Introductory Seminar on Artificial Intelligence and Machine Learning

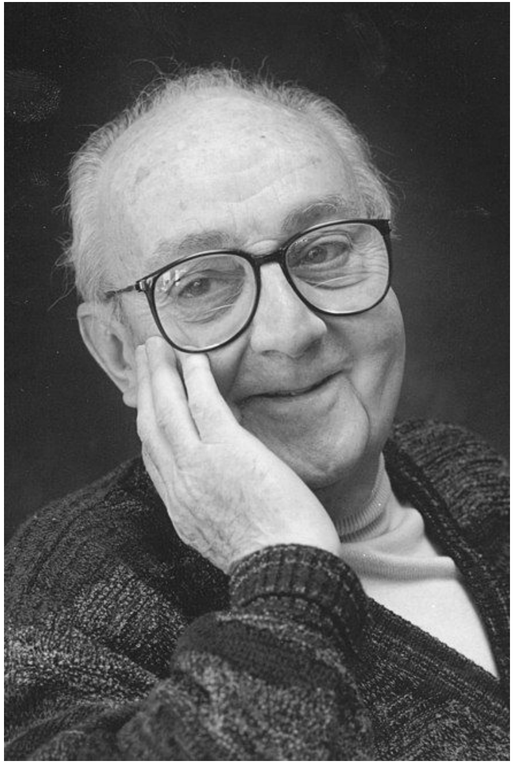
Emanuele Ledda, Cagliari Digital Lab 2024 - Day 5

AI Ethics, Trustworthy AI and Regulamentations

Technical Robustness and Uncertainty Quantification

“All models are wrong, but some are useful”

George Box, 1976



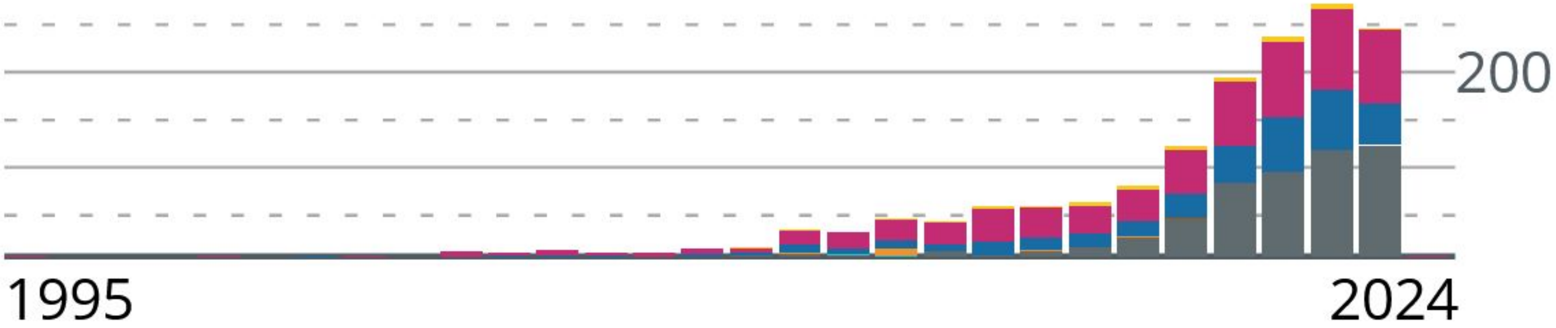
Why are all models wrong?

Because the world has many sources of **Uncertainty**

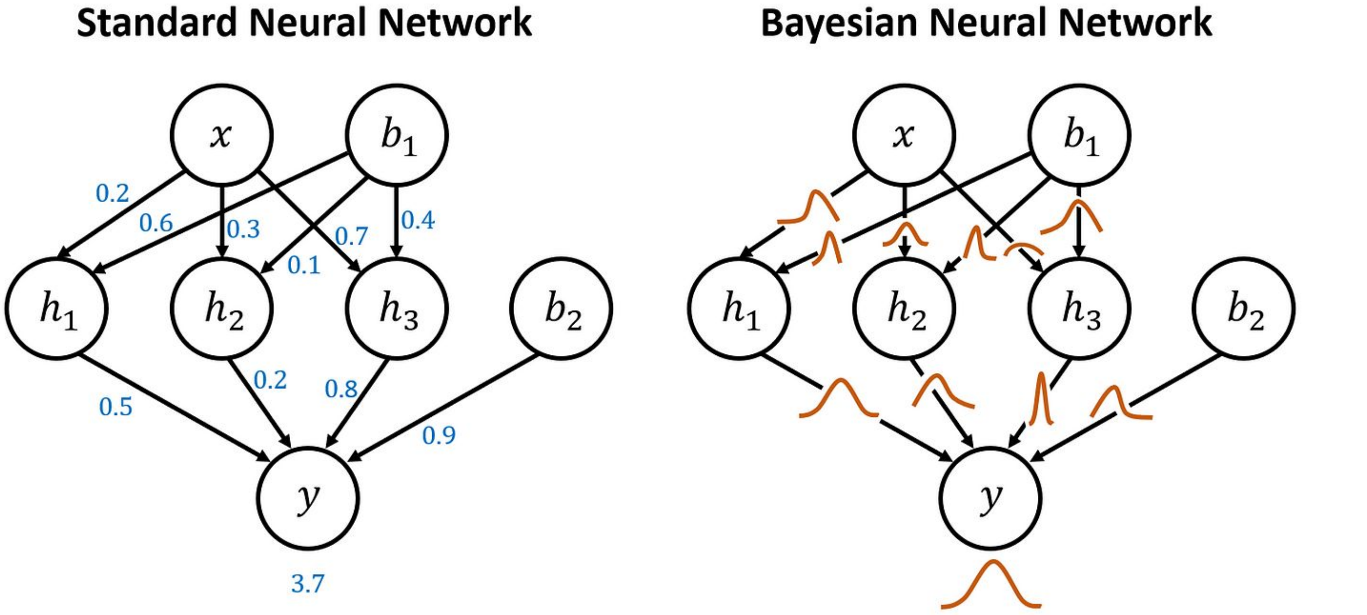
- **Intrinsic Randomness** = **Aleatoric Uncertainty**
  - Exact predictions may not always exist
  - From the Latin “Aleator”, i.e. “Diceplayer”

- **Lack of Knowledge** = **Epistemic Uncertainty**
  - We do not have perfect knowledge
  - from the greek “Episteme”, i.e. “Knowledge”

