Università
di Genova

Pattern Recognition
and Applications Lab

University of
Cagliari, Italy

From known knowns to unknown unknowns and Trustworthy AI

Fabio Roli

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…

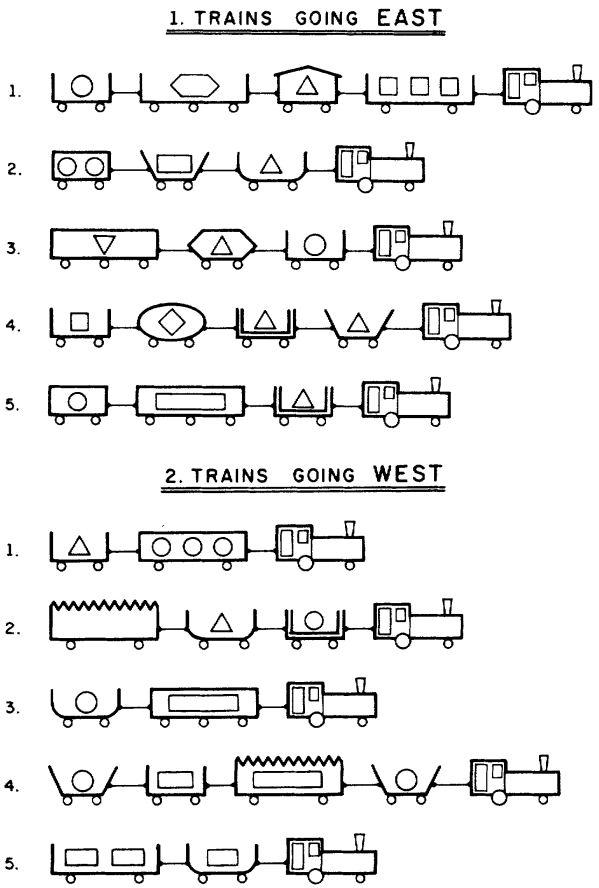


Fig. 4.

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

Machine learning of "known knowns"

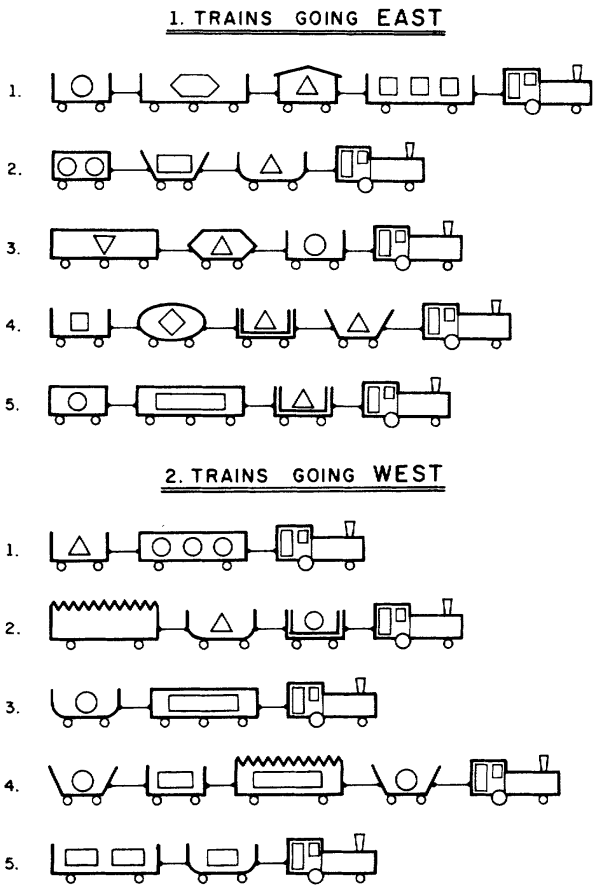


Fig. 4.

