RC3: Resilient Computing and Cybersecurity Center

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

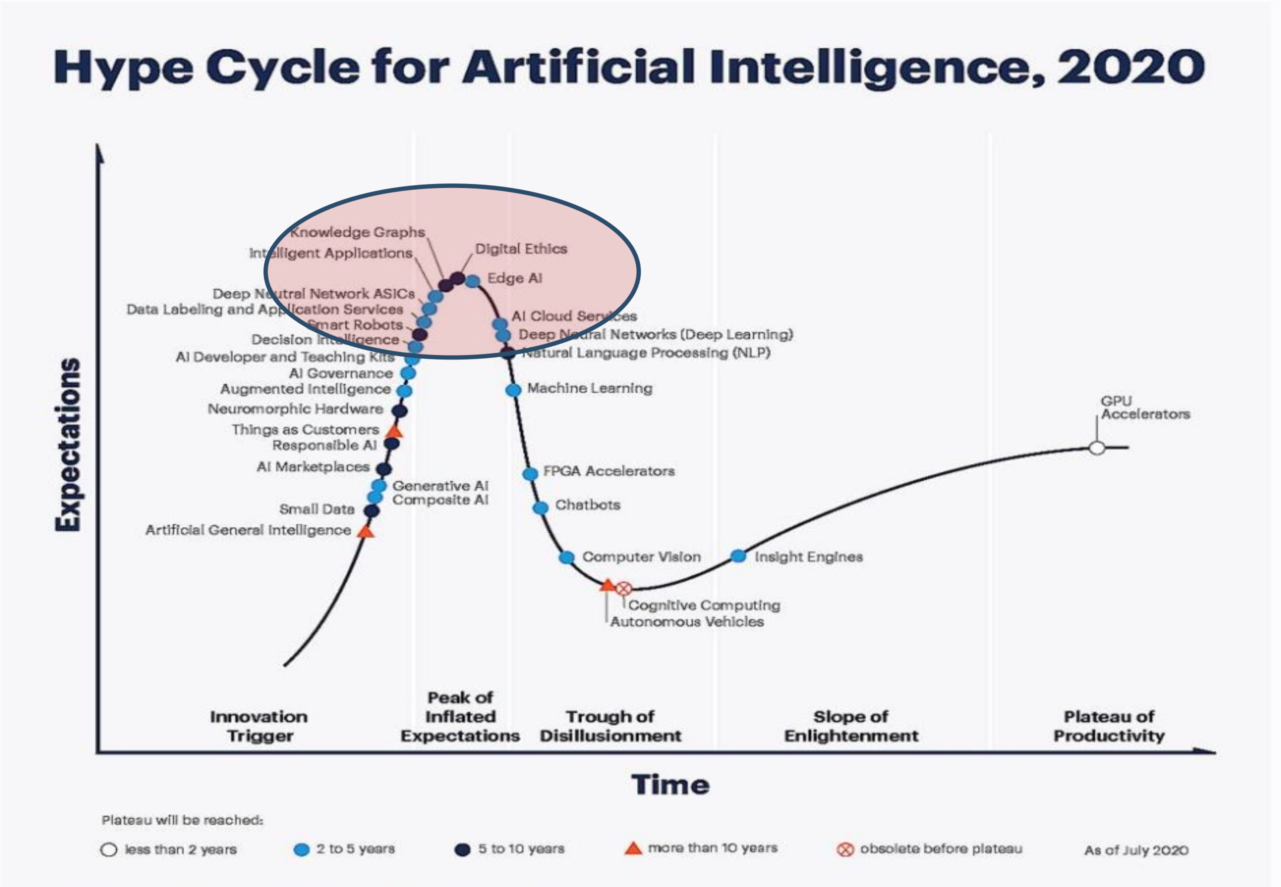
AI for Cybersecurity vs. Cybersecure AI: a chicken and egg problem?

https://rc3.kaust.edu.sa

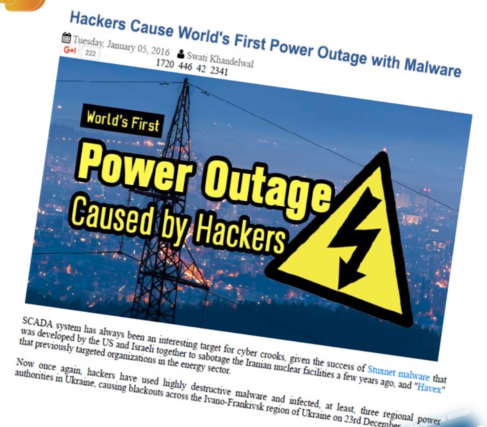
Paulo Esteves-Veríssimo, Professor, Director
King Abdullah University of Science and Technology, CEMSE
Resilient Computing and Cybersecurity Center – RC3
paulo.verissimo@kaust.edu.sa
https://cemse.kaust.edu.sa/people/person/paulo-verissimo

Int'l Workshop on Workshop on Cyber Resilient Systems, 4-7th June –Cascais, Portugal

Enter AI, ML, DNN, ...