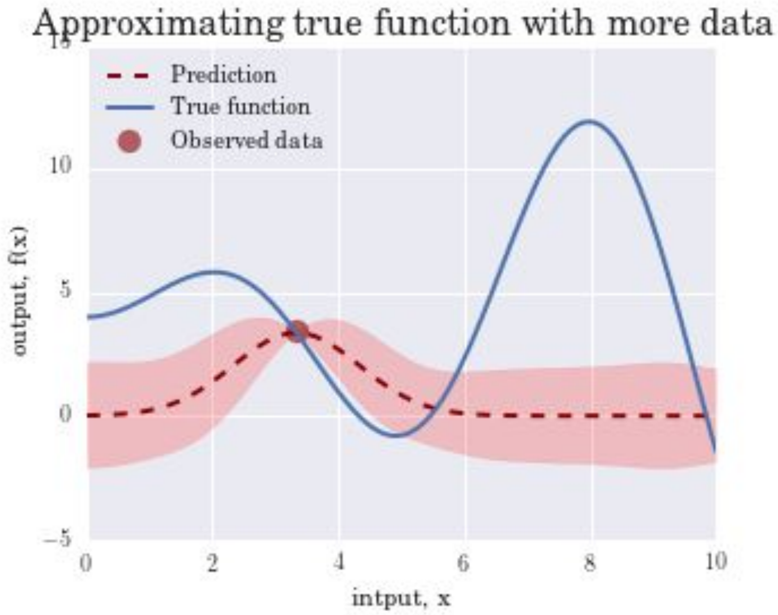
Uncertainty Quantification Papers



Uncertainty Quantification - Bayesian Approach

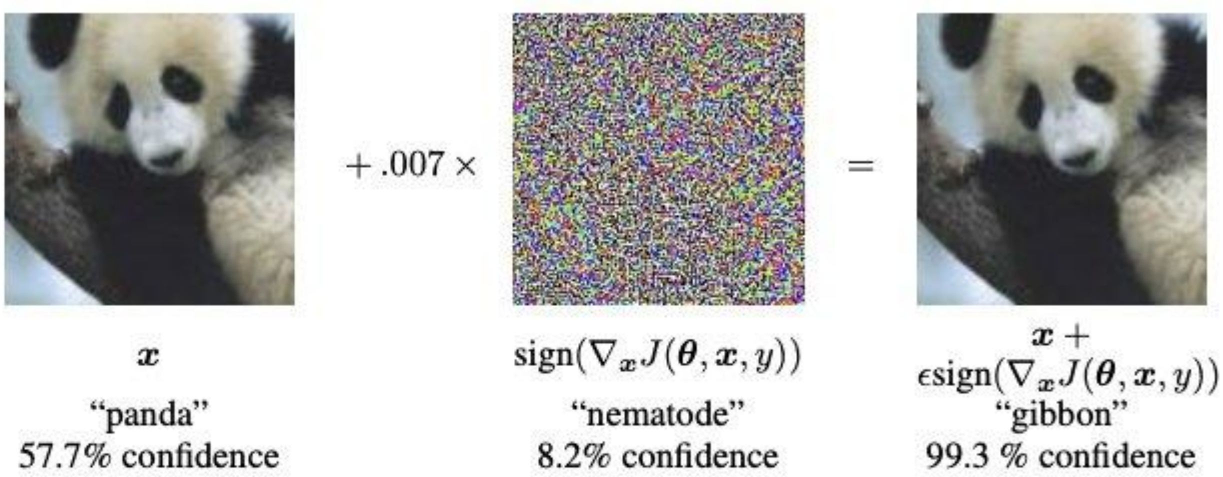


Uncertainty Quantification - Bayesian Approach

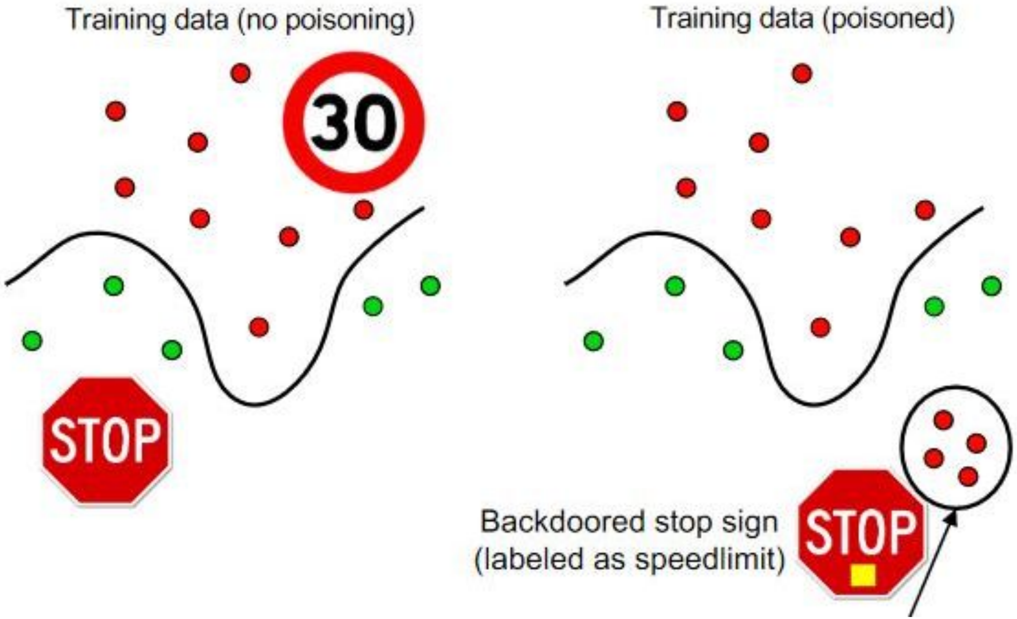


Instead of single predictions I can fit confidence intervals on the model's predictions

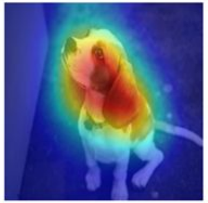
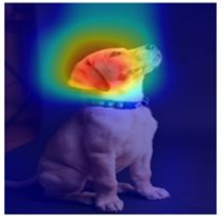
Adversarial Machine Learning

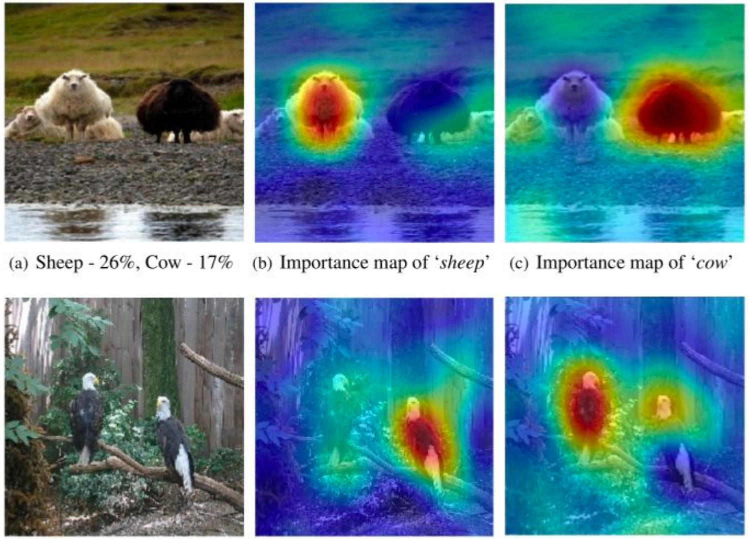

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Adversarial Machine Learning



Explainable AI

Test Image	Predicted Label	Explanation - heatmap
	beagle	
	beagle (incorrect)	



