No Trust without Regulation!
The European Challenge on Regulation, Liability and Standards

François Terrier
Program Director of List Institute
CEA’s AI Senior Fellow
From where I talk...
✓ Low-carbon energy (nuclear & renewable)
✓ Digital technology
✓ Medicine of the future (technology)
✓ Defence and national security
Smart Digital Systems Institute

- Society
- Environnement
- Sovereignty

- Artificial Intelligence
- Digital Trust
- Computing solutions
- Industry of the Future

1,000+ employees
200+ PhD & Postdocs
200+ industrial partners
Main CEA List’s axes of research in AI

Science for AI

Algorithm frugality:
Data management, optimisation, embedded

System Trust:
Methods and tools

Energy efficiency:
Components, board, architectures, Microelectronics technologies

AI for Science and Society

Applications focussed on our expertises:
- Vision, Image, Video
- Natural Language Processing
- Security
- Energy
- Mobility
- Health
- Material science
- IoT, Telecom.
- Industry of the future

~200 pers List
The well-known context...
AI is coming with new and huge challenges

More and more expectations on trust and frugality

A lot of new applications

Automatization

A lot of data, computation

Trust challenge

No common sense

Fragile

Frugality challenge

Fragile + Noise =

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has its limits

From Cloud to Edge...
Efficient and impressing!!!
on elementary task as
  • Perception
  • Reasoning

\[ \text{No common sense} \]

\[ \text{Fragile} \]
Scolar bus + Ostrich =
\[ \text{Nothing (recognized) behind?} \]

\[ \text{Attackable} \]

\[ \text{Miss-used} \]

Report on Tesla first accident - Recommendation
Incorporate system safeguards that limit the use of automated vehicle control systems to those conditions for which they were designed. (H-17-41)
Need of policy
Outside of Europe:
still at stage of recommandations...

The AI RMF is intended for voluntary use in addressing risks in the design, development, use, and evaluation of AI products, services, and systems.

https://futureoflife.org/open-letter/pause-giant-ai-experiments/

https://oecd.ai/en/ai-principles

...innovative and trustworthy and that respects human rights and democratic values. (May 2019)
European approach to ethics and regulation

**European commission vision**

2018: High-Level Expert Group on Artificial Intelligence (set up by the European Commission)

2019: Refinement for key sectors

2020: Approach for excellence and Trust

2021: Proposal to the parliament

**Expert group analyse the subject of TRUST**

- Manufacturing
- Health
- Justice
- E-Government

Trustworthy AI should be:
- (1) lawful - respecting all applicable laws & regulations
- (2) ethical - respecting ethical principles and values
- (3) robust - both from a technical perspective while taking into account its social environment

**Human-centric AI:**
- AI system builder is responsible
  - robustness, safety, privacy, transparency...
- Human right must be respected and not subject to automated decision only

**European commission vision**

- Lawful
- Ethical
- Robust

**Toward a european regulation**

**Ethics imperatives & defense**

Europe makes a step forward!
A global policy set up by Europe

A complete approach with 3 pillars: regulation, liability, conformity

Conformity (Standards)

Regulation

Data Act

AI liability

Liability

ETHICS GUIDELINES FOR TRUSTWORTHY AI

Quality & Technical
Toward an European regulation for AI deployment respecting the european values

• A strong European legislation, *(i.e applicable as it is to any system or service provided in any EU country)*

**Adopted by EU Parliament, on June 14th 2023: 499 votes in favour, 28 against and 93 abstentions ➔ Now it goes to each national parliaments**

- List of prohibited AI
- Risk classification
- Rules for high risk AI systems
- Transparency obligations
- Support to innovation

---

Toward an European regulation: centered on the usage

**FORBIDDEN USAGES**

- **COGNITIVE BEHAVIOURAL MANIPULATION**
  - AI systems that deploy harmful manipulative 'subliminal techniques'
  - AI systems that exploit specific vulnerable groups (physical or mental disability)

- **SOCIAL SCORING & Cie by PUBLIC AUTHORITIES**
  - AI systems used by public authorities, or on their behalf, for social scoring purposes
  - Predictive policing systems (based on profiling, location or past criminal behavior)
  - Emotion recognition in law enforcement, border management, the workplace, educational institutions

- **REAL TIME & REMOTE BIOMETRIC IDENTIFICATION**
  - “Real-time” and “Post” remote biometric identification systems in publicly accessible spaces (except for serious crimes after judicial authorization)
  - Biometric categorization using sensitive characteristics (e.g. gender, race, ethnicity, citizenship status, religion, political orientation)
  - Untargeted scraping of facial images from internet or CCTV to create facial recognition databases (= violating human rights and right to privacy).
Toward an European regulation: centered on the usage and risk analysis...

3 levels of risks depending of the usage domain + 1 Specific case

- High risk systems: Safety obligations
- Generative AI: Dedicated Transparency obligations
- Limited risk: Transparency obligations
- Low or minimal risk: No obligations
High risk systems

- FOR ALL DOMAINS WITH EXISTING REGULATIONS:
  ➔ General principles + requirement of corresponding EU regulations ➔ strong req.
  - Ex.: transportation, health, energy, toys, lifts...

