Università di Genova

Pattern Recognition and Applications Lab

University of Cagliari, Italy

Short History of AI: from Foundations to the EU AI Act

Fabio Roli

Machine Learning Security Course, University of Cagliari, 14.12.2023

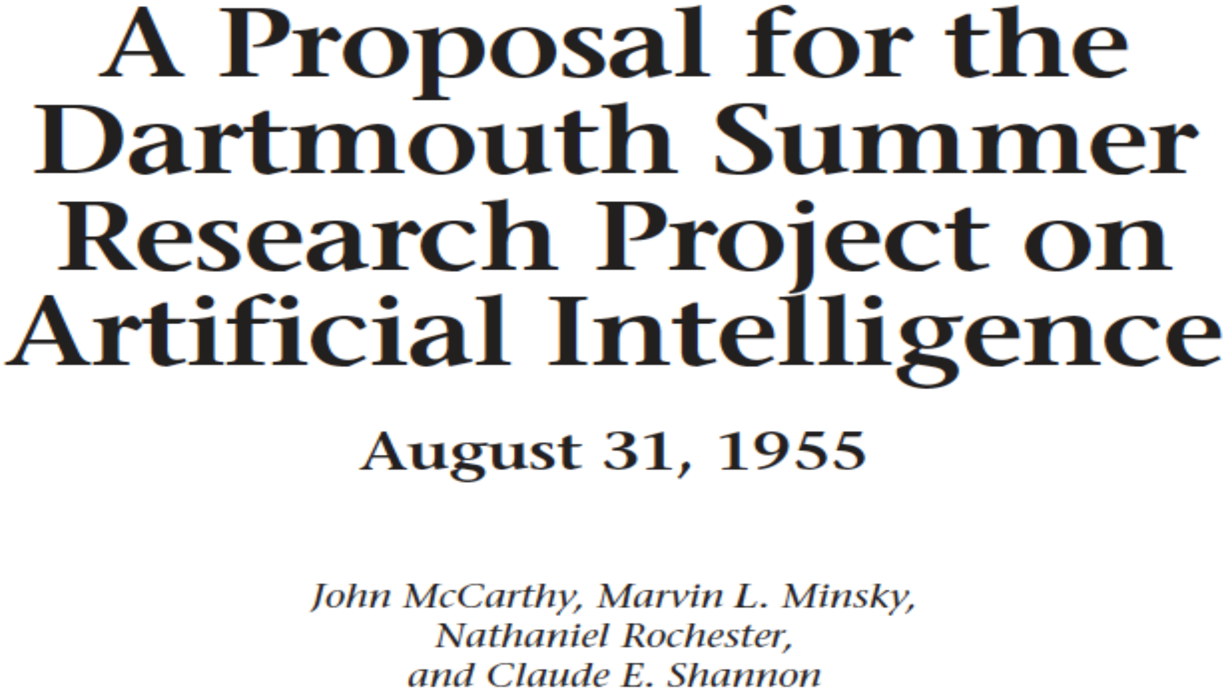
My two take-home messages for today

- Modern AI: which kind of AI is? And how we got here over the last 70 years.

- Why the last December 9th Europe decided to rule AI? And what is the EU AI Act in short

The beginning of the story, summer 1955...

[http://www.aaai.org/ojs/index.php/aimagazine/article/view/1904]



A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

John McCarthy, Marvin L. Minsky,
Nathaniel Rochester,
and Claude E. Shannon

The beginning of the story, summer 1956...

"The study will proceed on the basis of the conjecture that, in principle, ***every aspect*** *of* ***learning*** *or* ***any*** *other* ***feature*** *of* ***intelligence*** *can be described so precisely that a* ***machine can*** *be constructed to* ***simulate it****"*

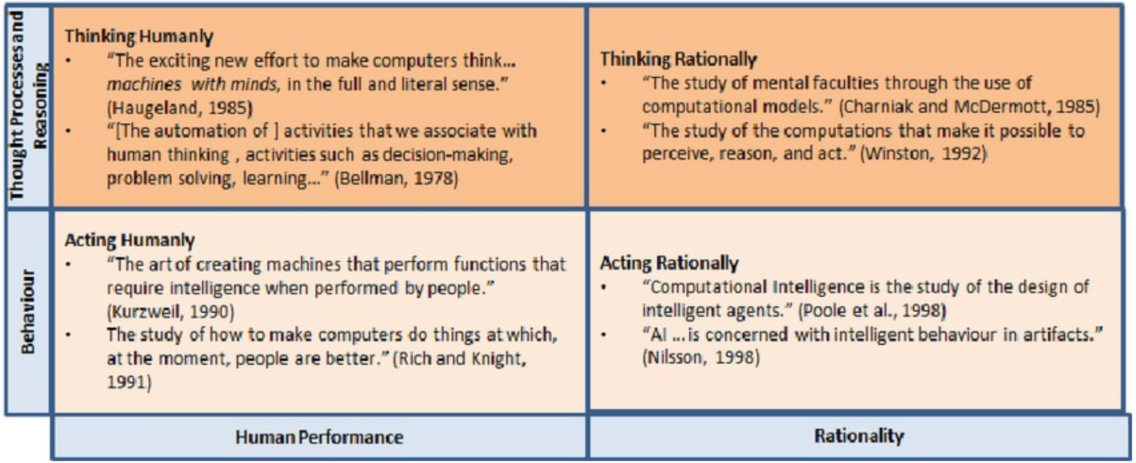
"We propose a ***two-month*** *study done by* ***ten people****...*
...We believe that in one summer, ***significant progress*** *can be made in the development of* ***artificial intelligence****"*

But what does it mean for a machine to be intelligent?

Possible goals of Artificial Intelligence

- There are four objectives historically pursued in the implementation of AI systems (intelligent machines), which are shown in the table below.

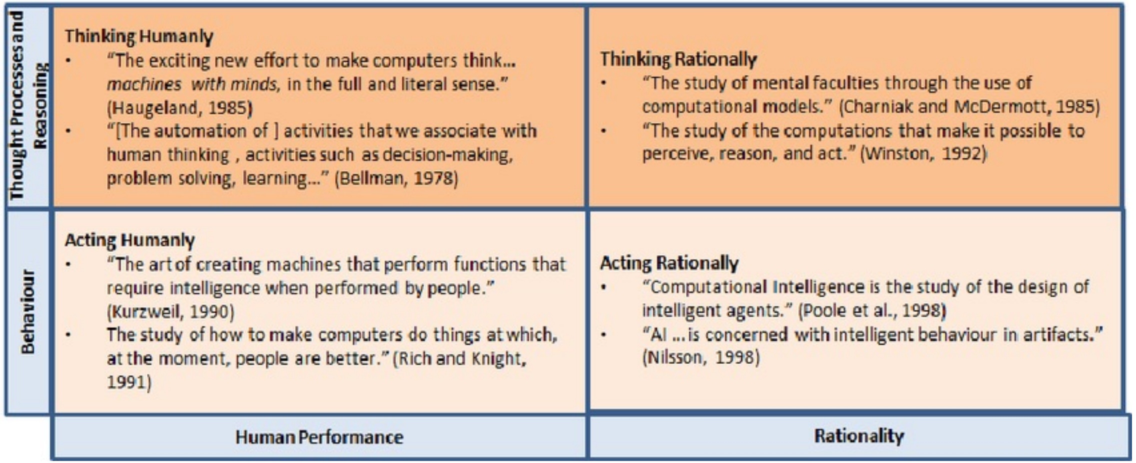
- The rows correspond to the distinction between **thought** and **behavior**, the columns to the distinction between **human performance** and **rationality**.



	Human Performance	Rationality
Thought Processes and Reasoning	**Thinking Humanly** • "The exciting new effort to make computers think... *machines with minds*, in the full and literal sense." (Haugeland, 1985) • "[The automation of] activities that we associate with human thinking , activities such as decision-making, problem solving, learning..." (Bellman, 1978)	**Thinking Rationally** • "The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) • "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Behaviour	**Acting Humanly** • "The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) • The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	**Acting Rationally** • "Computational Intelligence is the study of the design of intelligent agents." (Poole et al., 1998) • "AI ...is concerned with intelligent behaviour in artifacts." (Nilsson, 1998)

S.Russel, P.Norvig Artificial Intelligence: A Modern Approach, 2022

What about the intelligence of ChatGPT?



	Human Performance	Rationality
Thought Processes and Reasoning	**Thinking Humanly** • "The exciting new effort to make computers think… *machines with minds*, in the full and literal sense." (Haugeland, 1985) • "[The automation of] activities that we associate with human thinking , activities such as decision-making, problem solving, learning…" (Bellman, 1978)	**Thinking Rationally** • "The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) • "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Behaviour	**Acting Humanly** • "The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) • The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	**Acting Rationally** • "Computational Intelligence is the study of the design of intelligent agents." (Poole et al., 1998) • "AI …is concerned with intelligent behaviour in artifacts." (Nilsson, 1998)

Turing test

A. Turing, “Computing Machinery & Intelligence,” *Mind*, Vol. 59(236), 1950.

How to decide if a machine is intelligent?

Suppose you put in a room a human and a machine that claims being intelligent

Another human being, the "judge," can communicate with them in written and spoken form, but without seeing them

The judge asks a series of questions ("challenges") to the two interlocutors and then decides who is the human

The judge's error is the proof of the intelligence of the machine

The judge could challenge the interlocutors to read a handwritten text....or to quickly recognize faces....or to solve crossword puzzles...

The most recent definition of AI system (09/12/2023)…

The EU AI Act should regulate **AI systems** (the high risk ones)

But what is an AI system?

Artificial intelligence system (AI system) *means a system that is designed to operate with a certain level of* ***autonomy*** *and that, based on machine and/or human-provided data and inputs,* ***infers*** *how to achieve a given set of* ***human-defined objectives*** *using machine learning and/or logic- and knowledge based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions , influencing the environments with which the AI system interacts.*