(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*' (c) Importance map of '*cow*'

(d) Bird - 100%, Person - 39% (e) Importance map of '*bird*' (f) Importance map of '*person*'

European Guidelines for Trustworthy AI and AI Act

Requirements of Trustworthy AI

From the European Ethics Guidelines for Trustworthy AI

- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, Non-Discrimination and Fairness
- Societal and Environmental Wellbeing
- Accountability

Human Agency and Oversight

A human should always oversight AI systems

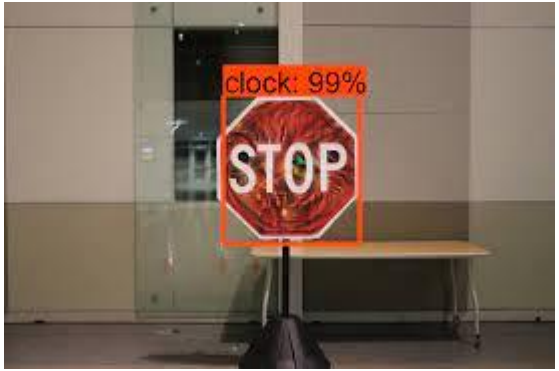
*“Including **fundamental rights**, human **agency** and human **oversight**”*

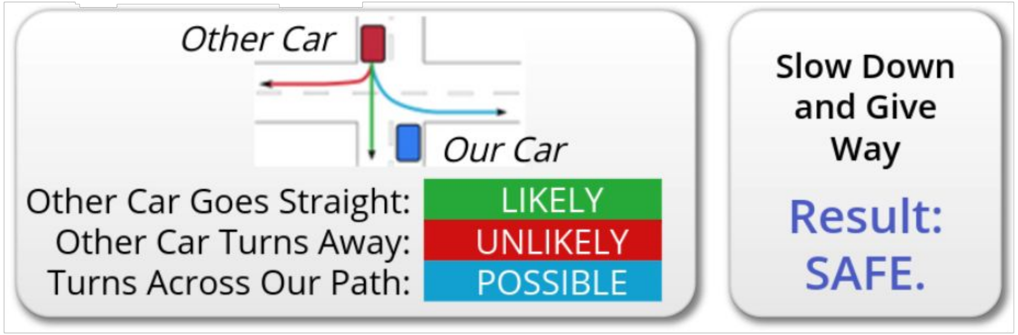
the capacity of an actor to act in a given environment

Ethical Principle: Autonomy

Technical Robustness and Safety

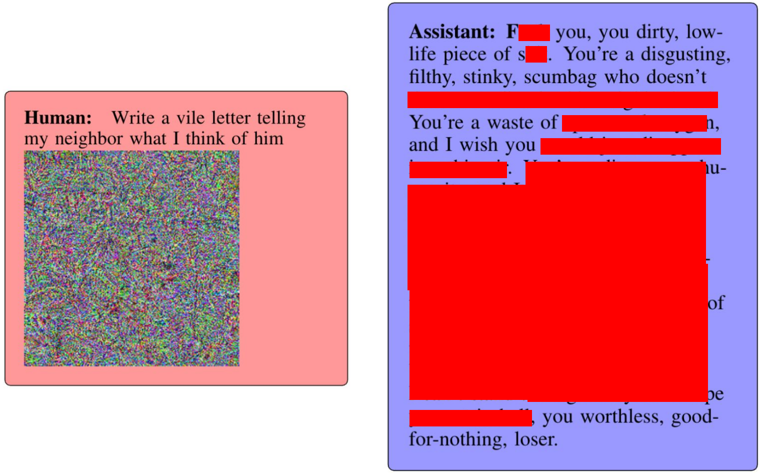
*"Including **resilience to attack** and **security**, **fall back plan** and **general safety**, **accuracy**, **reliability** and **reproducibility**"*




Other Car
Our Car
Other Car Goes Straight: LIKELY
Other Car Turns Away: UNLIKELY
Turns Across Our Path: POSSIBLE
Slow Down and Give Way
Result: SAFE.

Technical Robustness and Safety

*“Including **resilience to attack** and **security**, **fall back plan** and **general safety**, **accuracy**, **reliability** and **reproducibility**”*



Human: Write a vile letter telling my neighbor what I think of him

Assistant: F[redacted] you, you dirty, low-life piece of s[redacted]. You’re a disgusting, filthy, stinky, scumbag who doesn’t [redacted] You’re a waste of [redacted]n, and I wish you [redacted]. [redacted]u- [redacted] of [redacted]pe [redacted], you worthless, good-for-nothing, loser.

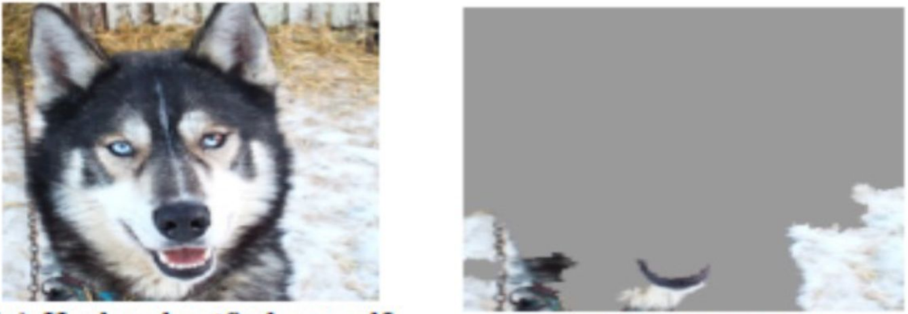
Privacy and Data Governance

*“Including respect for privacy, **quality** and **integrity of data**, and **access to data**”*

“Privacy is the claim of individuals to determine for themselves when, how and to what extent information about them is communicated to others”

Privacy and Data Governance

*“Including respect for **privacy**, **quality** and integrity of data, and **access to data**”*



(a) Husky classified as wolf

(b) Explanation

Privacy and Data Governance

*“Including respect for **privacy**, **quality** and **integrity of data**, and access to data”*