AI vs. Security vs. Safety



Hackers Cause World's First Power Outage with Malware

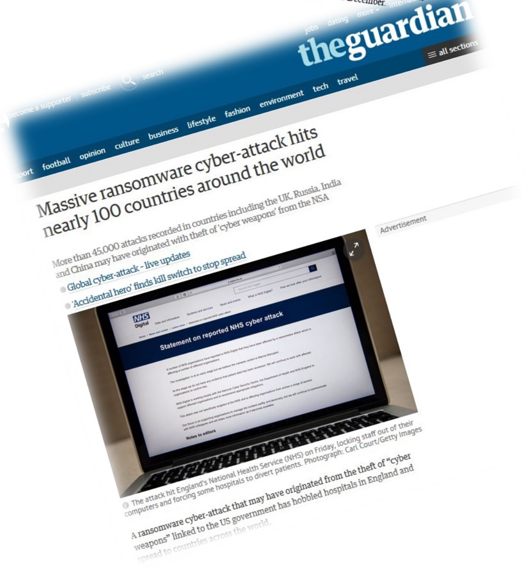
Tuesday, January 05, 2016 Swati Khandelwal

G+1 222

1720 446 42 2341

SCADA system has always been an interesting target for cyber crooks, given the success of Stuxnet malware that was developed by the US and Israeli together to sabotage the Iranian nuclear facilities a few years ago, and "Havex" that previously targeted organizations in the energy sector.

Now once again, hackers have used highly destructive malware and infected, at least, three regional power authorities in Ukraine, causing blackouts across the Ivano-Frankivsk region of Ukraine on 23rd December.



jobs dating more International edition
theguardian
all sections
become a supporter subscribe search
sport football opinion culture business lifestyle fashion environment tech travel
Advertisement

Massive ransomware cyber-attack hits nearly 100 countries around the world

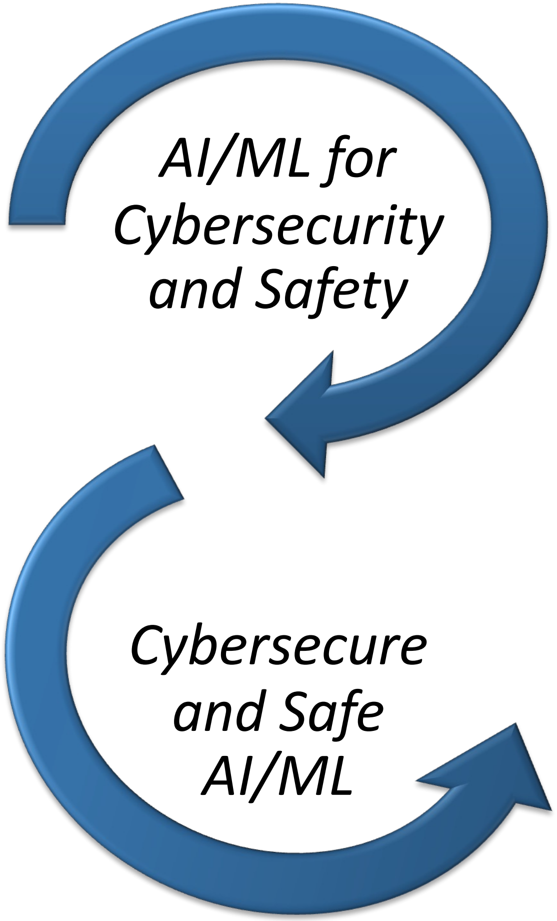
More than 45,000 attacks recorded in countries including the UK, Russia, India and China may have originated with theft of 'cyber weapons' from the NSA

- Global cyber-attack - live updates
- 'Accidental hero' finds kill switch to stop spread

NHS Digital
Statement on reported NHS cyber attack
Notes to editors

The attack hit England's National Health Service (NHS) on Friday, locking staff out of their computers and forcing some hospitals to divert patients. Photograph: Carl Court/Getty Images

A ransomware cyber-attack that may have originated from the theft of "cyber weapons" linked to the US government has hobbled hospitals in England and spread to countries across the world.