Eastbound Trains:

$$\exists\, car_1\ [length(car_1)=short]\ [car\text{-}shape(car_1)=closed\ top]$$
$$::>\ [class=Eastbound] \qquad (22)$$

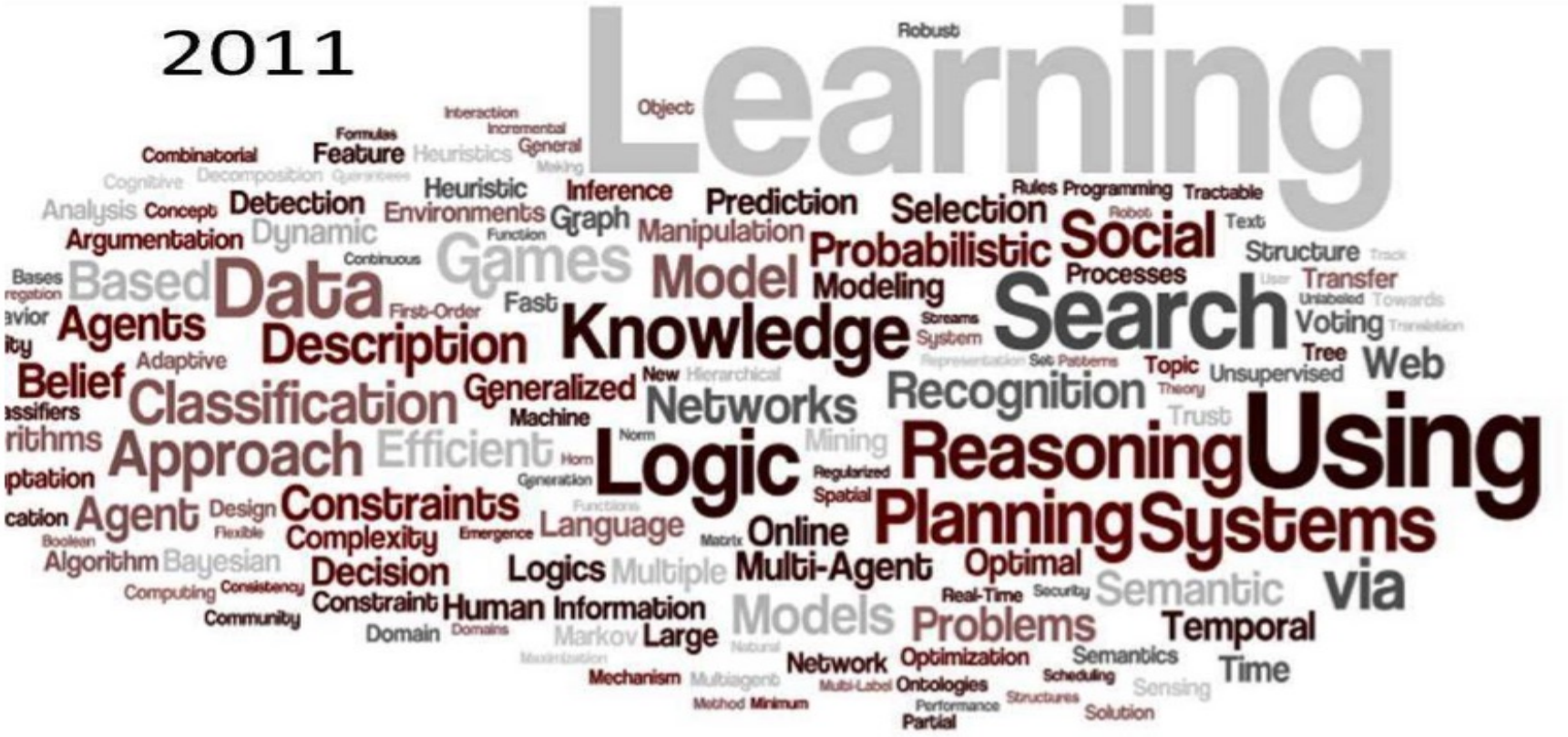
IF a train contains a car that is short and has a closed top, THEN it is an Eastbound train

- «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 50 years of research in AI, the main stream is **Big Data + Deep Learning + GPUs**