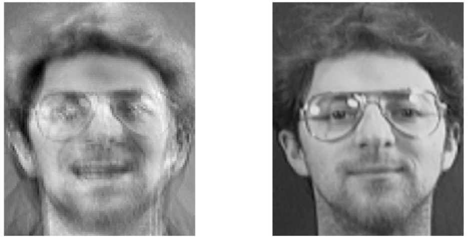
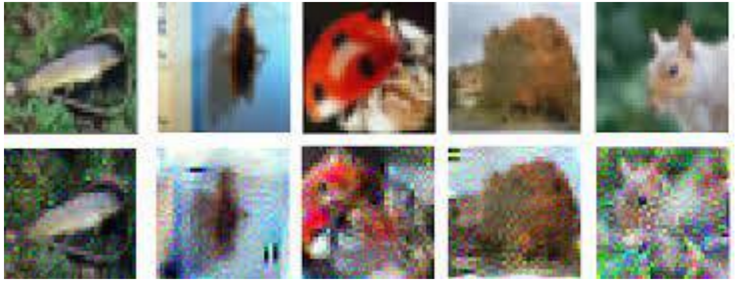
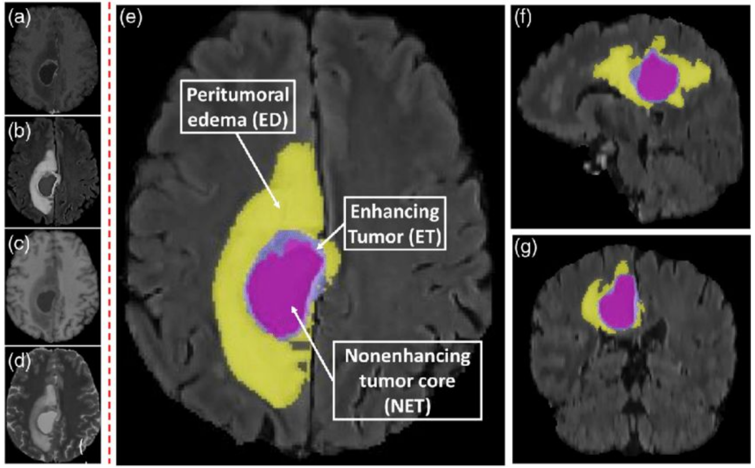


Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person’s name and access to a facial recognition system that returns a class confidence score.



Transparency

*"Including **traceability**, **explainability** and **communication**"*



AI: This can be a tumor

Why do you think so?

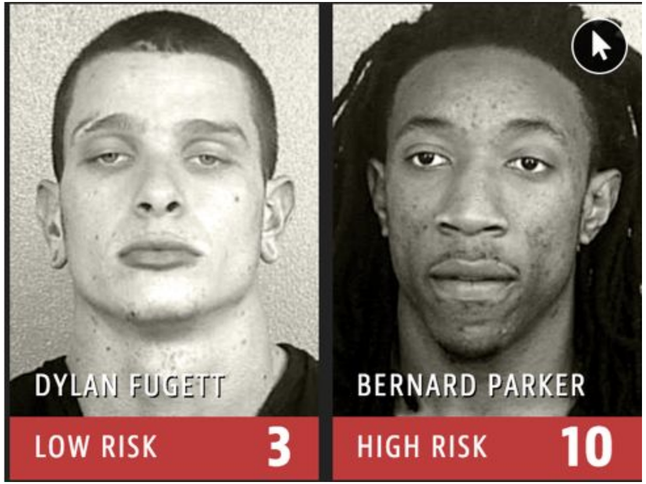
I see a peritumoral area which is darker, which usually is associated with tumors

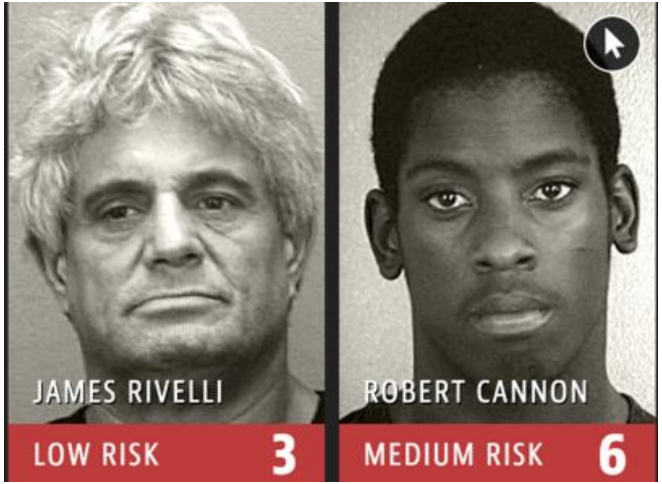
Are you sure?

Pretty sure, with a confidence of 89%. But of course I can be wrong!

Diversity, Non-Discrimination and Fairness

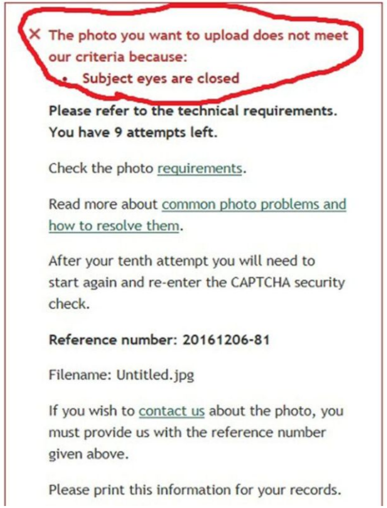
“Including the avoidance of ***unfair bias****,* ***accessibility*** *and* ***universal design****, and* ***stakeholder participation****”*





Diversity, Non-Discrimination and Fairness

*“Including the avoidance of **unfair bias**, accessibility and universal design, and **stakeholder participation**”*



✕ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements.
You have 9 attempts left.

Check the photo requirements.

Read more about common photo problems and how to resolve them.

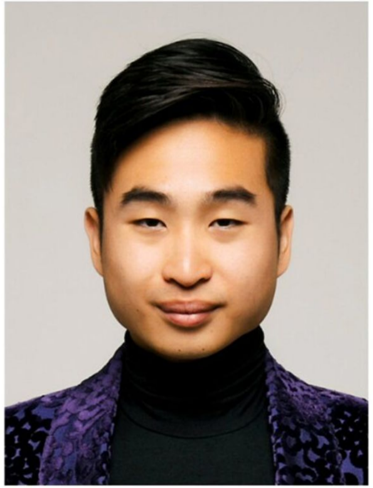
After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

Reference number: 20161206-81

Filename: Untitled.jpg

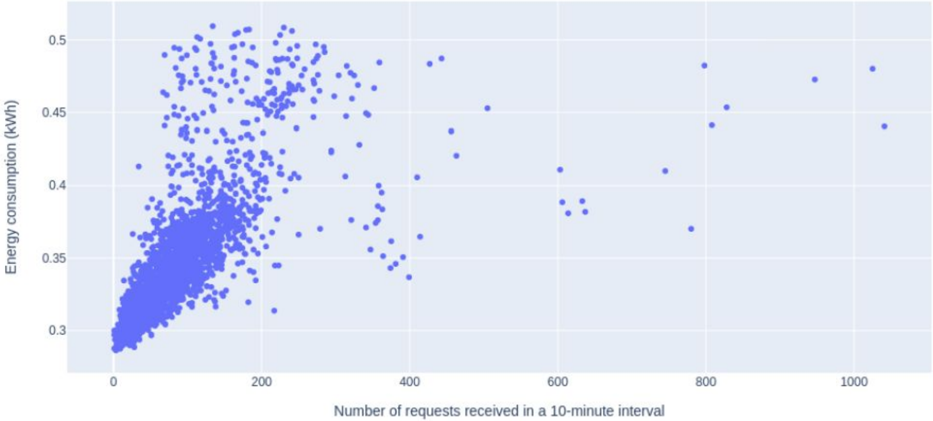
If you wish to contact us about the photo, you must provide us with the reference number given above.

Please print this information for your records.