But there is an open debate on the best definition for «AI system»…

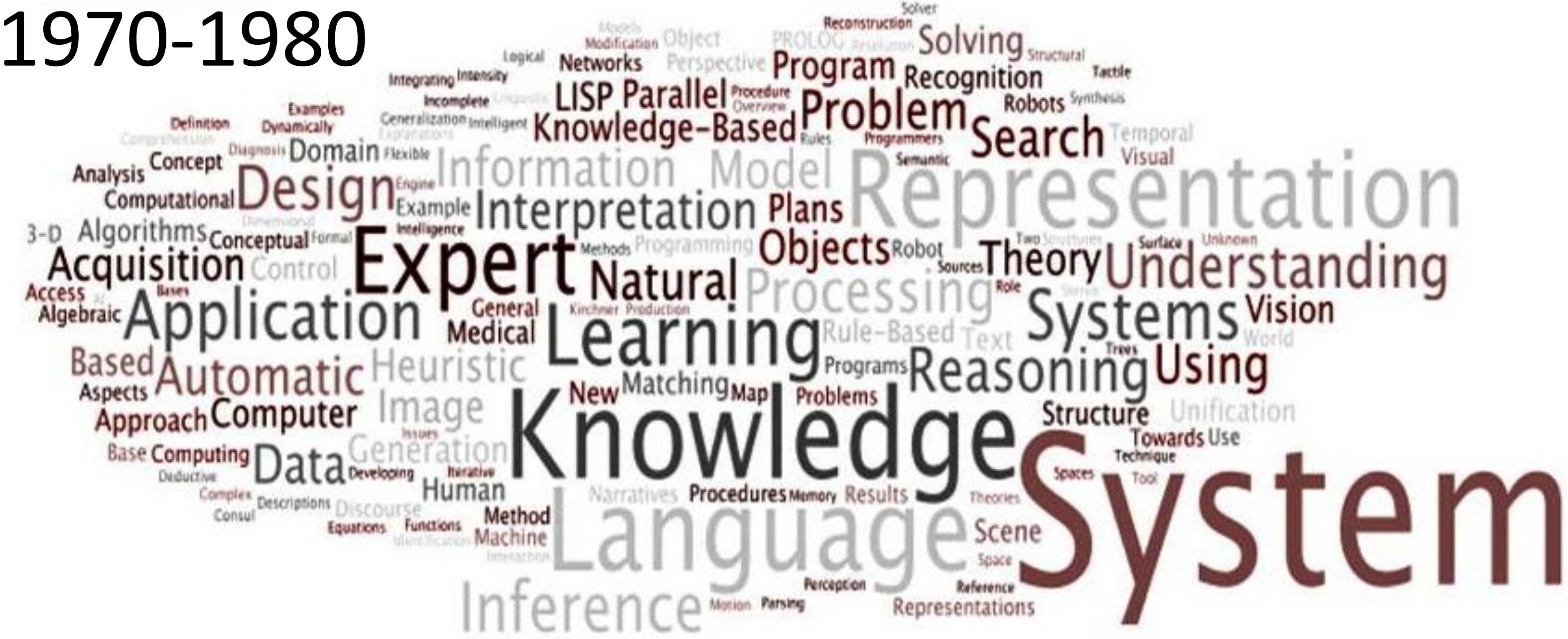
Here we are, but how we got here?

A short, and biased, history of AI…

What was AI mainstream in 1970-1980?

AI mainstream in 1970-1980

1970-1980



N. Cristianini, *On the current paradigm in artificial intelligence* (2014)

1980: The INDUCE algorithm by Ryszard S. Michalski

- A learning algorithm from examples that generates **symbolic descriptions** of object classes
  - The description language of object classes is a simple extension of the first-order predicate calculus

1. Learning starts from one example of a class (the «seed»)
2. Examples of other classes are «counter-examples»
3. Guided search for symbolic descriptions that «cover» all the examples of the given class and do not cover the counter-examples

1937 -2007

Ryszard S. Michalski, IEEE-T PAMI, 1980

Learning of object classes with INDUCE (1980)…

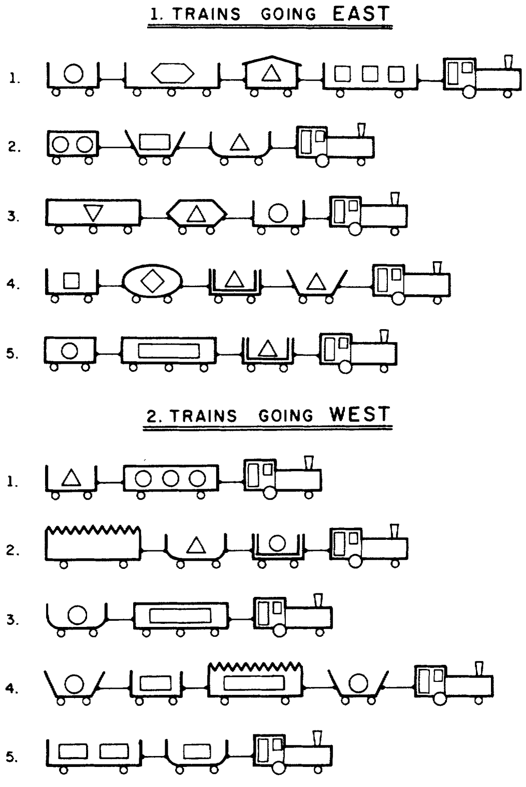
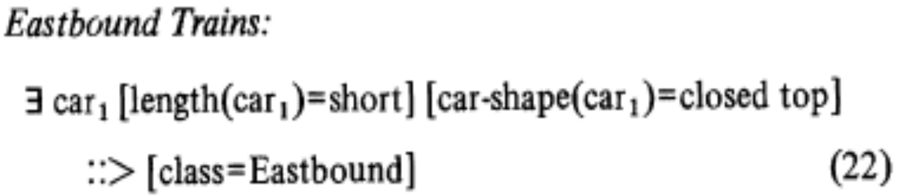


Fig. 4.



Eastbound Trains:

$\exists$ car$_1$ [length(car$_1$)=short] [car-shape(car$_1$)=closed top]

::> [class=Eastbound] (22)

IF a train contains a car that is short and has a closed top, THEN it is an Eastbound train

Machine learning of “known knowns”

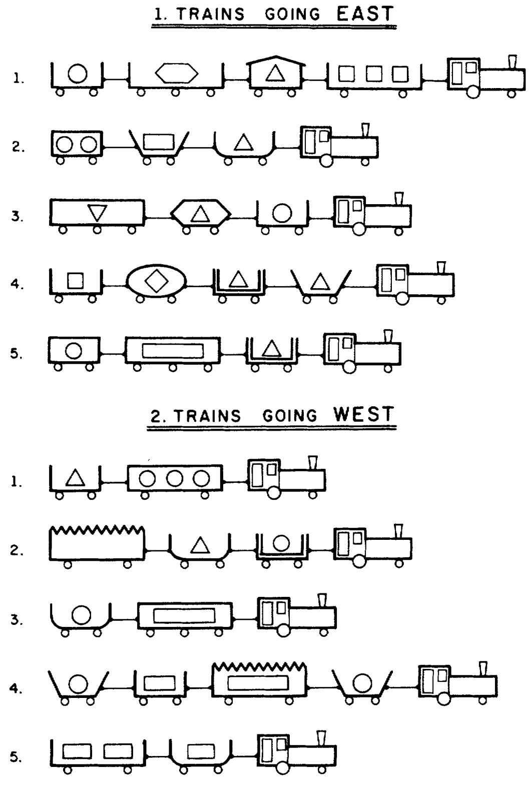
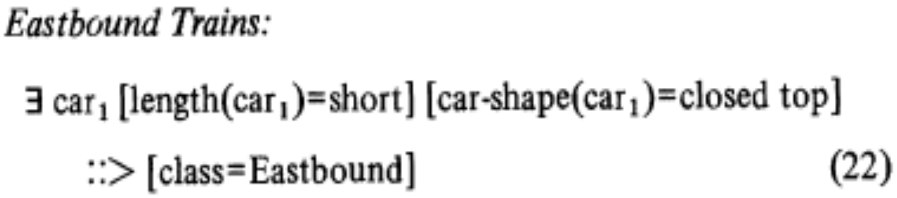


Fig. 4.



Eastbound Trains:

$$\exists\, car_1\, [length(car_1)=short]\, [car\text{-}shape(car_1)=closed\ top]$$

$$::> [class=Eastbound] \qquad (22)$$

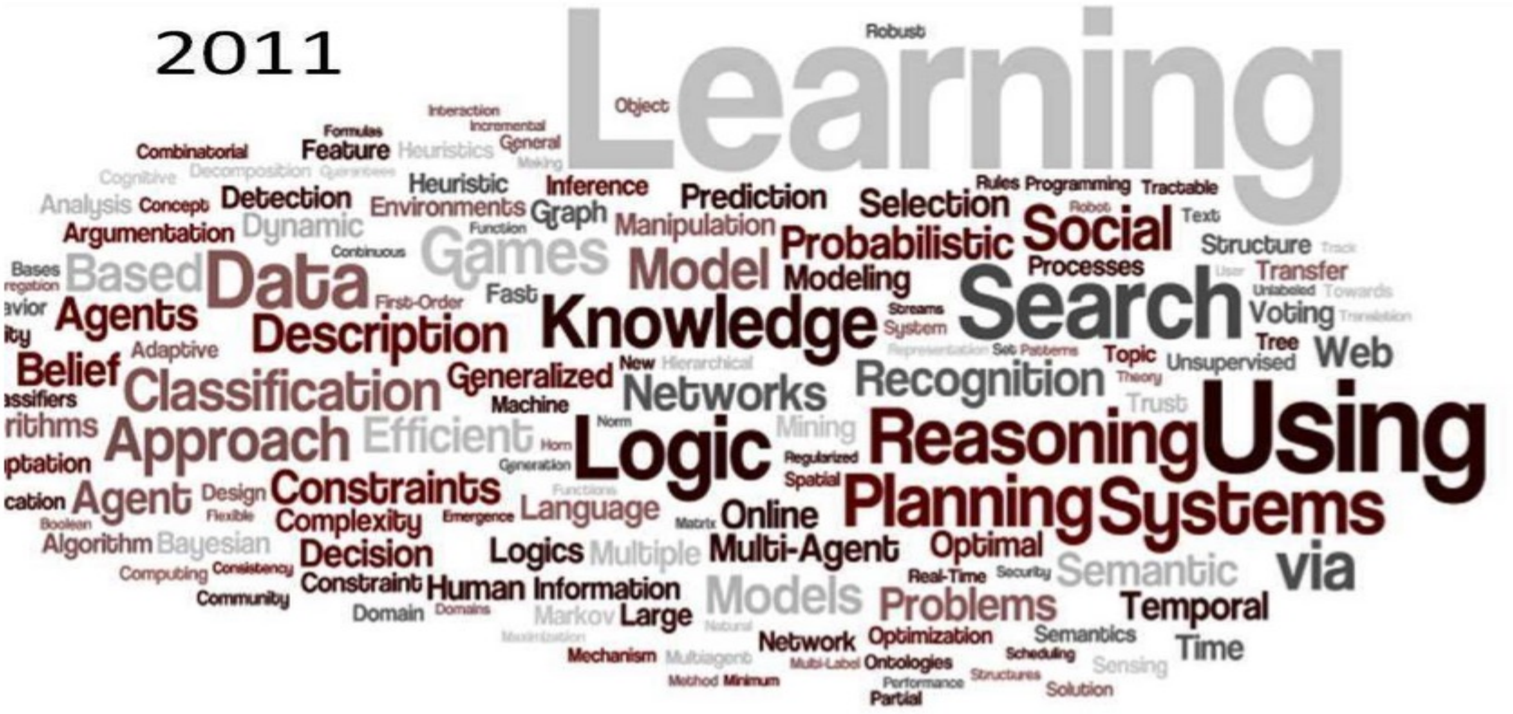
IF a train contains a car that is short and has a closed top, THEN it is an Eastbound train

- «micro worlds» that were perfectly known and predictable («noise-free»)
- The INDUCE algorithm dealt with **«known knowns»**

Credits: slide partially inspired from a speech by Donald Rumsfeld and a research project by Thomas Dietterich
https://futureoflife.org/ai-researcher-thomas-dietterich/

What is AI mainstream nowadays?

What is AI mainstream nowadays?