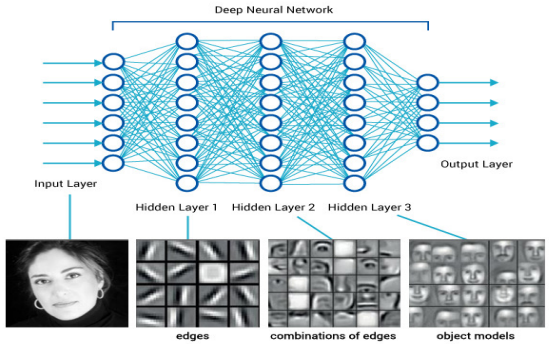
Big Data

Facebook 350 millions of images per day

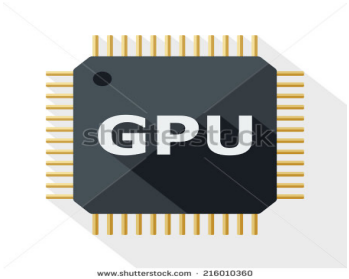
Walmart 2.5 Petabytes customer data hourly

YouTube 300 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
- Unfortunately, there is a whole world out there where these assumptions are violated…

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

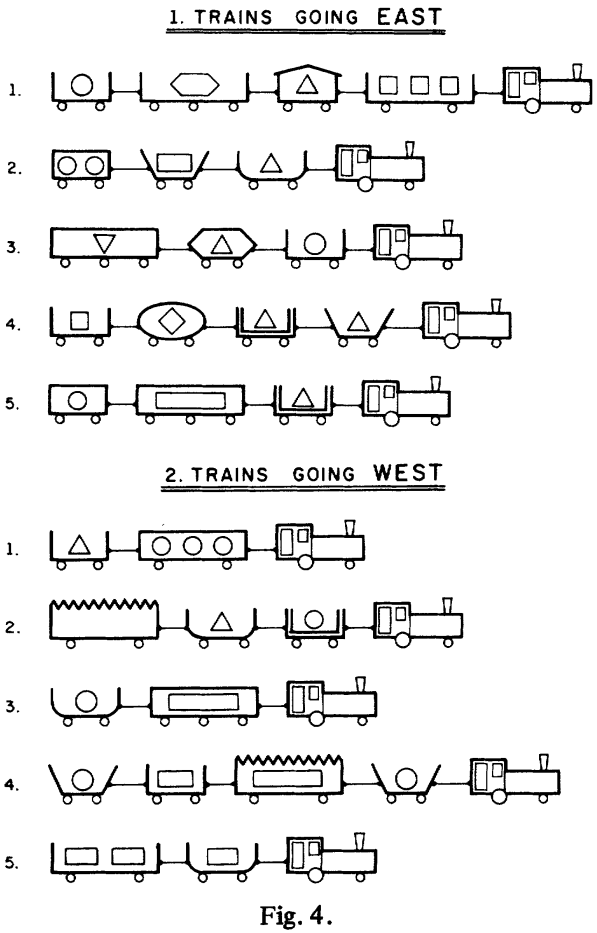
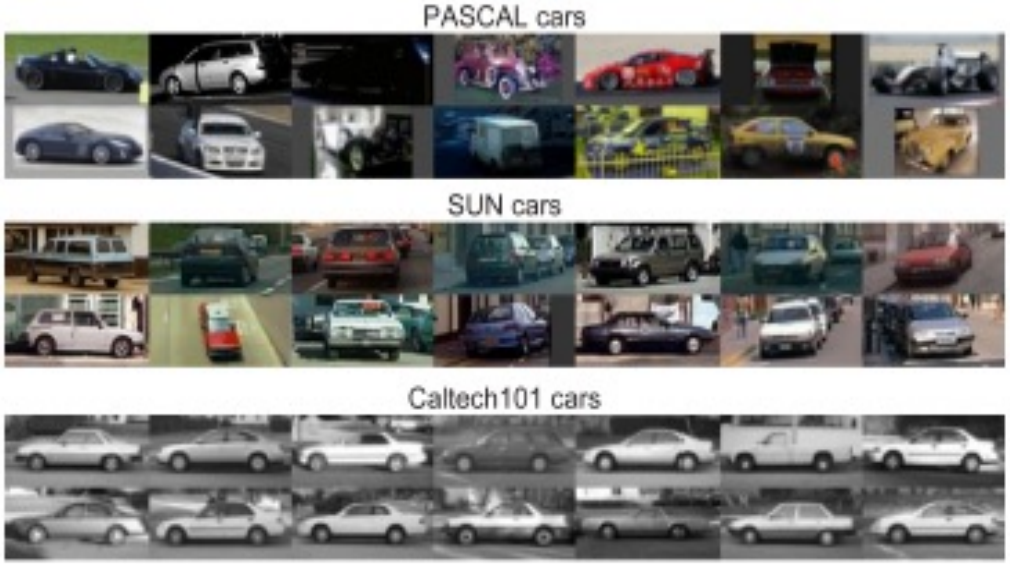


Fig. 4.