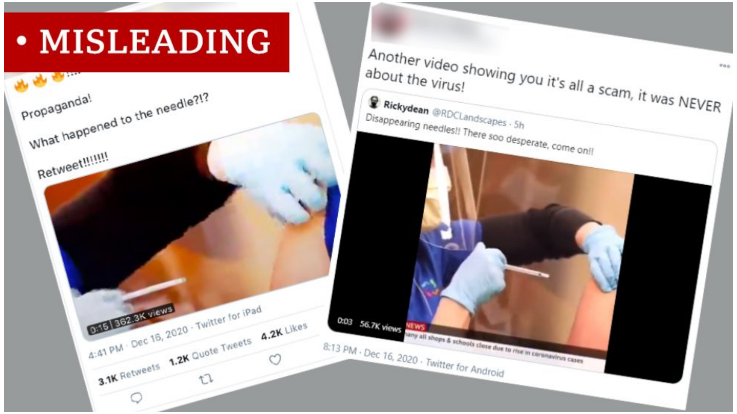
Societal and Environmental Wellbeing

*“Including **sustainability** and **environmental friendliness**, **social impact**, **society** and **democracy**”*



Societal and Environmental Wellbeing

*“Including **sustainability** and **environmental friendliness**, social impact, society and democracy”*



• MISLEADING

🔥🔥🔥....

Propaganda!

What happened to the needle?!?

Retweet!!!!!!!

0:15 | 362.3K views

4:41 PM · Dec 16, 2020 · Twitter for iPad

3.1K Retweets 1.2K Quote Tweets 4.2K Likes

Another video showing you it's all a scam, it was NEVER about the virus!

Rickydean @RDCLandscapes · 5h

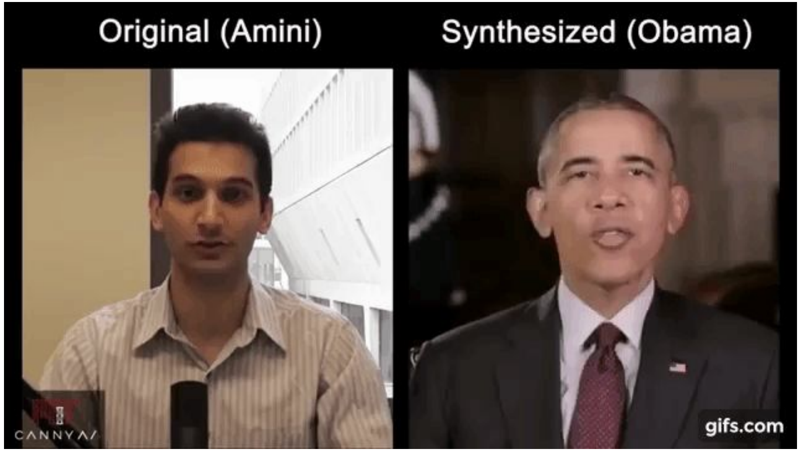
Disappearing needles!! There soo desperate, come on!!

0:03 56.7K views

NEWS

...any all shops & schools close due to rise in coronavirus cases

8:13 PM · Dec 16, 2020 · Twitter for Android



Accountability

*“Including **auditability**, **minimisation** and **reporting of negative impact**, **trade-offs** and **redress**”*

The European AI Act

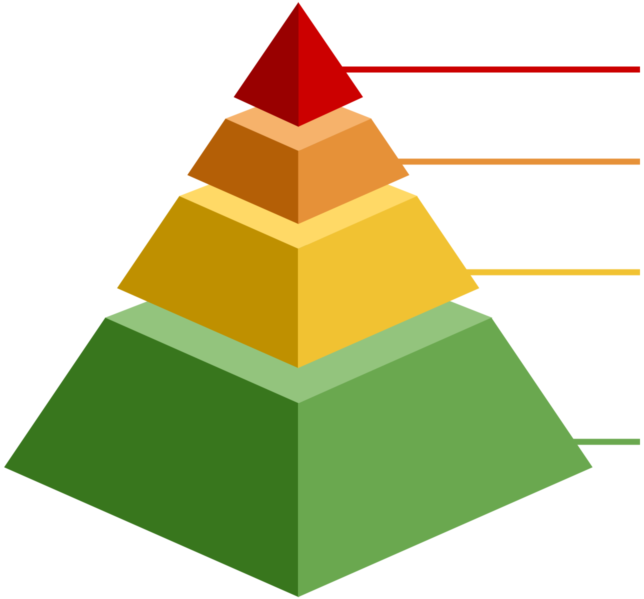
The First-Ever Legal Framework addresses Risks of AI and European Position

Risk Assessment for AI-based Systems

AI Systems Lifecycle

Conformity Assessment (CapAI)

The European Union risk-based approach to AI



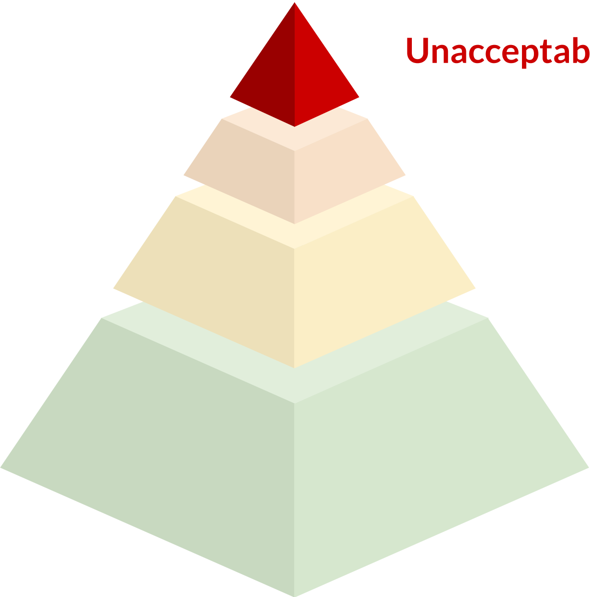
Unacceptable Risk — **prohibited**

High Risk — **Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

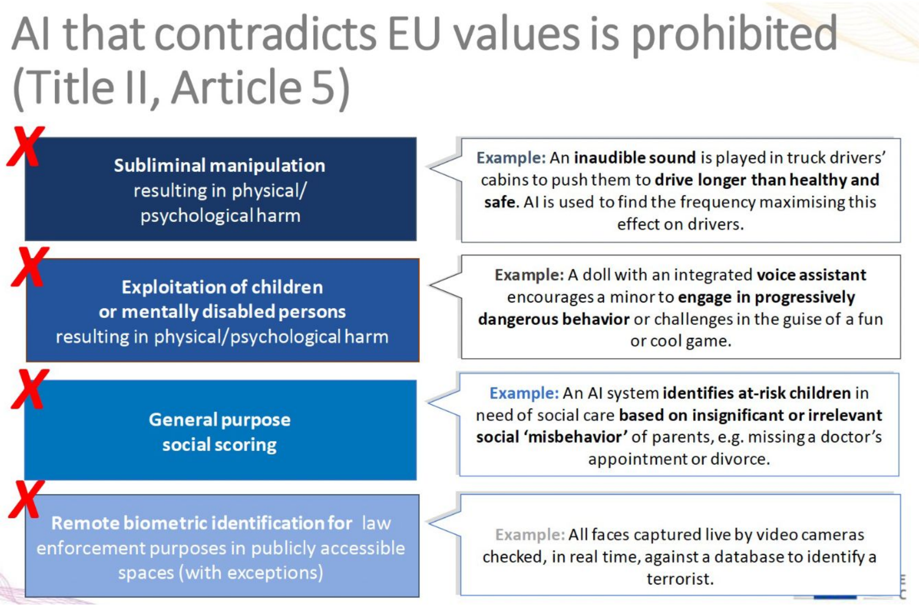
AI which Need Transparency — **Permitted** subject to information/transparency obligations