+ SPECIFIC APPLICATIONS AREAS (requiring registration) as:
  - Biometry & person categorization;
  - Management of critical infrastructure
  - Essential public and private services;
  - Education; Employment;
  - Recommendation systems (social media >45M users);
  - Democratic process influence;
  - Law enforcement; Migration etc.;
  - Administration of justice;

- For all others (applications not already governed by European legislation) ➔ self-assessment
Requirements

High-risk AI based systems

- **RISK ANALYSIS AND ESTIMATION (MAINTAINED REGULARLY) ACCORDING TO USE**
- Data set pertinent, representative, free of errors and complete
- Technical documentation establishing conformity to requirements
- Automatic recording of events (‘logs’)
- Sufficient transparency TO INTERPRET OUTPUT AND USE IT APPROPRIATELY

(continuously maintained during the system life)
Requirements

Generative AI based applications

- **Assess and mitigate possible risks** *(regarding to the possible uses and contexts of use)*
  (to health, safety, fundamental rights, the environment, democracy and rule of law)

- **Register** the models in the EU database before release on the EU market

- Comply with **transparency** requirements:
  - **Disclosing that the content was AI-generated**
  - **Ensure safeguards against generating illegal content**
  - **Provide publically detailed summaries of the copyrighted data used for training**
  - **Provide capabilities measuring and logging resource consumption** *(over their entire lifecycle)*

**Challenge is how to adopt a risk oriented approach when usages are not known in advance**
Requirements

Limited-risk AI based systems (not based on Generative AI)

- Such as systems that interacts with humans (i.e. chatbots), emotion recognition systems, biometric categorisation systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes) → limited set of transparency obligations:
  
  ➢ Interacting with natural persons: ensure natural persons are informed of the AI nature
  ➢ Disclose that the content has been artificially generated or manipulated
AI liability: a more protective law

The specific characteristics of AI make it particularly difficult to meet the burden of proof for a successful claim (e.g. opacity/lack of transparency, explainability, autonomous behaviour, continuous adaptation, limited predictability)

- Adaptation of law to allow for compensation for damages without the need to prove a fault
  - Reduce liability rules uncertainty and risk of legal fragmentation
  - Causality not mandatory (except for High Risk AI because they have to provide transparency and thus give the means to establish causality links)

- Responsibility to the provider of product and services (depending on context of use)
  - Will depend to impact analysis depending of the risks entailed by the uses
  - Will depend to quality of development, transparency, effective oversight by natural persons
European AI Regulation

Trust → Certification → Regulation → Standards

THE AI ACT

EU Artificial Intelligence Act: Risk levels

- Unacceptable risk: Prohibited
- High risk: Conformity assessment
- Limited risk: Transparency obligation
- Minimal risk: No obligation

Will influence regulations of the other countries → « the RGPD effect... »

2018 « HLEG » → 2021 « AI Act » → 2022 Normalisation → 2024... !

October 2022, request for: Producing standards supporting the regulation

December 2024: Agreed and harmonized standard

End of « technical work: December 2023!»

A very fast process

Standardisation

CENELEC

European organism

CEN

AI national commission

Coordinate French contributions

Normalisation

2018 « HLEG » → 2021 « AI Act » → 2022 Normalisation → 2024... !
Standards organization

Horizontal and general standard for AI Act

Harmonised with the existing ones

Mainly identify & reuse existing elements

- Risk management systems
- Governance and quality of datasets
- Record keeping through logging capabilities
- Transparency and information provisions
- Human oversight
- Accuracy specifications
- Robustness specifications
- Cybersecurity specifications
- Quality management systems
- Conformity assessment

A terminology and conceptual BACKBONE FOR DOMAIN STANDARDS

Defines the « What » = Concepts = Not the « How »

Credits: Henri Sohier, Co-project leader, “AI Trustworthiness characterization”, CEN-CENELEC JTC21 WG4
Labelling approach to complete certification

A virtual example

Label for system trust and/or company trust (or more)
Name of the original labels or standards

System attributes in case of system trust

Optional if it can be a self-declaration

Link to more information (the summary can never show everything)
What about safety?
TRUST challenge: a set of characteristics

ENGINEERING VIEW POINT

SAFETY, CERTIFICATION,
Quality, Reliability,
Security, Privacy
Robustness, Accuracy
Traceability, Interpretablity

USAGE VIEW POINT

ETHICS, SOCIETAL IMPACT,
Accountability,
Fairness, Explainability

TRANSPARENCY

SAFETY, CERTIFICATION,
Quality, Reliability,
Security, Privacy
Robustness, Accuracy
Traceability, Interpretablity

AI community see AI components as (physical) systems

AI community see AI components as SW

AI Act
Mastering the process → the specification
Operational Design Domain

« Operating conditions under which a given system is specifically designed to function »
(including environmental, time-of-day restrictions)

Various definitions and naming depending on the domains (concept being refined and integrated by the related standards)

- E.g.: SAE J3016, BSI PAS 1883, SAE AIR6988 (2021), {ConOps} EASA (2021), {SOD, OSED, ODD} SAE AS6983 / EUROCAE ED-xxx (Draft 3a – June 2022)

- « Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristic. » SAE-J3016

✓ from the outside of the system
Voluntary restriction within which the expected nominal functioning of an AI-based system is ensured.