N. Cristianini, *On the current paradigm in artificial intelligence* (2014)

XD: eXtreme Data-driven Learning

Here we are

After 70 years of research in AI, the main stream is **Big Data + Deep Learning + GPUs**

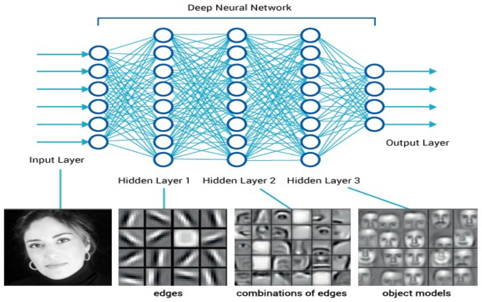
Big Data

Facebook millions of images per day

Walmart, Petabytes customer data hourly

YouTube thousands of hours of videos per minute

Deep Learning



GPU

Machine learning of “known unknowns”

- **Underlying assumption:** past data is *representative* of future data (IID data)

- **Underlying assumption:** we know all the object classes to recognize

- The success of modern AI is on tasks for which we collected enough representative training data for the object classes to recognize

The current mainstream of "known unknowns"

Big Data + GPU + Deep Learning

How we got here ?

Machine learning as an experimental science…

«Machine learning is a scientific discipline and, like the fields of AI and computer science, has both theoretical and empirical aspects.

[…]

Although experimental studies are not the only path to understanding, we feel they constitute one of machine learning's brightest hopes for rapid **scientific progress**, and we encourage other researchers to join in this evolution.»

Pat Langley

Machine Learning as an Experimental Science

(1988)

The rise of benchmark data sets

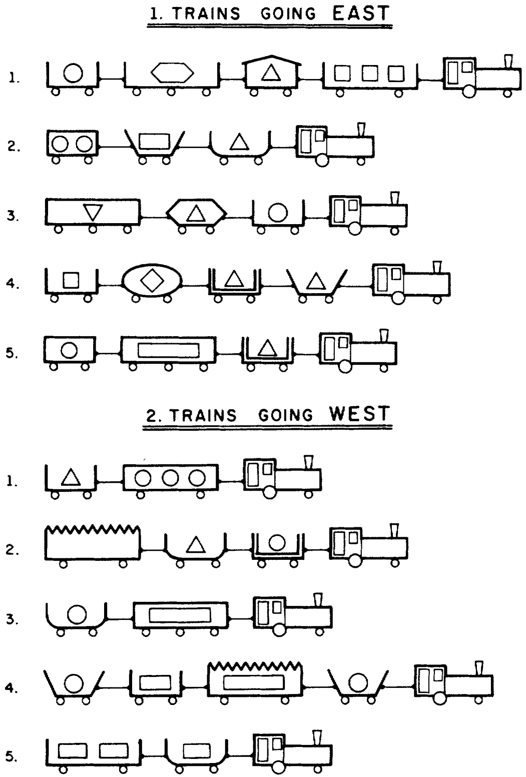
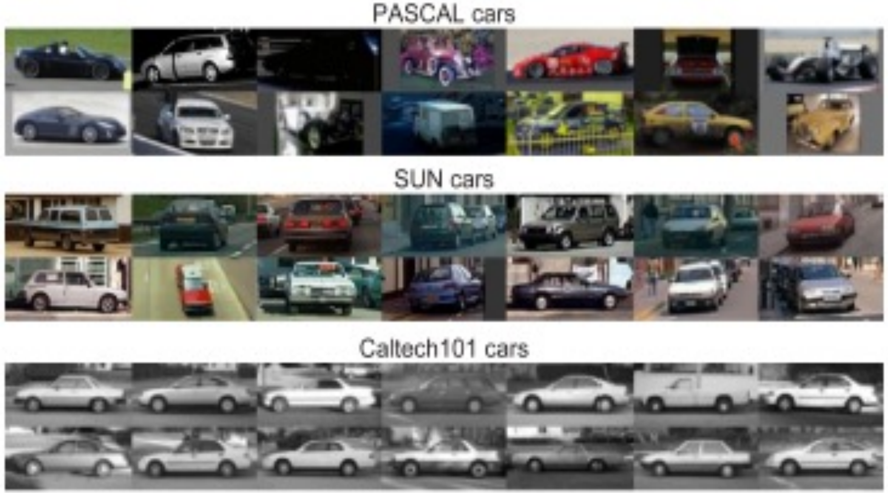
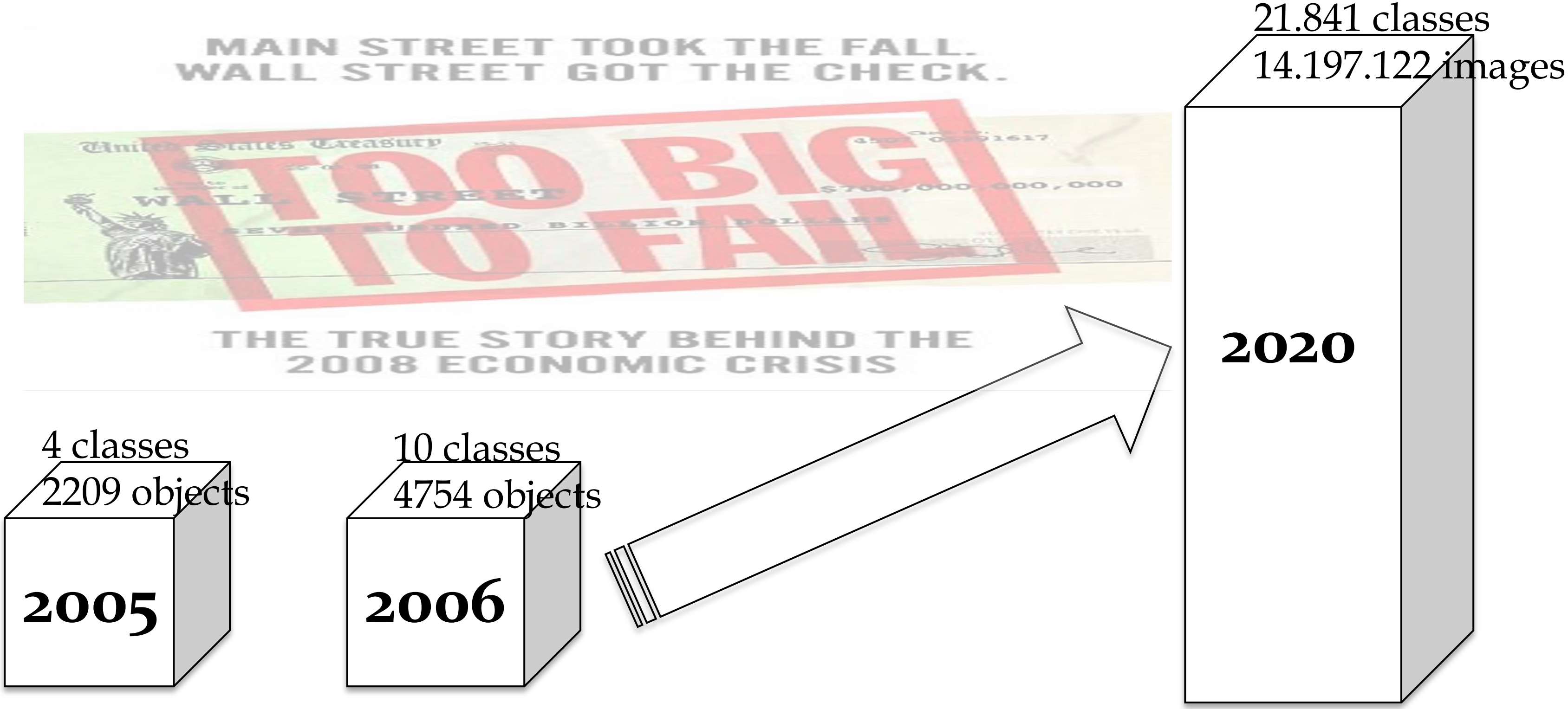


Fig. 4.



Bigger is better…


MAIN STREET TOOK THE FALL.
WALL STREET GOT THE CHECK.
United States Treasury
WALL STREET
SEVEN HUNDRED BILLION DOLLARS
$700,000,000,000
TOO BIG TO FAIL
THE TRUE STORY BEHIND THE 2008 ECONOMIC CRISIS

4 classes
2209 objects

2005

10 classes
4754 objects

2006

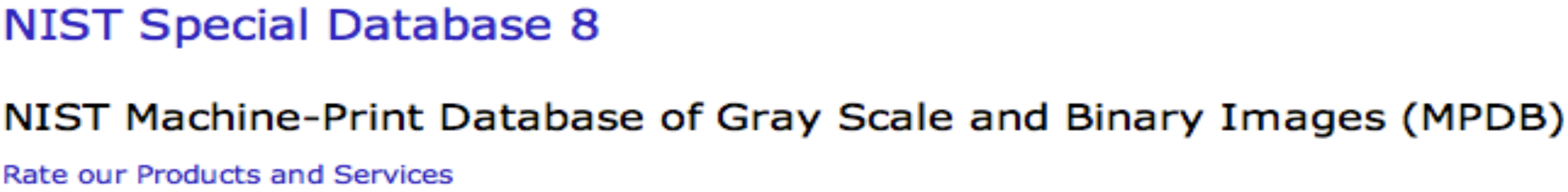
21.841 classes
14.197.122 images

2020

The PASCAL VOC data set

image-net.org

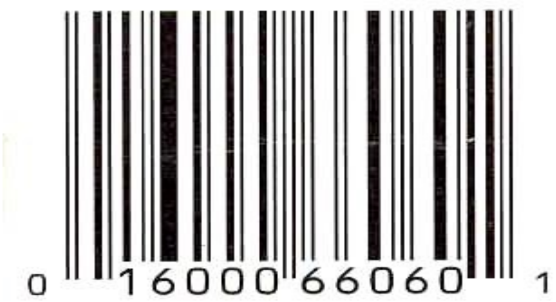
Splendors of benchmark data sets

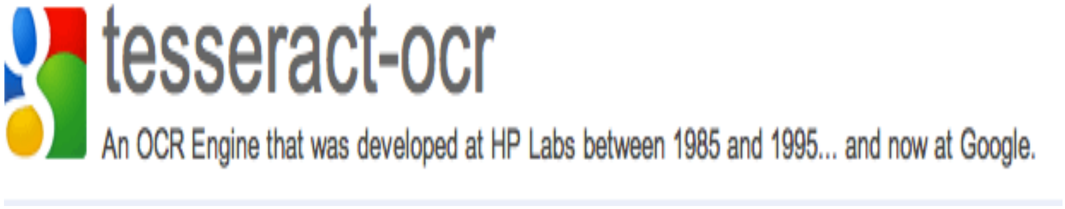


NIST Special Database 8

NIST Machine-Print Database of Gray Scale and Binary Images (MPDB)