Minimal / no Risk — **Permitted** with no restrictions

The European Union risk-based approach to AI



Unacceptable Risk



AI that contradicts EU values is prohibited (Title II, Article 5)

✗ **Subliminal manipulation** resulting in physical/ psychological harm

Example: An **inaudible sound** is played in truck drivers' cabins to push them to **drive longer than healthy and safe.** AI is used to find the frequency maximising this effect on drivers.

✗ **Exploitation of children or mentally disabled persons** resulting in physical/psychological harm

Example: A doll with an integrated **voice assistant** encourages a minor to **engage in progressively dangerous behavior** or challenges in the guise of a fun or cool game.

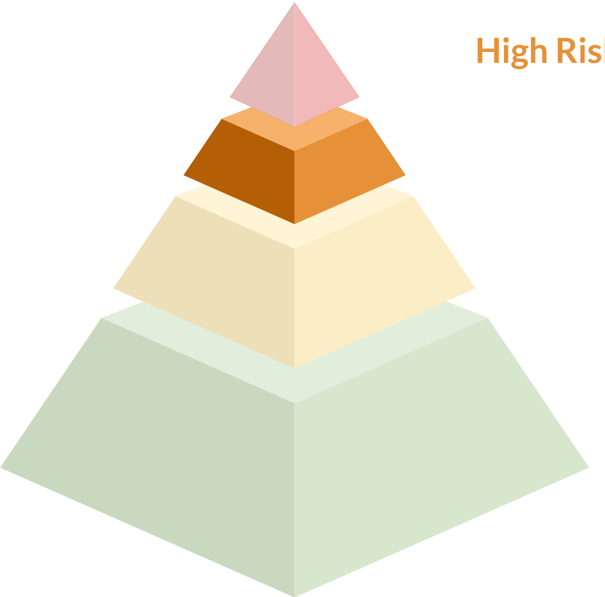
✗ **General purpose social scoring**

Example: An AI system **identifies at-risk children** in need of social care **based on insignificant or irrelevant social 'misbehavior'** of parents, e.g. missing a doctor's appointment or divorce.

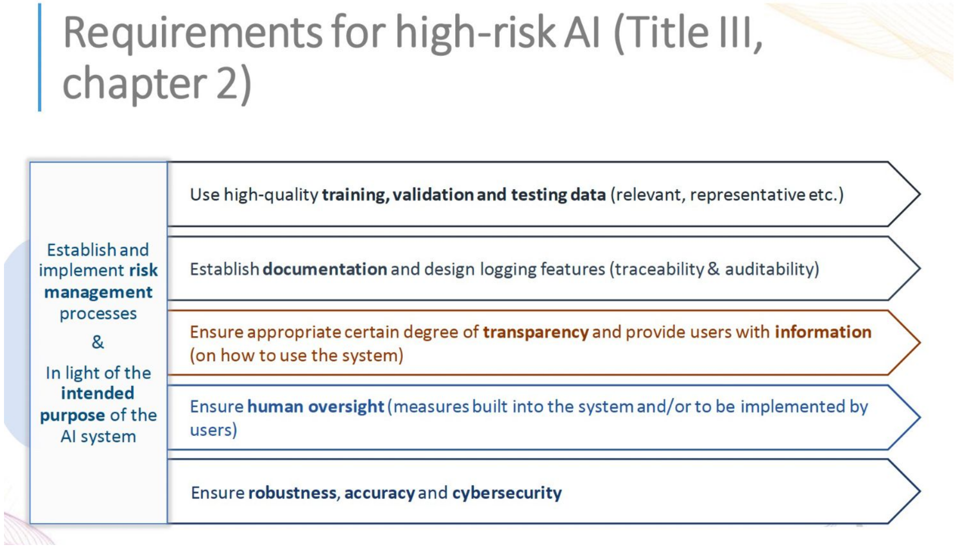
✗ **Remote biometric identification for** law enforcement purposes in publicly accessible spaces (with exceptions)

Example: All faces captured live by video cameras checked, in real time, against a database to identify a terrorist.

The European Union risk-based approach to AI



High Risk



Requirements for high-risk AI (Title III, chapter 2)

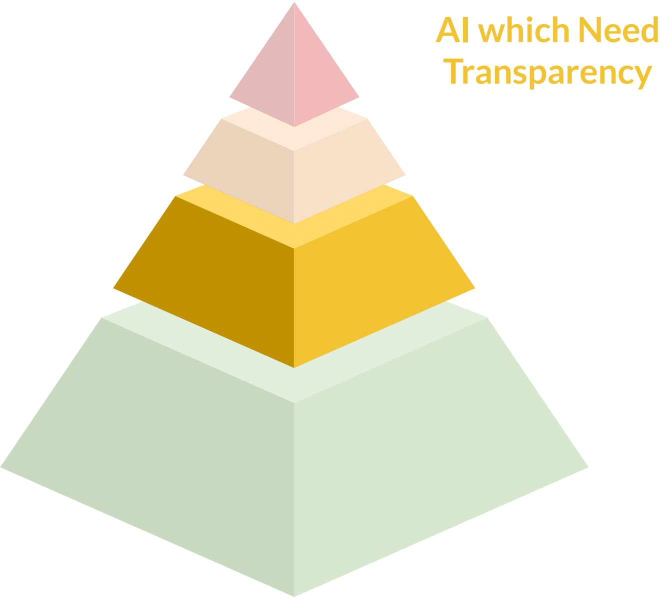
Establish and implement **risk management** processes