AI/ML for Cybersecurity and Safety

Cybersecure and Safe AI/ML

TESLA'S AUTOPILOT HAS HAD ITS FIRST DEADLY CRASH

Woman dead after being struck by self-driving Uber

FSD 10.4 tries to turn into incoming traffic, AGAIN!

Homogeneous ML/DNN-based systems cannot give strong assurance guarantees

- Status-quo
  - *Autonomous cars use ML-powered multi-sensor perception and complex control logic, and sometimes redundant modules to which they hand over in case of problems.*
- Assurance
  - *Infeasible to provide reliable figures/conclusions*
  - *Impossible to certify under current best practices*

Tesla radar did not recognize a camel, causing an accident in the UAE

Can we leverage the best of the *security and dependability* fields?

- In particular, dependability teaches us that our techniques:
  - (i) should identify the uncertainties and weaknesses exhibited at **component level**,
  - and (ii) craft mechanisms that address them, to produce predictably correct **system-level** results.
- *Result (ii) always conditioned by how well we did (i)*

Hybrid ML-based systems may help

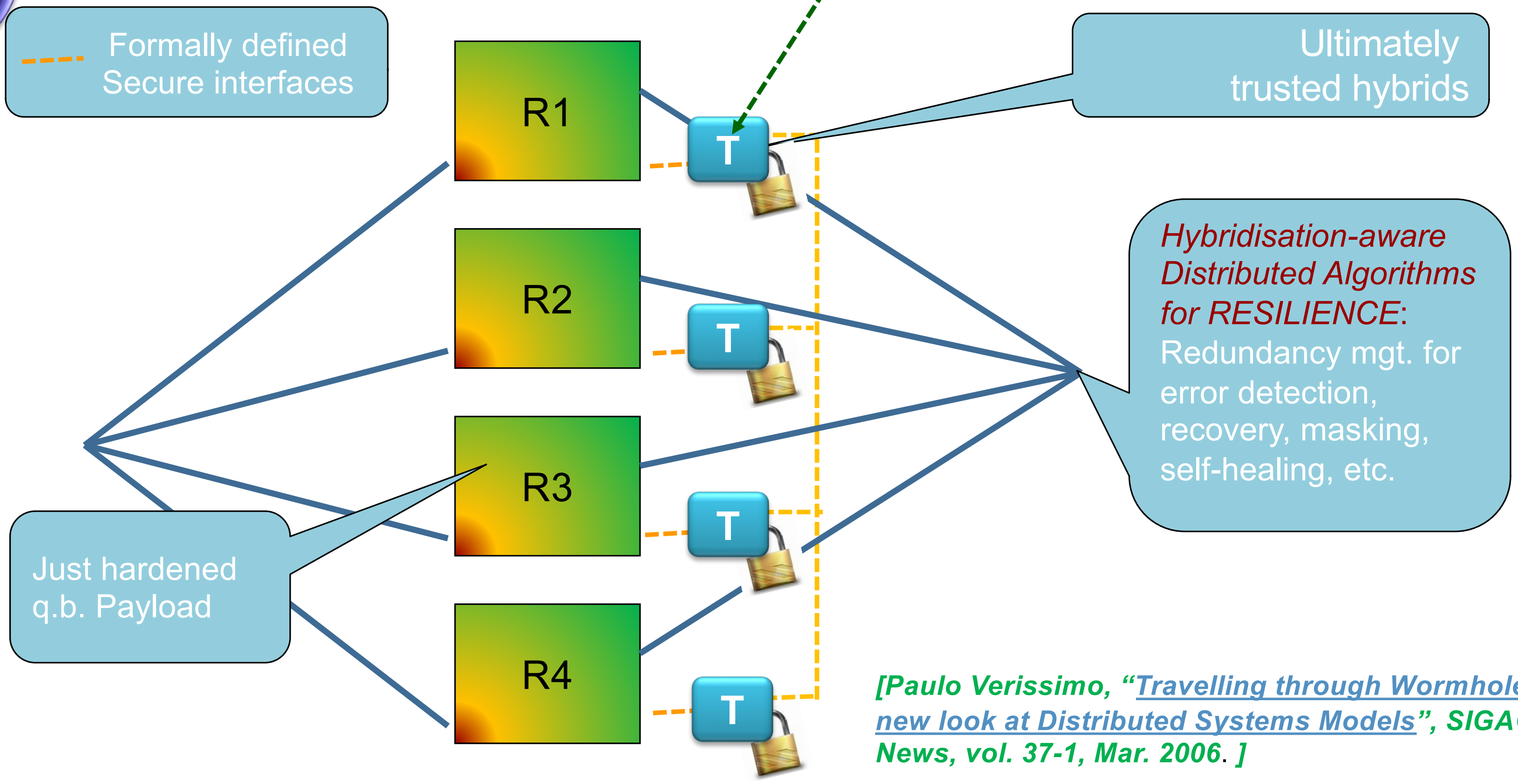
an autonomous vehicles ecosystem example