Rate our Products and Services



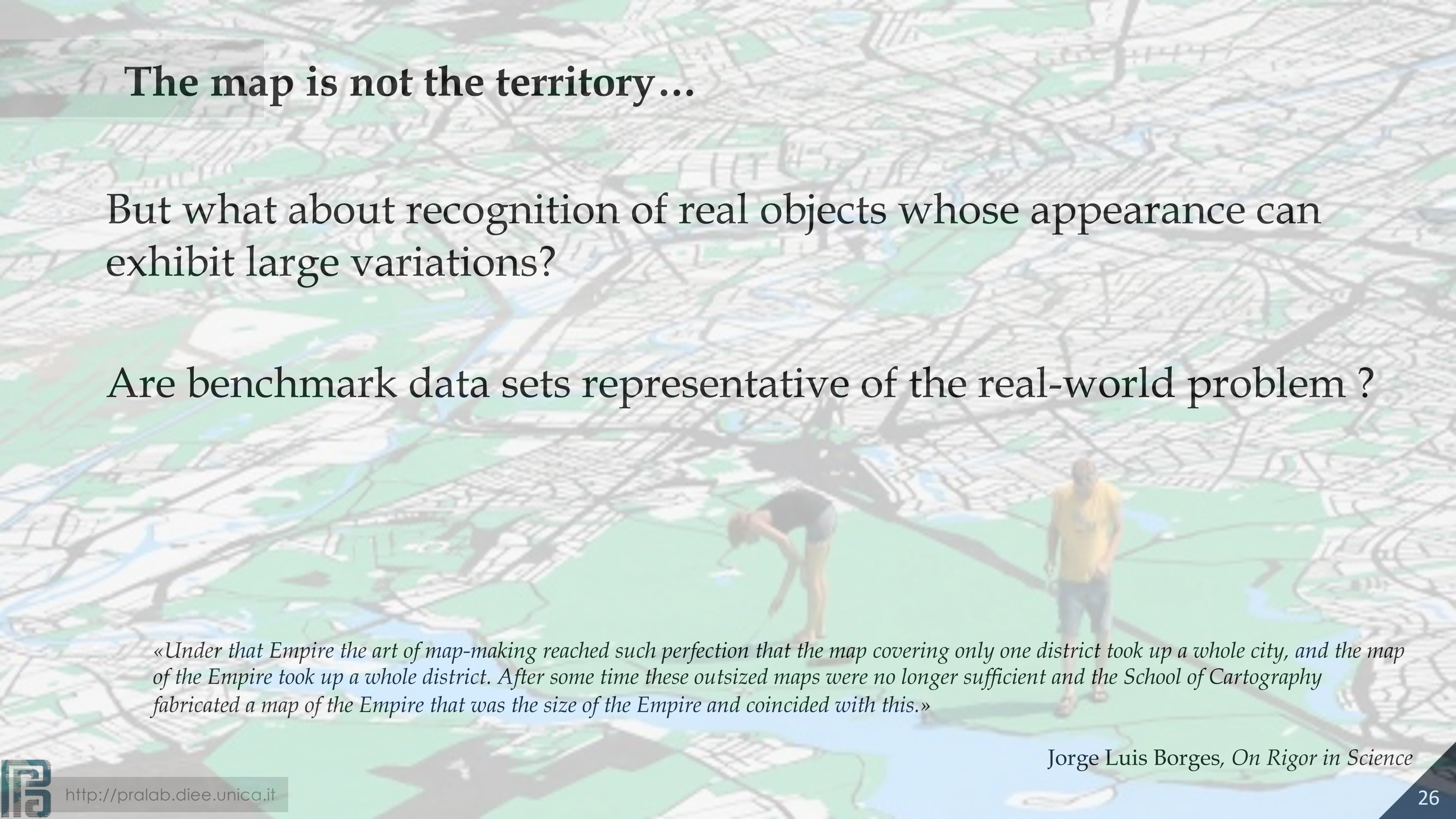


tesseract-ocr

An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google.

Theo Pavlidis, *The challenge of general machine vision*, 2014

The map is not the territory…



But what about recognition of real objects whose appearance can exhibit large variations?

Are benchmark data sets representative of the real-world problem ?

> *«Under that Empire the art of map-making reached such perfection that the map covering only one district took up a whole city, and the map of the Empire took up a whole district. After some time these outsized maps were no longer sufficient and the School of Cartography fabricated a map of the Empire that was the size of the Empire and coincided with this.»*

Jorge Luis Borges, *On Rigor in Science*

Miseries of benchmark data sets

…So, what is the value of current data sets when used to train algorithms for object recognition that will be deployed in the real world? …

Antonio Torralba, Alexei A. Efros, CVPR 2011

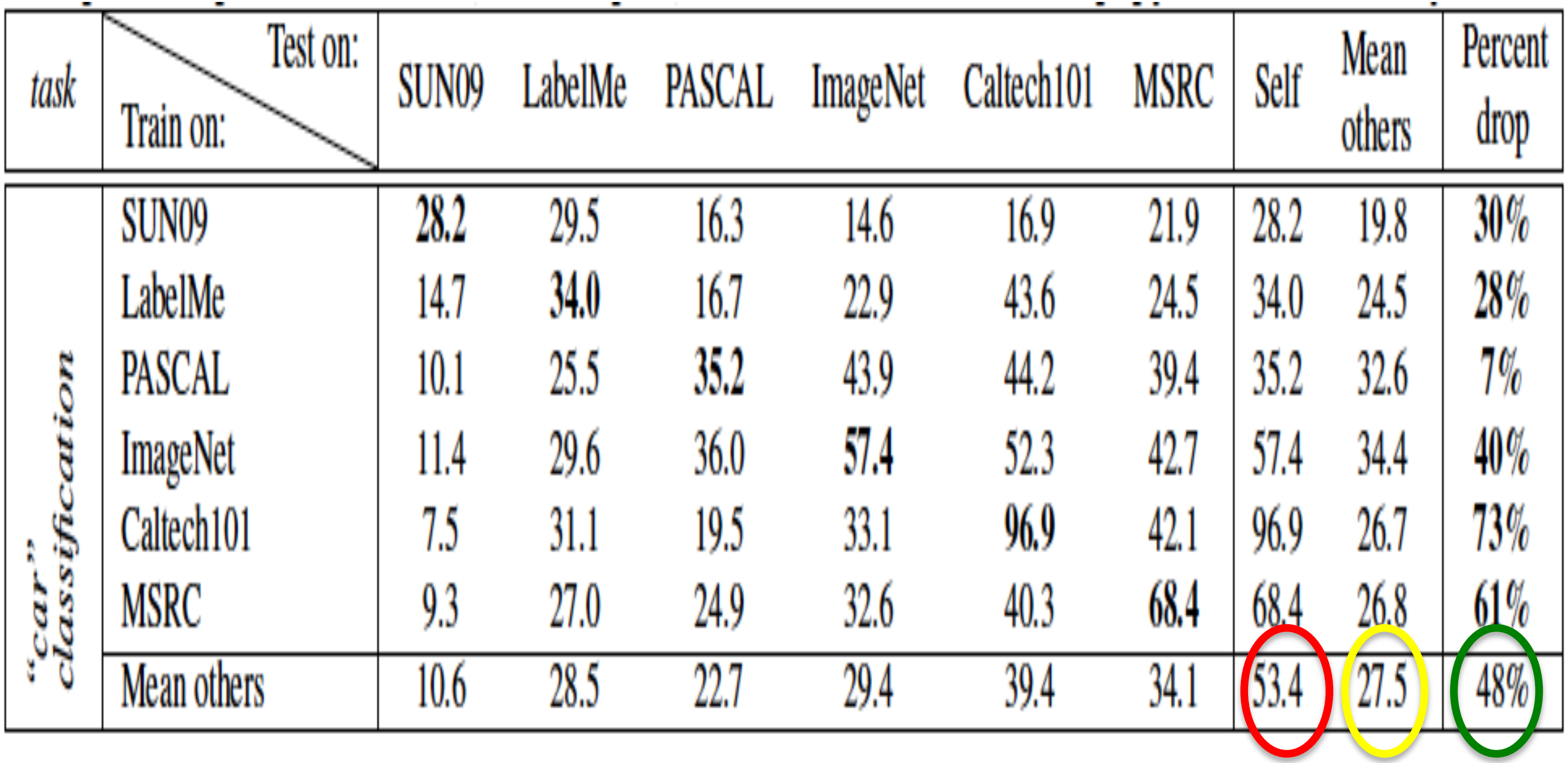
Miseries of benchmark data sets

…So, what is the value of current data sets when used to train algorithms for object recognition that will be deployed in the real world? …

…The answer that emerges can be summarized as: "better than nothing, but not by much"…

Antonio Torralba, Alexei A. Efros, CVPR 2011

The curse of biased data sets



task	Test on: / Train on:	SUN09	LabelMe	PASCAL	ImageNet	Caltech101	MSRC	Self	Mean others	Percent drop
“car” classification	SUN09	**28.2**	29.5	16.3	14.6	16.9	21.9	28.2	19.8	30%
	LabelMe	14.7	**34.0**	16.7	22.9	43.6	24.5	34.0	24.5	28%
	PASCAL	10.1	25.5	**35.2**	43.9	44.2	39.4	35.2	32.6	7%
	ImageNet	11.4	29.6	36.0	**57.4**	52.3	42.7	57.4	34.4	40%
	Caltech101	7.5	31.1	19.5	33.1	**96.9**	42.1	96.9	26.7	73%
	MSRC	9.3	27.0	24.9	32.6	40.3	**68.4**	68.4	26.8	61%
	Mean others	10.6	28.5	22.7	29.4	39.4	34.1	53.4	27.5	48%

Torralba, Efros (CVPR 2011)

Bigger is better?

Maybe current data sets are not large enough to represent well problems of the real world?

Should we make them bigger?

Can we sample real world images?

Estimate No. 1: The number of meaningful/valid images on a 1200 by 1200 display is at least as high as 10^{400}.

Estimate No. 2: 10^{25} (greater than a trillion squared) is a very conservative lower bound to the number of all possible discernible images.

«These numbers suggest that it is impractical to construct training or testing sets of images that are dense in the set of all images unless the class of images is restricted.»

Theo Pavlidis

The Number of All Possible Meaningful or Discernible Pictures (2009)

Data beats theory

«By the mid-2000s, with success stories piling up, the field had learned a powerful lesson: **data can be stronger than theoretical models**.
A new generation of intelligent machines had emerged, powered by a small set of statistical learning algorithms and large amounts of data.»

Nello Cristianini
The road to artificial intelligence: A case of data over theory
(New Scientist, 2016)

The unreasonable effectiveness of data

«Perhaps when it comes to ***natural language*** *processing and related fields, we're doomed to complex theories that will never have the elegance of physics equations. But if that's so, we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data»*

Alon Halevy, Peter Norvig, and Fernando Pereira
The unreasonable effetiveness of data
IEEE Intelligent Systems 2009

A look back...

Nello Cristianini
The road to artificial intelligence: A case of data over theory
(New Scientist, 2016)

The "fathers" of artificial intelligence in the 1950s thought that a machine could reproduce human intelligence with search algorithms and symbolic reasoning.

Today we know that they were wrong!

The "fathers" of artificial intelligence in the 1950s thought that building intelligent machines would give us a better understanding of how the human brain works.

Today we know that they were wrong!

Today, the behavior of our computers may appear intelligent. But in fact, these are statistical algorithms that discover patterns, correlations. Without understanding the causes... It is, however, intelligence, in the behavioral sense of Turing! It is, however, a great scientific and technological advance!

Which way we have taken towards modern AI?

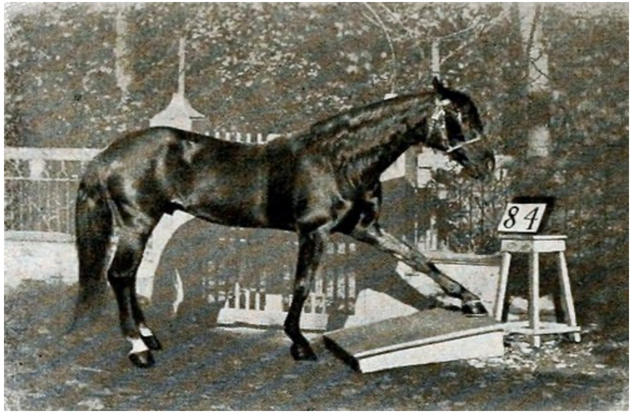
- Why we left the road of the fathers of AI?