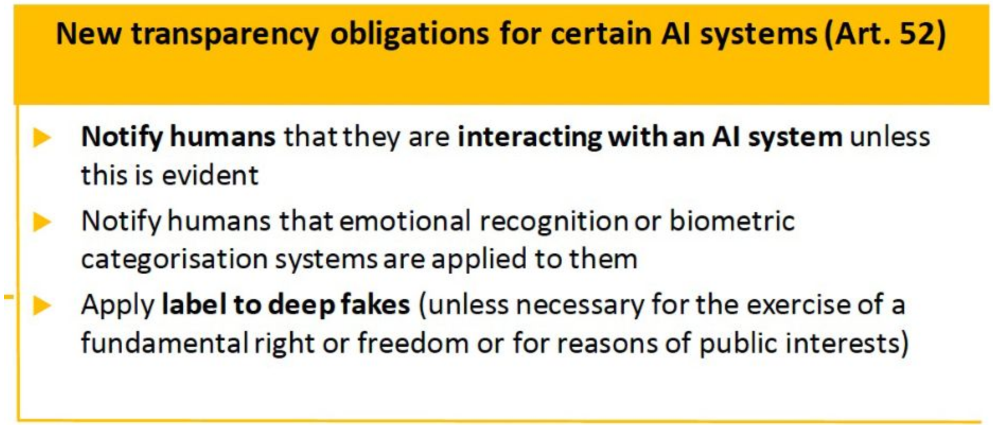
&

In light of the **intended purpose** of the AI system

- Use high-quality **training, validation and testing data** (relevant, representative etc.)
- Establish **documentation** and design logging features (traceability & auditability)
- Ensure appropriate certain degree of **transparency** and provide users with **information** (on how to use the system)
- Ensure **human oversight** (measures built into the system and/or to be implemented by users)
- Ensure **robustness**, **accuracy** and **cybersecurity**

The European Union risk-based approach to AI

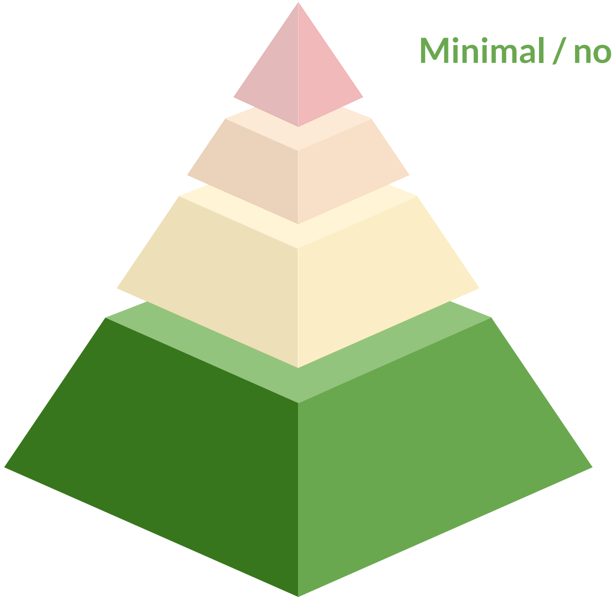




New transparency obligations for certain AI systems (Art. 52)

- **Notify humans** that they are **interacting with an AI system** unless this is evident
- Notify humans that emotional recognition or biometric categorisation systems are applied to them
- Apply **label to deep fakes** (unless necessary for the exercise of a fundamental right or freedom or for reasons of public interests)

The European Union risk-based approach to AI

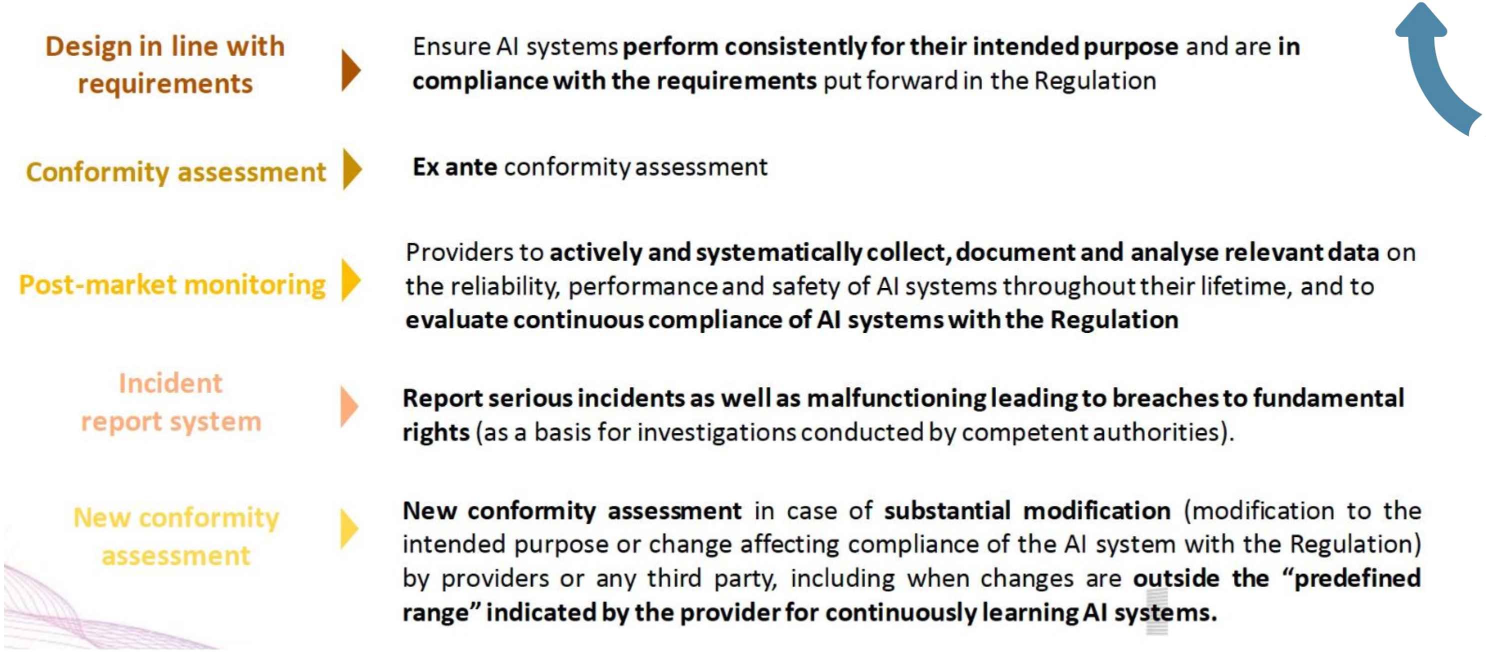


Minimal / no Risk

Possible voluntary codes of conduct for AI with specific transparency requirements (Art. 69)

- No mandatory obligations
- Commission and Board to encourage drawing up of codes of conduct intended to foster the **voluntary application of requirements to low-risk AI systems**

Lifecycle of AI Systems



Design in line with requirements

Ensure AI systems **perform consistently for their intended purpose** and are **in compliance with the requirements** put forward in the Regulation

Conformity assessment

Ex ante conformity assessment

Post-market monitoring

Providers to **actively and systematically collect, document and analyse relevant data** on the reliability, performance and safety of AI systems throughout their lifetime, and to **evaluate continuous compliance of AI systems with the Regulation**