- Redundancy – *these components cooperate redundantly to achieve the end goal of safety*
  - *Replication*
  - *Reconfiguration, hand-over*
  - *Take-over*
  - *Diversity, for malicious faults*
- Hybrid architecture - *Autonomous cars having different realms running under different assumptions*
  - *Hybrid system and fault assumptions ("hierarchy of functions")*
  - *Modular*
  - *Distributed.*
- Assurance – *enablers of the goal*
  - *Recent Hybrid Logic of Events allows verifying architecturally hybrid systems by proof assistants.*
  - *Trusted-trustworthy hybrids anchor the global trustworthiness, through proof of the Lifting predicate*

Hybridisation-aware distributed algorithms, models, and architectures

Hybridisation-aware algorithms: models, architect., and control

Leveraging trusted-trustworthy components and TEE, with the right set of simple functions (failure detectors, monotonic counters, reliable timers and clocks, PRG, signatures, indelible logs, binary consensus)



Formally defined Secure interfaces

Ultimately trusted hybrids

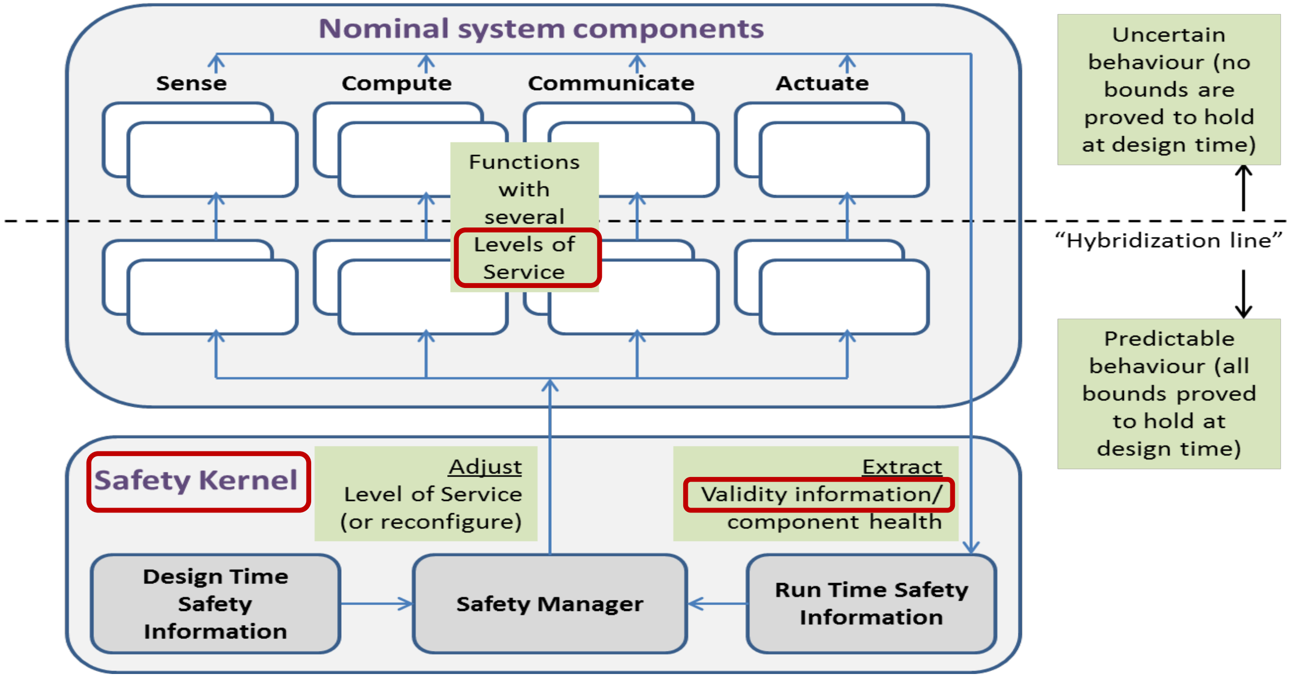
Hybridisation-aware Distributed Algorithms for RESILIENCE: Redundancy mgt. for error detection, recovery, masking, self-healing, etc.