E.g.: « Very heavy rainfall »

✓ from the inside of the system
Description of measurable foreseeable operating conditions within which a AI-based component must operate.

E.g.: « Signal variation from rain detection sensor »

Credits: M. Adedjouma et al. – www.confiance.ai
A general domain analysis completed during design

1. Road structure;
2. Road users, including...;
3. Animals;

... *


Topics to address

- Influence factors
  - Measurable characteristics
  - Marging & limits

E.g.: Cruising in its lane
E.g.: Capability to perceive lane markings
E.g.: Distance covered without lane detection; Color of markings...
E.g.: Multiple lane markings colored in roadworks areas

Credits: M. Adedjouma et al. – www.confiance.ai
ODD-based Hazard Identification

A tool-supported framework to automate the hazard identification process using an ontological model and a specification of the AI-based system ODD.

G. Ollier et al., Using Operational Design Domain in Hazard Identification, EDCC 2022
Example on ongoing work on dataset collection, labeling, classification vs the ODD

Credits: M. Addejouma, P. Toukam, F.-M. Ngole Mboula et al. – www.confiance.ai
Formal methods...
SAFETY: FORMAL METHODS AND AI... ???

Formal methods a long time ago: The prophecy

1979: “Program verification is bound to fail. We can’t see how it’s going to be able to affect anyone’s confidence about programs”, in Social processes and proofs of theorems and programs”, Communications of ACM.

The prophets - By Richard De Millo, Richard Lipton, and Alan Perlis.
- Distinguished Professor of Computing at the Georgia Tech
- VP and CTO of Hewlett-Packard
- ACM, Carnegie Mellon, Princeton, Georgia Tech
- Yale, Berkeley, Yale, Purdue
- First Turing Award recipient
- ACM, Carnegie Mellon, Yale, Purdue
- Knuth Prize winner

Fast forward a few decades

What about AI and the hardest = ML: we have been here before…

Formal methods
- Symbolic AI
- Perturbation robustness
- Properties verification
- Model interpretation

Performance tests could be safety tests only if they are well defined regarding to the risks and their probabilities…

What happens when a test fails from a safety point of view?
→ usually we ask to correct it, but it seems not applicable in ML...

The key question is: how are they defined & what is the coverage?
Goal: evaluate the AI component against perturbations

- AI component as a black box
- Formalised perturbations according ODD
- Automatic test generation (sample)
- Compare output with the expected one

  - E.g: « Metamorphism »
  (geometric transformation = « any computable math formulae »)

Welding control

How the application is robust against image degradations?

Metamorphic testing, example

E.g.: evaluate robustness against luminosity variations

Credits: Confiance.ai – AIMOS - www.confiance.ai/ia-a-epreuve-du-bruit
Proof of safety properties on neural networks

ACAS-XU: a complex problem (15 M of states) → function as set of 45 connected neural networks

Is tractable through formal methods analysing the NN
- E.g. PyRAT: pyrat-analyzer.com

If the intruder is near and approaching from the left, the system advises « strong right ».

Input constraints:
- $250 \leq \rho \leq 400$
- $0.2 \leq \theta \leq 0.4$
- $-3.141592 \leq \psi \leq -3.141592 + 0.005$
- $100 \leq v_{own} \leq 400$
- $0 \leq v_{int} \leq 400$
**Formal verification of ML based systems**

A use case: Detection of mooring line breaks

1) Formal verification of safety properties (*safe operating domain*)
   → exhaustive computation of input domains for which the system detects failures

2) Robustness evaluation of the sensitivity to disturbances on the inputs:
   - Global approach: computation of interval of acceptable perturbations


Fake example: always safe with waves under 5 m
Formal methods challenges

They require initial « formalisation » of properties or test objectives (ODD)
If properties are not formalised (e.g. « recognize a pedestrian »),
we need to go to internal robustness properties (e.g. « metamorphism »)

For global verification: OK with structured data (leading to « small » models)
  ■ E.g.: command and control, tabular approaches...
    
    But for « large » models, very huge computation is needed.

They apply well on local verification = around a particular data input
  ■ E.g.: welding control application, the image size lead to a large NN,
    → practical computation allowed to explore until 5% of variation around any given input data

Research is very active and provides continuously new results
to deal with more and more complex AI components
Imagine new usages & understand real needs

Applications

Design, V&V Certification

Deployment technologies HW/SW

Performance Cost

New tools
New process

TRUST & FRUGALITY will make the difference

francois.terrier@cea.fr

Thanks!