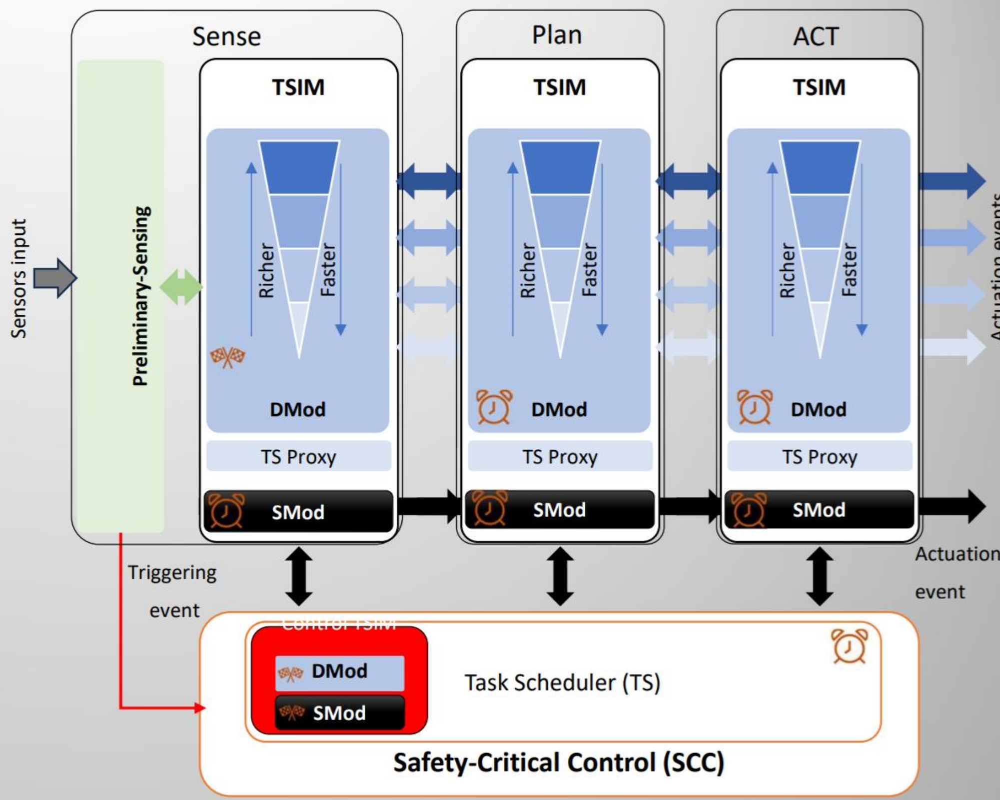
In AV control

Savvy Architecture

- Preliminary Sensing
  - Detect an Event
  - Define Time-to-Event (T2E)

- Safety-Critical Control (SCC)
  - Define Time-to-Hazard (T2H)
  - Set T2E and T2H timers
  - Schedule Tasks over Time-Sensitive Intelligent Modules (TSIM)

- Timer T2H << T2E:
  - TSIM tunes ML model to deliver before T2H
- Timer T2H = T2E
  - Fail-operational: SCC takes over



Crucial non-technical enablers:

- Resilience technologies (sustainability through threats)
- **Laws and regulations** (Europe is advanced here)

Move fast break things?---

King Abdullah University of Science and Technology | Resilient Computing and Cybersecurity Center

TAKE-AWAYS:

Resilient Computing and Cybersecurity Center

Ecosystem mindset

Laws and regulations, "no Far-West"

AV systems (AI/ML or other) cannot ignore distributed real-time systems and control theory

Accidents and attacks, safety and security

Reconciliation of uncertainty with predictability must be an inherent design predicate, not an after thought, a question of "training better"

Modular and technology neutral resilience solutions, from mechanical to cyber world

Q&A on Paulo's talk

- Good Q&A on the talk from Long
- Colleagues asked about whether some of the issues can be seen as perception failures rather than safety failures (though with the acknowledgment that the perception failures can lead to safety failures).
- Concerns that the (parts of) the automotive industry are not treating the safety issues seriously enough – and the philosophy of "move fast and break things" should not be used in safety-critical environments (including automotive cars).
- Several comments regarding the reconciliation of uncertainty with predictability, and ensuring that this is an inherent design predicate.

Thank you!

- Correction/editions/clarifications are welcome (from authors and audience).