Just hardened q.b. Payload

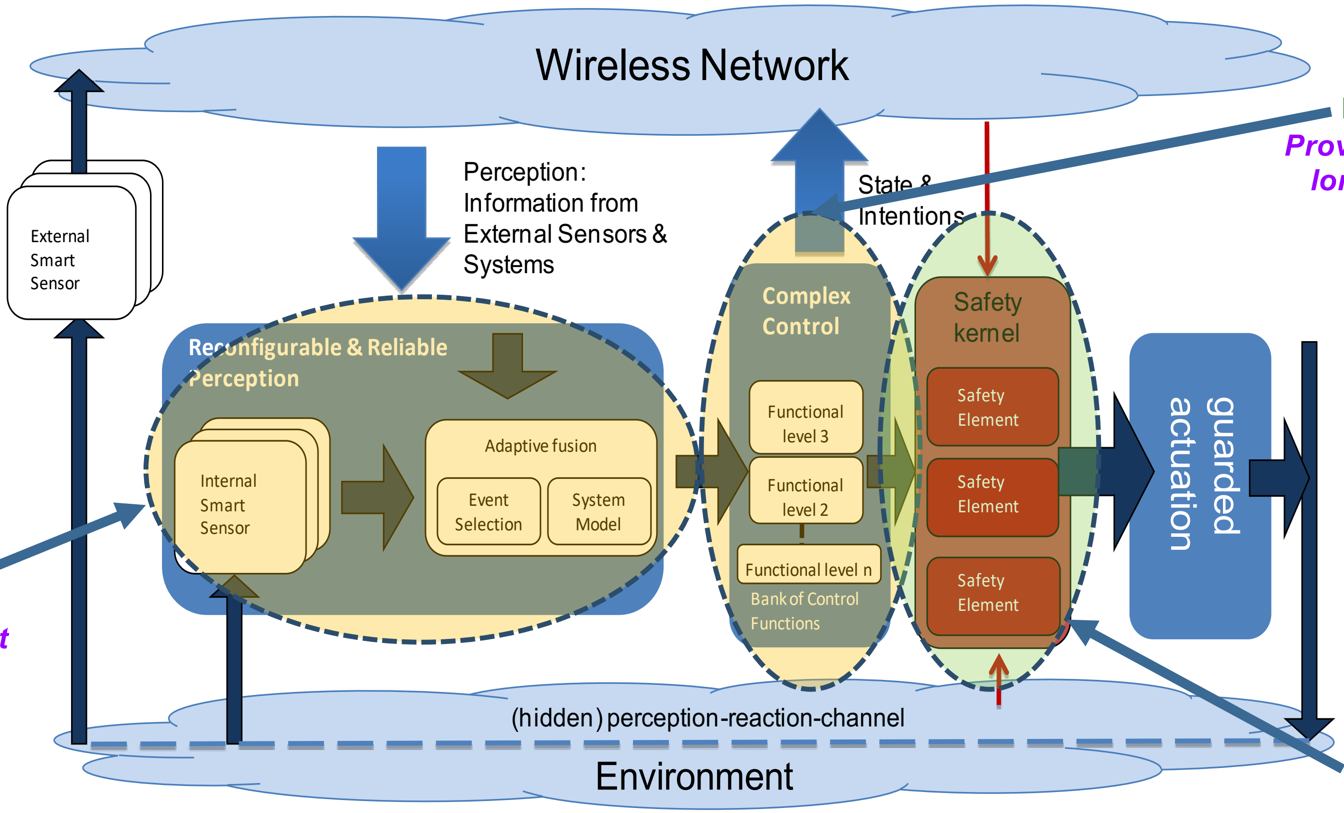
[Paulo Verissimo, "Travelling through Wormholes: a new look at Distributed Systems Models", SIGACT News, vol. 37-1, Mar. 2006.]

KARYON architecture: proof of concept of hybridisation for safety

- **Main Concepts:**
  - Level of Service
  - Data validity
  - Safety kernel


Nominal system components
Sense
Compute
Communicate
Actuate
Functions with several Levels of Service
Uncertain behaviour (no bounds are proved to hold at design time)
"Hybridization line"
Predictable behaviour (all bounds proved to hold at design time)
Safety Kernel
Adjust Level of Service (or reconfigure)
Extract Validity information/ component health
Design Time Safety Information
Safety Manager
Run Time Safety Information

Karyon Architecture



HYBRIDISATION BY LEVEL OF SERVICE

Proven functional safety, as long as assumptions for level are satisfied

ABSTRACT SENSOR MODEL

integrity attributes are evaluated against a set of different safety requirements

SAFETY KERNEL HYBRID

takes over, toward fail-operational or fail-safe termination