- Because it didn't work…

- But which way we have taken?

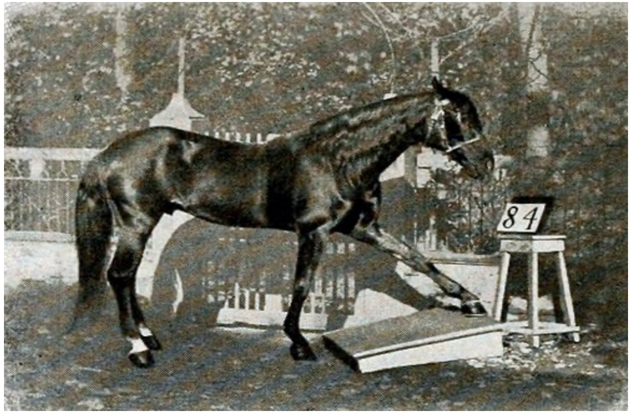
The shortcut…

Clever Hans…



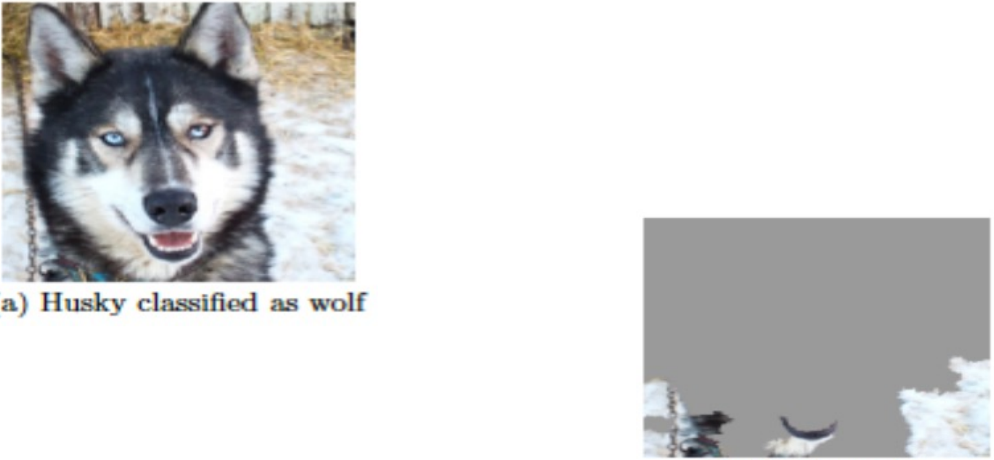
In the early 1900s, Wilhelm von Osten, a German horse trainer and mathematician, spread the news that his horse had learned to count...

Clever Hans…



Correlation is not causation !
Being good at discovering statistical correlations is not the same as knowing math...

Shortcut learning and spurious correlations…



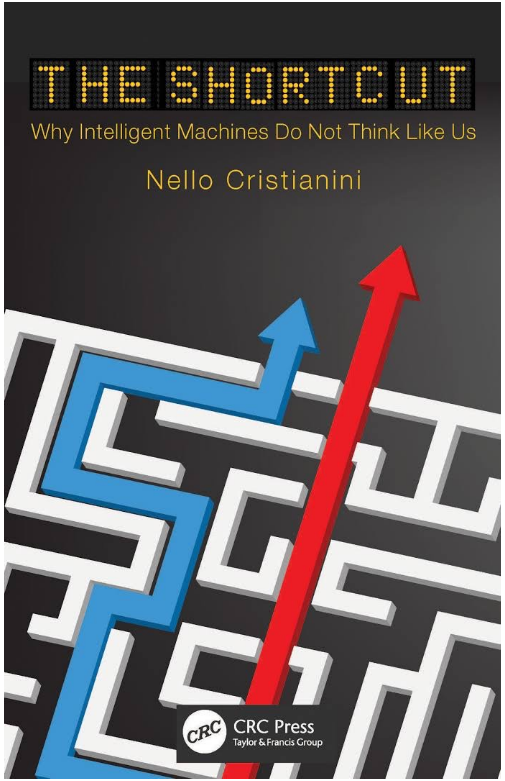
(a) Husky classified as wolf

(b) Explanation

[M.T. Ribeiro et al., KDD 2016]

Recommended reading

The shortcut (N. Cristianini, 2023): how machines became intelligent without thinking in a human way



Today AI (ChatGPT…) may appear intelligent…

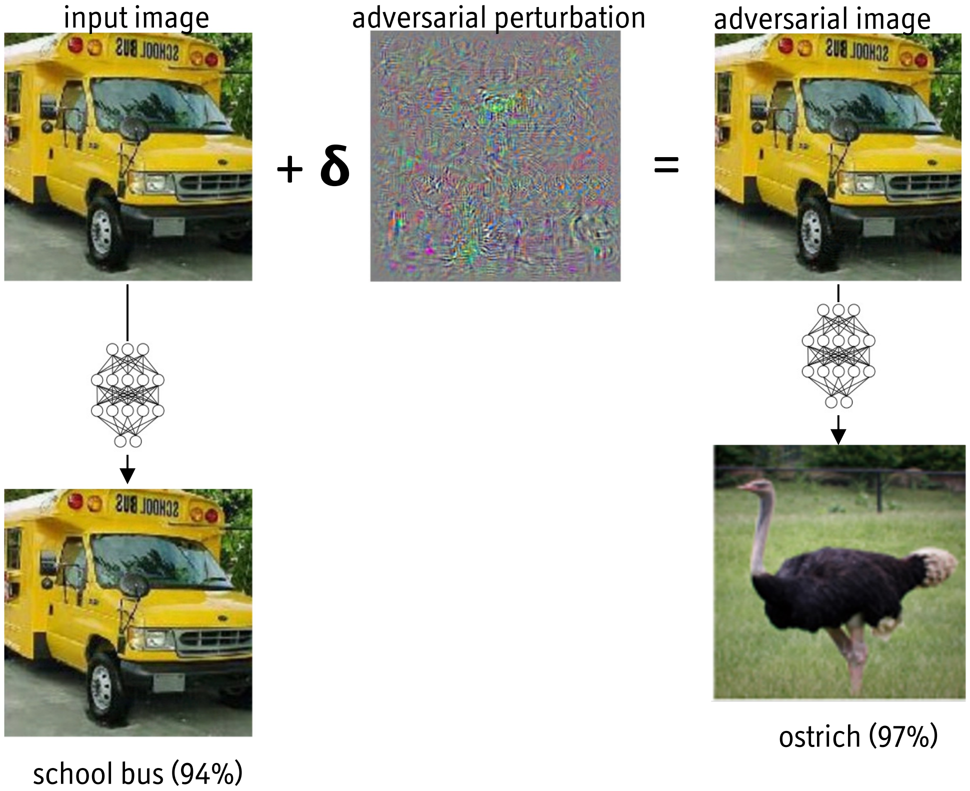
- But we know that we exploited a shortcut to get here…
- High accuracy does not always imply robustness, and correlation does not imply causation…
- This means that AI can exhbit a lack of robustness and several risks…

Adversarial examples: the workhorse of this course…

Minimize $\|\boldsymbol{\delta}\|$

so that $f(x+\boldsymbol{\delta})=l$

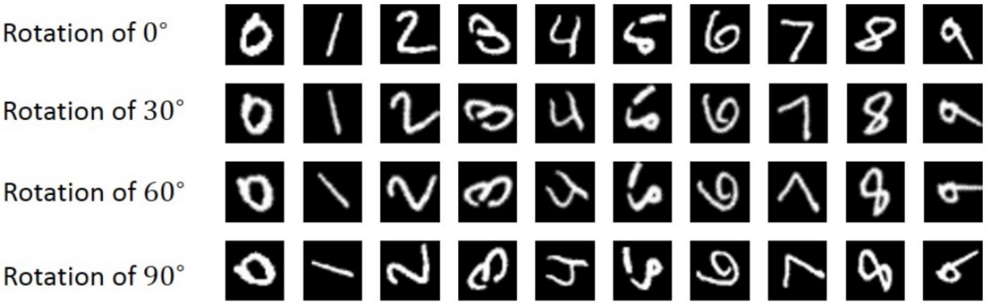
The adversarial image $x+\boldsymbol{\delta}$ is visually hard to distinguish from x
Informally speaking, the solution $x+\boldsymbol{\delta}$ is the closest image to x classified as l=*ostrich*

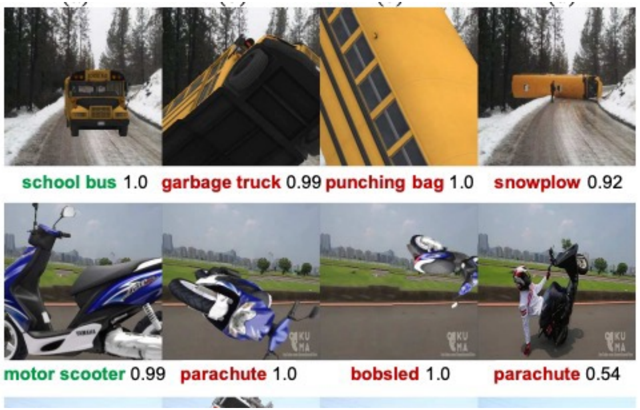


Biggio, Roli et al., Evasion attacks against machine learning at test time, **ECML-PKDD 2013**
Szegedy et al., Intriguing properties of neural networks, **ICLR 2014**

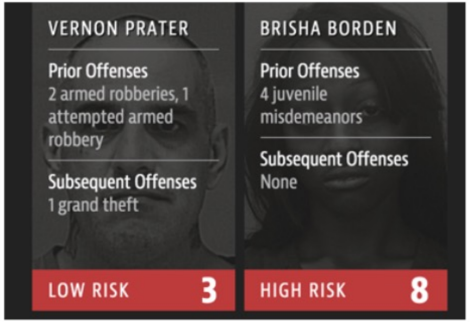
Other risks of AI…

Lack of Robustness…





Bias, Discrimination, Fairness….



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

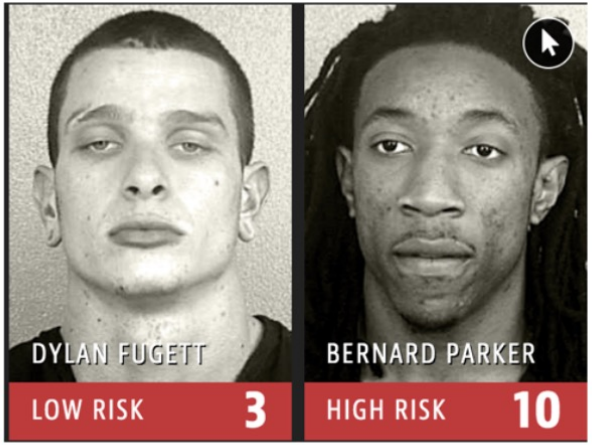
LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8



DYLAN FUGETT

LOW RISK 3

BERNARD PARKER

HIGH RISK 10

Given this state of affairs…

With the aim of **preserving** the **benefits** of AI-based technologies and **reducing** the **risks** for European citizens, the European Parliament voted in plenary the proposal of the European Artificial Intelligence Act (the **EU AI Act**) in June 2023

This regulatory framework on AI is expected to be in force by law at the **end of 2025**.