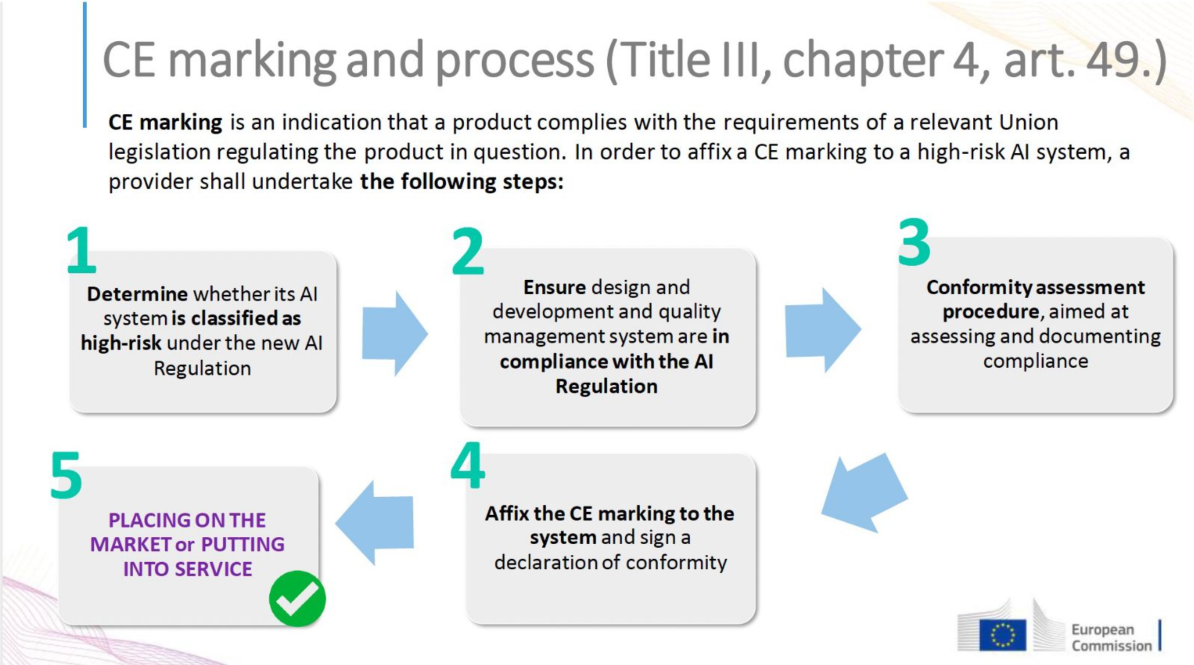
Incident report system

Report serious incidents as well as malfunctioning leading to breaches to fundamental rights (as a basis for investigations conducted by competent authorities).

New conformity assessment

New conformity assessment in case of **substantial modification** (modification to the intended purpose or change affecting compliance of the AI system with the Regulation) by providers or any third party, including when changes are **outside the "predefined range" indicated by the provider for continuously learning AI systems.**

Conformity Assessment of AI



CE marking and process (Title III, chapter 4, art. 49.)

CE marking is an indication that a product complies with the requirements of a relevant Union legislation regulating the product in question. In order to affix a CE marking to a high-risk AI system, a provider shall undertake **the following steps:**

1 **Determine** whether its AI system **is classified as high-risk** under the new AI Regulation

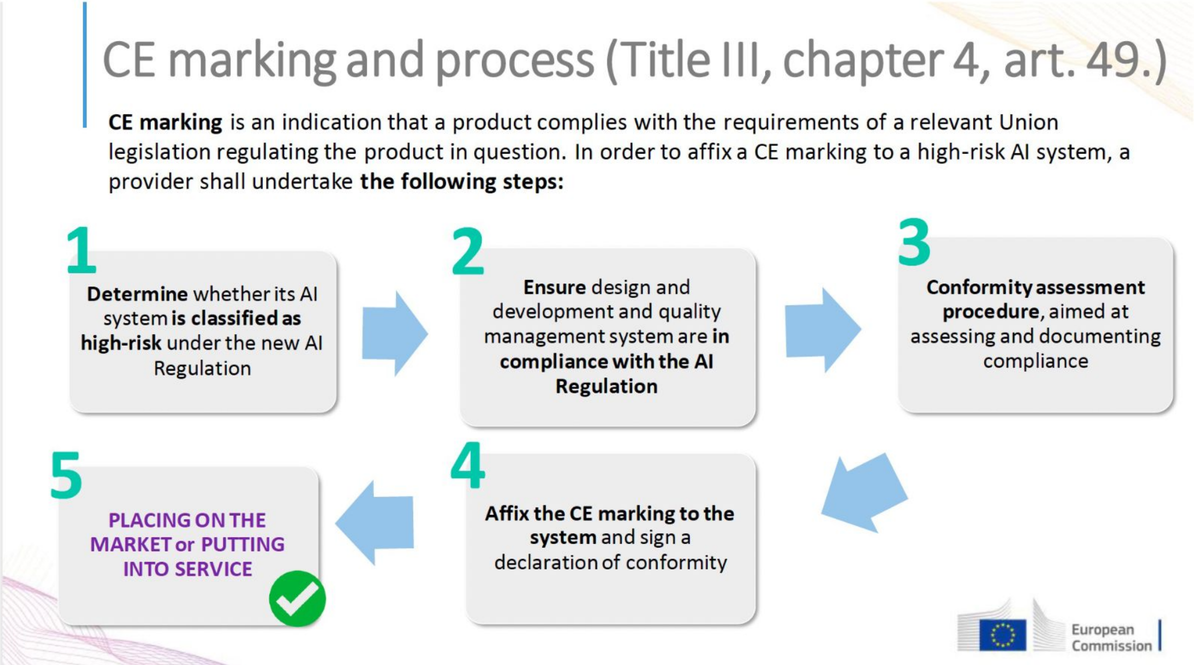
2 **Ensure** design and development and quality management system are **in compliance with the AI Regulation**

3 **Conformity assessment procedure**, aimed at assessing and documenting compliance

4 **Affix the CE marking to the system** and sign a declaration of conformity

5 PLACING ON THE MARKET or PUTTING INTO SERVICE

Conformity Assessment of AI

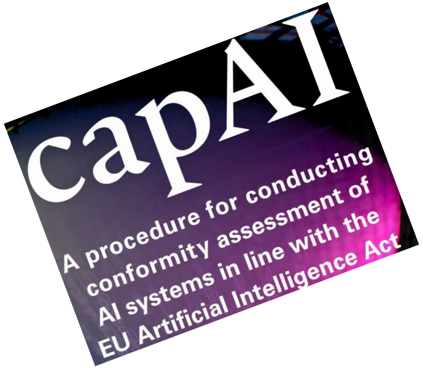


CE marking and process (Title III, chapter 4, art. 49.)

CE marking is an indication that a product complies with the requirements of a relevant Union legislation regulating the product in question. In order to affix a CE marking to a high-risk AI system, a provider shall undertake **the following steps:**

1. **Determine** whether its AI system **is classified as high-risk** under the new AI Regulation
2. **Ensure** design and development and quality management system are **in compliance with the AI Regulation**
3. **Conformity assessment procedure**, aimed at assessing and documenting compliance
4. **Affix the CE marking to the system** and sign a declaration of conformity
5. PLACING ON THE MARKET or PUTTING INTO SERVICE ✓

European Commission



capAI

A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act

That's All!

DLAB
CTE · CAGLIARI