The EU AI Act identifies **high-risk AI systems** that can gain access to the EU market only after a **"conformity assessment"** (a sort of CE mark for AI) that validates and tests the security and trustworthiness of the AI system.

The EU AI Act for Trustworthy AI

AI is good …

- For citizens
- For business
- For the public interest

… but creates some risks

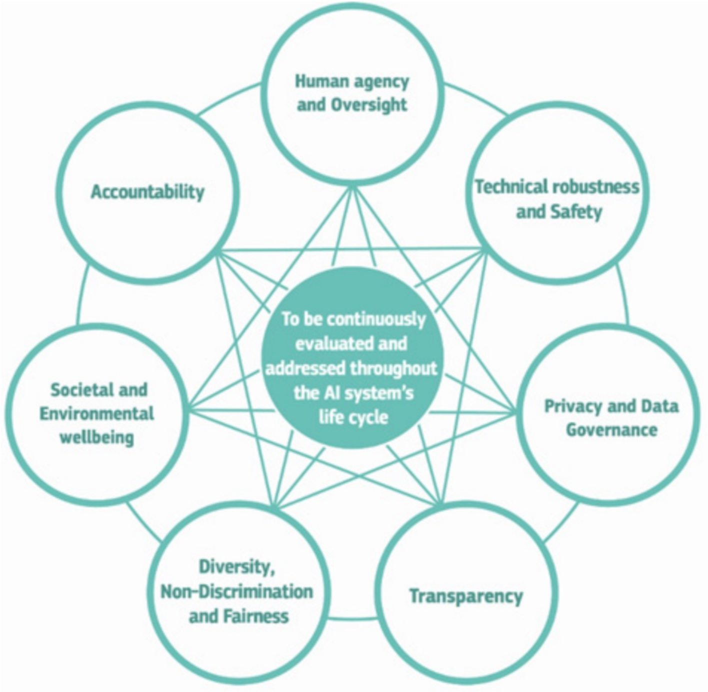
- For the safety of consumers and users
- For fundamental rights

European Commission

[Lucilla Sioli, A European Strategy for Artificial Intelligence, 2021]

The 7 European key requirements for Trustworthy AI

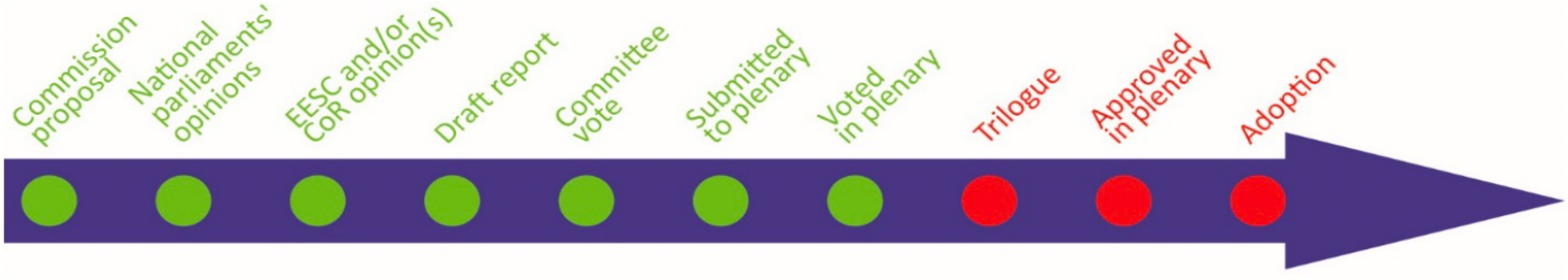
1. **Human agency and oversight**, including fundamental rights, human agency and human oversight
2. **Technical robustness and safety**, including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3. **Privacy and data governance**, including respect for privacy, quality and integrity of data, and access to data
4. **Transparency**, including traceability, explainability and communication
5. **Diversity, non-discrimination and fairness**, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
6. **Societal and environmental wellbeing**, including sustainability and environmental friendliness, social impact, society and democracy
7. **Accountability**, including auditability, minimisation and reporting of negative impact, trade-offs and redress


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

[https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html]

The European AI Act: work in progress…

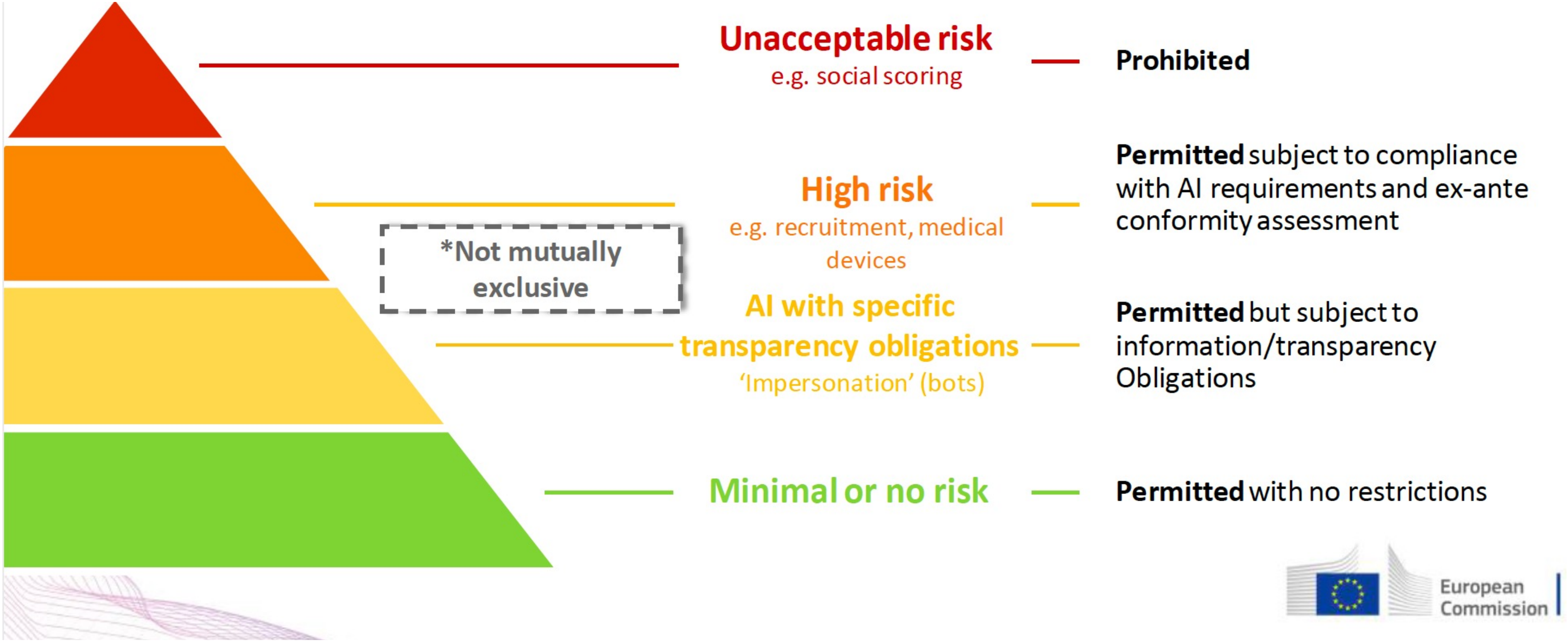
- The European Parliament **voted in plenary** the proposal of the EU AI Act in **June 2023.**



Trilogue: negotiation bringing together representatives of the European Parliament, the Council of the European Union and the European Commission.

Last December 9, the the European Parliament, the Council of the European Union and the European Commission approved a new, and final, version of the EU AI Act. More details later…

The EU risk-based approach to AI regulation



Unacceptable risk
e.g. social scoring
— **Prohibited**

High risk
e.g. recruitment, medical devices
— **Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

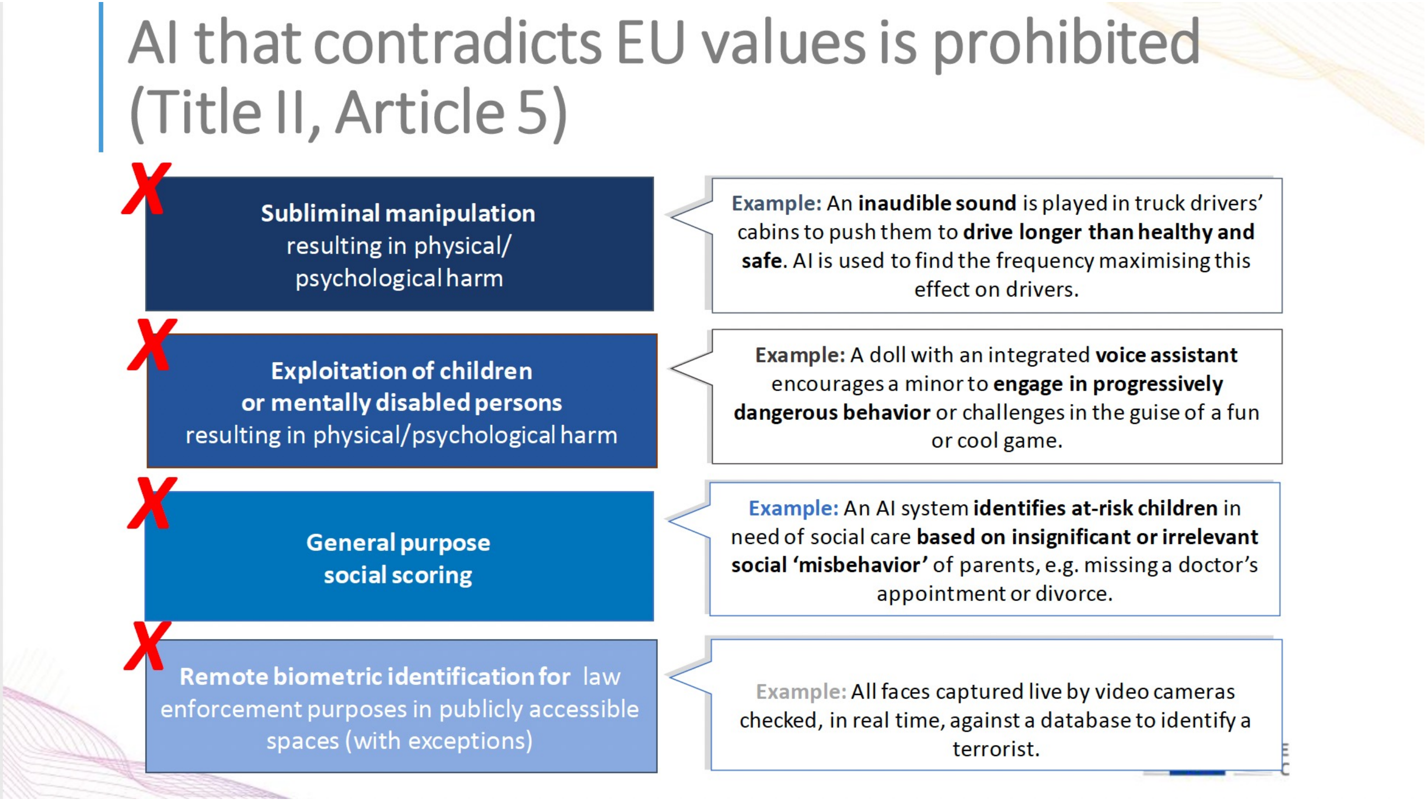
*Not mutually exclusive

AI with specific transparency obligations
'Impersonation' (bots)
— **Permitted** but subject to information/transparency Obligations

Minimal or no risk
— **Permitted** with no restrictions

[Lucilla Sioli, A European Strategy for Artificial Intelligence, 2021]

Prohibited applications of AI



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.

[Lucilla Sioli, A European Strategy for Artificial Intelligence, 2021]

High-risk AI systems

- Biometric identification and categorisation of natural persons, to the extent these do not fall under the aforementioned prohibited practices.
- Management and operation of critical infrastructures, such as AI systems used in safety-relevant components of the management of utilities and traffic.
- Education and vocational training, such as AI systems used to assess students in educational settings, or assign people to training offerings.
- Employment and worker management, such as AI systems used for the recruitment or assessment of employees, including questions such as promotion, performance management and termination.
- Access to essential services, such as AI systems that govern the access to private and public sector services and related actions, including the assessment of creditworthiness, credit scoring, or establishing the order of priority of access to such services. (Note: this aspect applies particularly to AI systems used in the financial services sector).

CapAI, Luciano Floridi et al. 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

High-risk AI systems

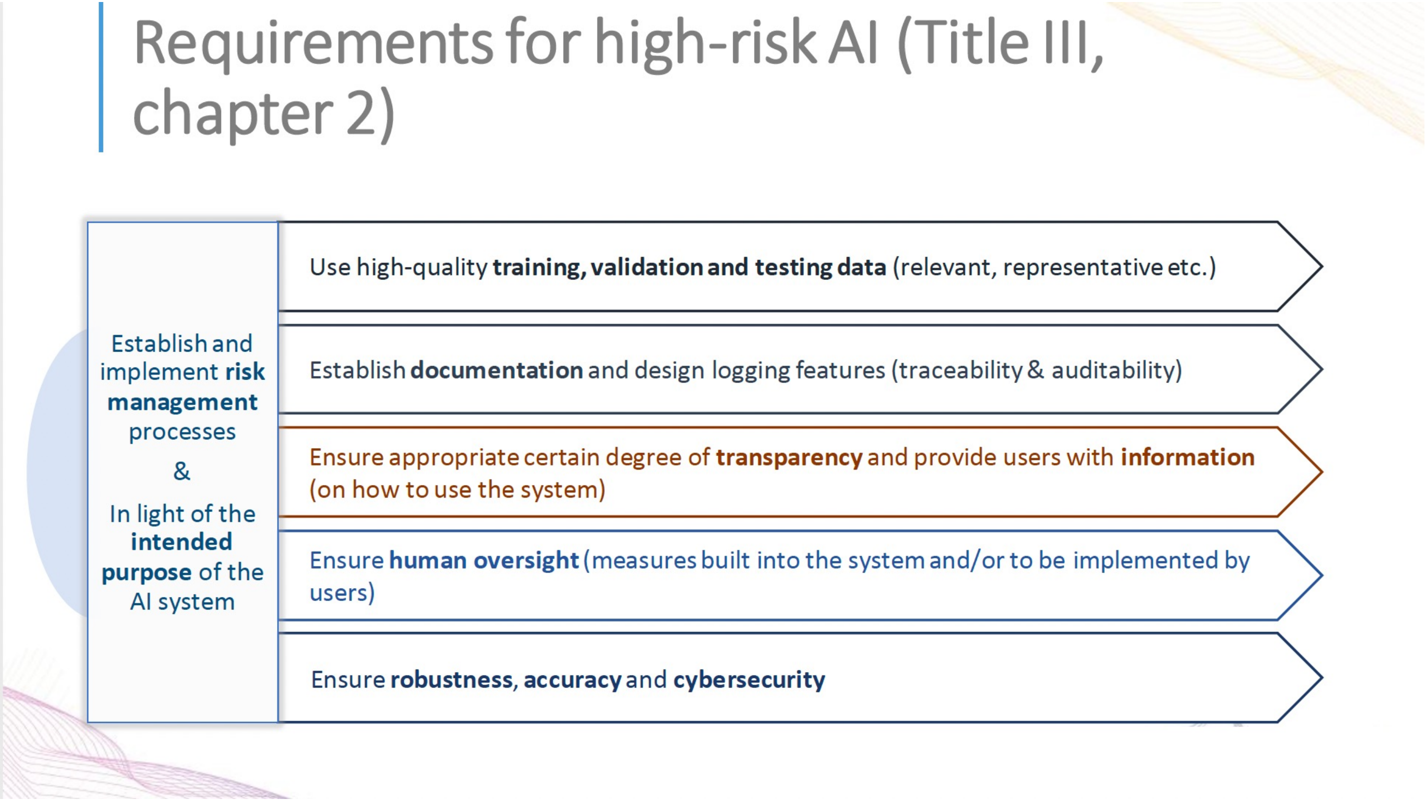
- Law enforcement, which includes a broad range of AI systems used, among other things, to assess the risk of any individual committing an offence, or of re-offending; predicting the likelihood of criminal offences (e.g., predictive policing and profiling), as well as the detection and investigation of fraudulent content;

- Border control management, including AI systems used for the control and management of borders, migration and asylum processes, such as validating travel documents and assessing the eligibility for asylum.

- Administration of justice and democratic processes, including any AI system used to assist in the judicial process by assessing and interpreting facts, and/or making legal recommendations in response to facts.

CapAI, Luciano Floridi et al. 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

Requirements for high-risk AI



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

Establish and implement **risk management** processes

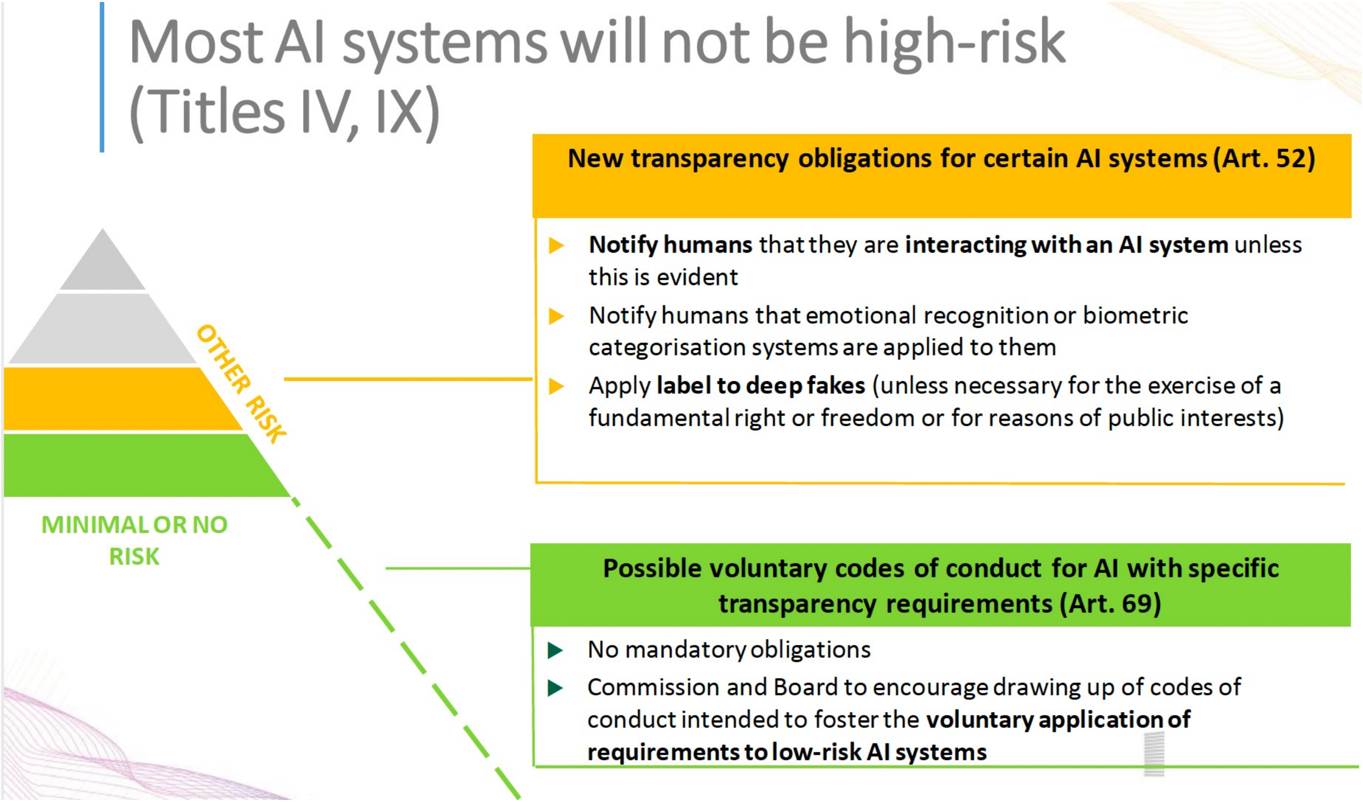
&

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**

[Lucilla Sioli, A European Strategy for Artificial Intelligence, 2021]

The EU risk-based approach to AI regulation



OTHER RISK
MINIMAL OR NO RISK

Most AI systems will not be high-risk (Titles IV, IX)

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

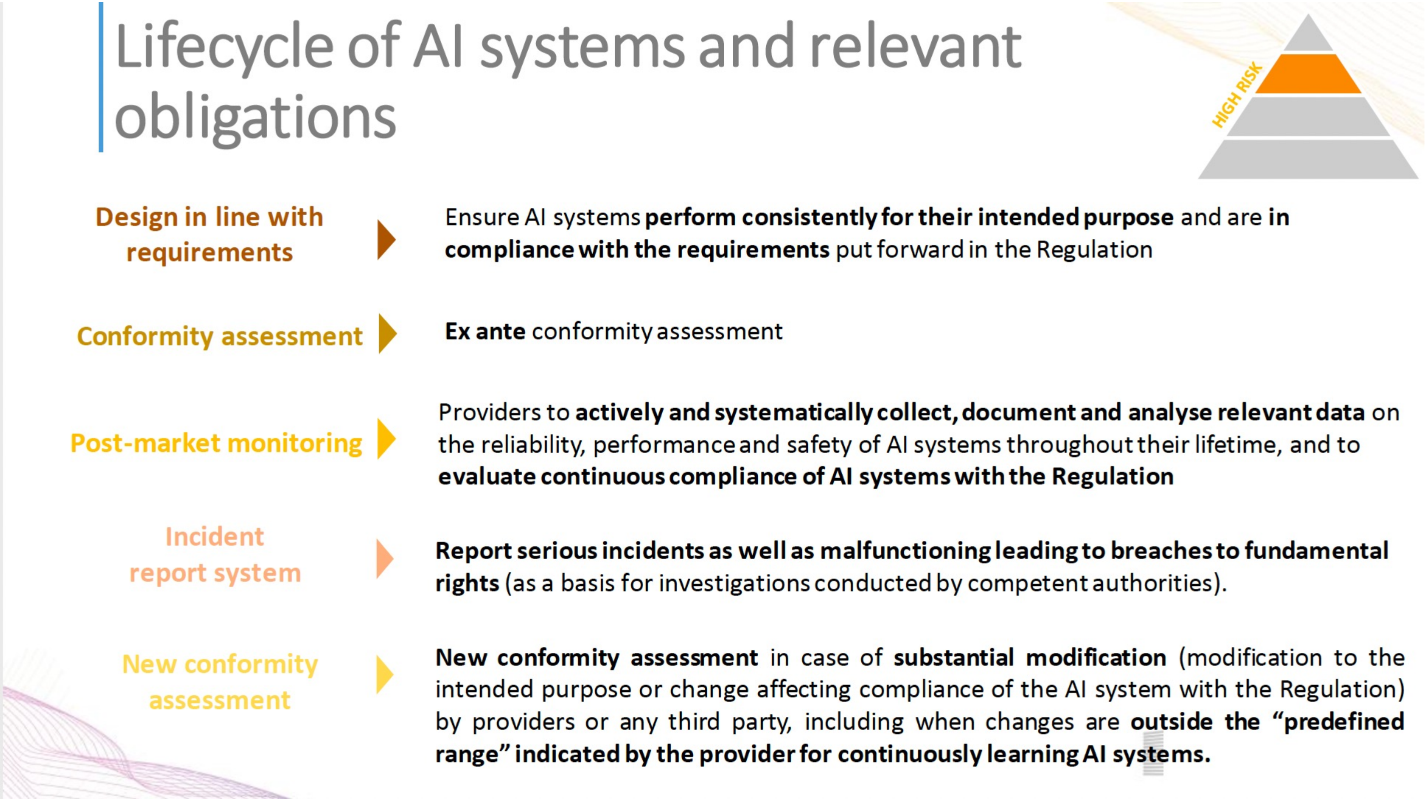
- **Notify humans** that they are **interacting with an AI system** unless this is evident
- Notify humans that emotional recognition or biometric categorisation systems are applied to them
- Apply **label to deep fakes** (unless necessary for the exercise of a fundamental right or freedom or for reasons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

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**

[Lucilla Sioli, A European Strategy for Artificial Intelligence, 2021]

Lifecycle of AI



Lifecycle of AI systems and relevant obligations

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.**

[Lucilla Sioli, A European Strategy for Artificial Intelligence, 2021]

Last news (December 9th, 2023)

After a long negotiations, Parliament and Council of Europe reached a new, provisional, agreement on the Artificial Intelligence Act…

Last news (December 9th, 2023)

Banned applications of AI

- Biometric categorisation systems that use sensitive characteristics (e.g. political, religious, philosophical beliefs, sexual orientation, race);
- Untargeted scraping of facial images from the internet or CCTV footage to create facial recognition databases;
- Emotion recognition in the workplace and educational institutions;
- Social scoring based on social behaviour or personal characteristics;
- AI systems that manipulate human behaviour to circumvent their free will;
- AI used to exploit the vulnerabilities of people (due to their age, disability, social or economic situation).

Exceptions: the use of remote biometric identification systems (sometimes referred to as live facial recognition) in publicly accessible spaces is allowed where strictly necessary for **law enforcement** purposes.

Last news (December 9th, 2023)

Foundational models, large language models (e.g., ChatGPT)

High impact/high risk

If the computing power used for training foundational models/large language models exceeds **10^{25} FLOPs** (floating point operations per second), then these systems will have to comply with additional regulations (that should be in force within 1 year) before that they can be put on the market

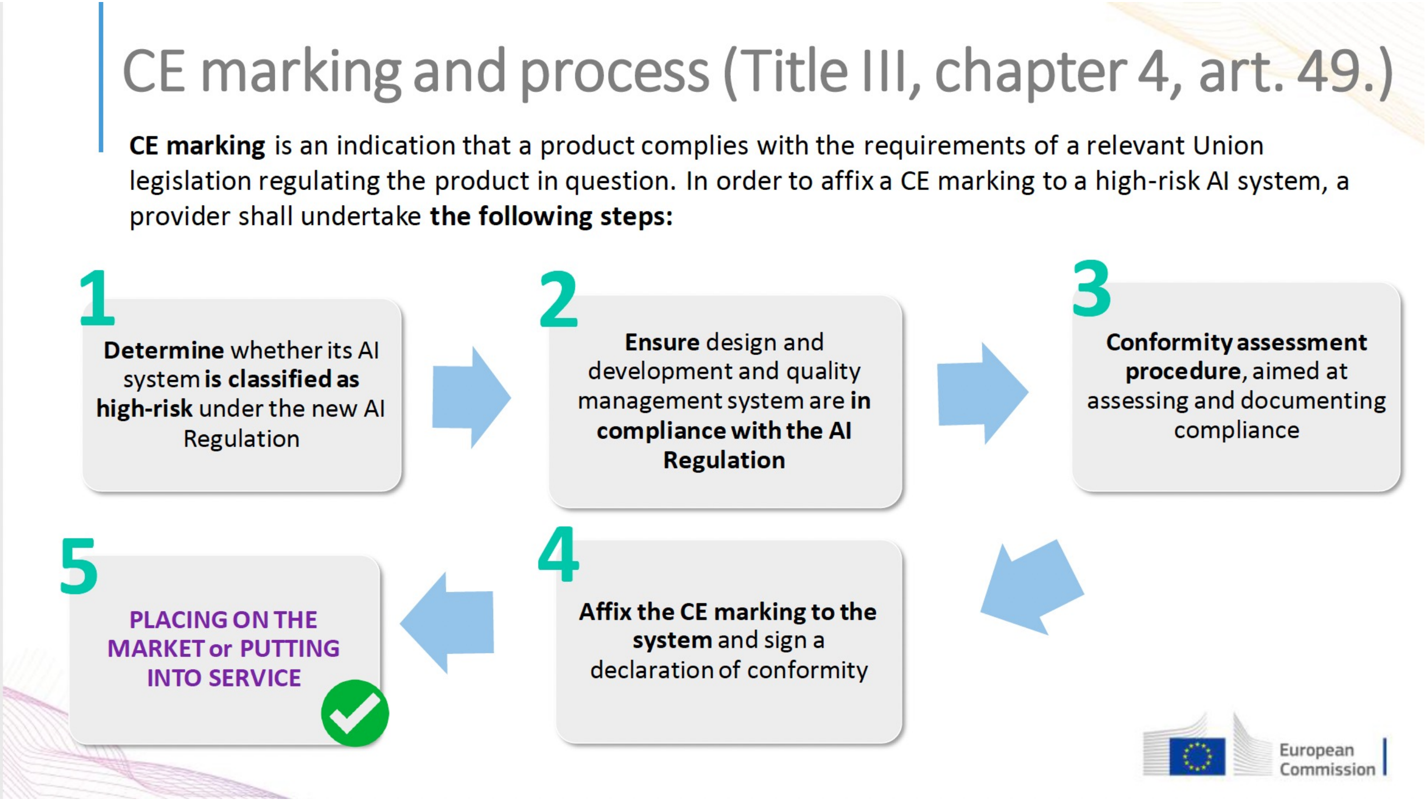
Who should be in charge for comformity assessment?

- The EU AI act proposal (June 2023) requires Member States to designate one or more competent authorities, including a **national supervisory authority** and **CABs (conformity assessment bodies)**, which would be tasked with supervising the application and implementation of the regulation

- National **market surveillance authorities** would be responsible for assessing operators' compliance with the obligations and requirements for high-risk AI systems. They would have access to confidential information (including the source code of the AI systems) and subject to binding confidentiality obligations.

- Article 11 of the EU AI Act
  - *Producers must provide the technical documentation shall be drawn up in such a way to demonstrate that the high-risk AI system complies with the requirements and provide national competent authorities and notified bodies with all the necessary information to assess the compliance of the AI system with those requirements. It shall contain, at a minimum, the elements set out in Annex IV.*

CE marking



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 ✓

What are the penalties for non-conformance to AIA?

The penalties set out in the AIA for non-conformance are, in principle, very similar to those set out in the GDPR.

Three main levels:

- Non-compliance with regard to prohibited AI practices, and/or the data and data governance obligations set out for high-risk AI systems can incur **a penalty of up to €30m**, or **6% of total worldwide turnover** in the preceding financial year (whichever is higher).
- Non-compliance of an AI system with any other requirement under the AIA than stated above can incur a penalty of **up to €20m, or 4% of total worldwide turnover** in the preceding financial year (whichever is higher).
- Supply of incomplete, incorrect or false information to notified bodies and national authorities in response to a request can incur a penalty of **up to €10m, or 2% of total worldwide turnover** in the preceding financial year (whichever is higher).

capAI, L. Floridi et al., 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

Are we ready for Conformity Assessment of AI Systems?

- Looks quite hard…both for Italian industry and competent authorities, including the **national supervisory authority**…

- The EU AI Act gives prescriptions, but, no practical guidelines and tools for conformity assessment in line with the act…There is a **lack** of **technical standards** so far.

- **Who actually decides** how **to check** the requirements of the EU AI Act and how these requirements **are met**?

- Recently, ENISA (the European Union Agency for Cybersecurity) wrote:

  > *we should ensure that the actors performing conformity assessment on AI systems have standardised tools and competences, including on cybersecurity.*

[ENISA, Cybesecurity of AI and standardization, March 2023]

Are we ready for Conformity Assessment of AI Systems?

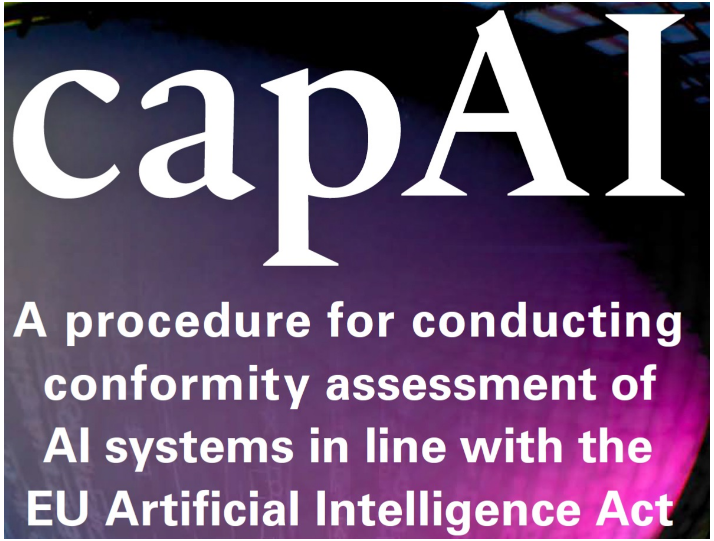
capAI, L. Floridi et al., 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

- Recently (in 2022), **capAI** has been published…

- **capAI** is a conformity assessment procedure for AI systems, to provide an independent, comparable, quantifiable, and accountable assessment of AI systems that conforms with the proposed the **EU AI Act** (AIA) regulation.

- The main purpose of capAI is to serve as a governance tool that ensures and demonstrates that the development and operation of an AI system are **trustworthy,** i.e., **legally compliant**, **ethically** sound, and **technically robust**, and thus **conform** to the **AIA**.

- But also capAI does not provide technical tools and practical assistance…just prescriptions and guidelines…


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

What is missing…

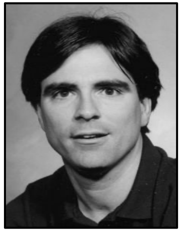
- Technical **standards** and **practical tools** for the semi-automatic production of documentation on the security, robustness, fairness, privacy,…, of AI systems in line with the EU AI Act.

- Technical **standards** and **practical tools** should allow companies or conformity assessment bodies to assess the compliance of AI systems with the EU AI Act

- So far, the EU AI Act says what to do but not how to do it in details…

Thanks for Listening!

Any questions?



Engineering isn't about perfect solutions; it's about doing the best you can with limited resources
(Randy Pausch, 1960-2008)