Bigger is better…

4 classes
2209 objects

2005

10 classes
4754 objects

2006

The PASCAL VOC data set

21.841 classes
14.197.122 images

2020

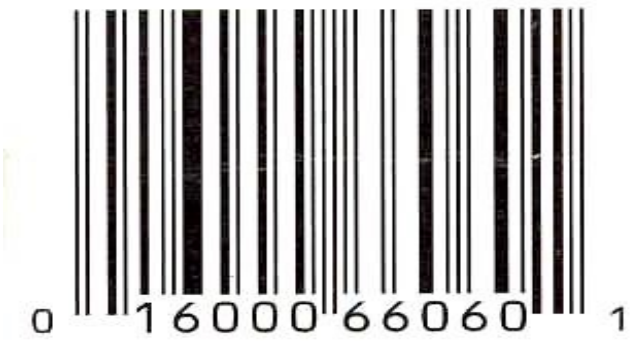
image-net.org

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.

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	Train on: \ Test 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

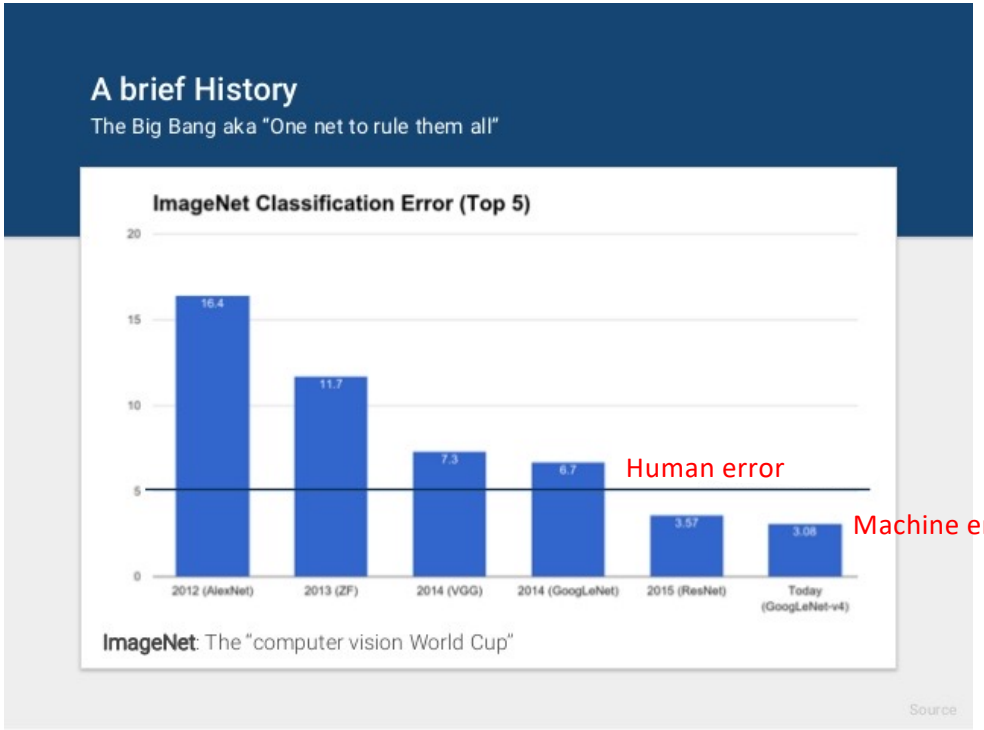
(New Scientist, 2016)

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

The bright side of AI: super human performances…



ImageNet Challenge

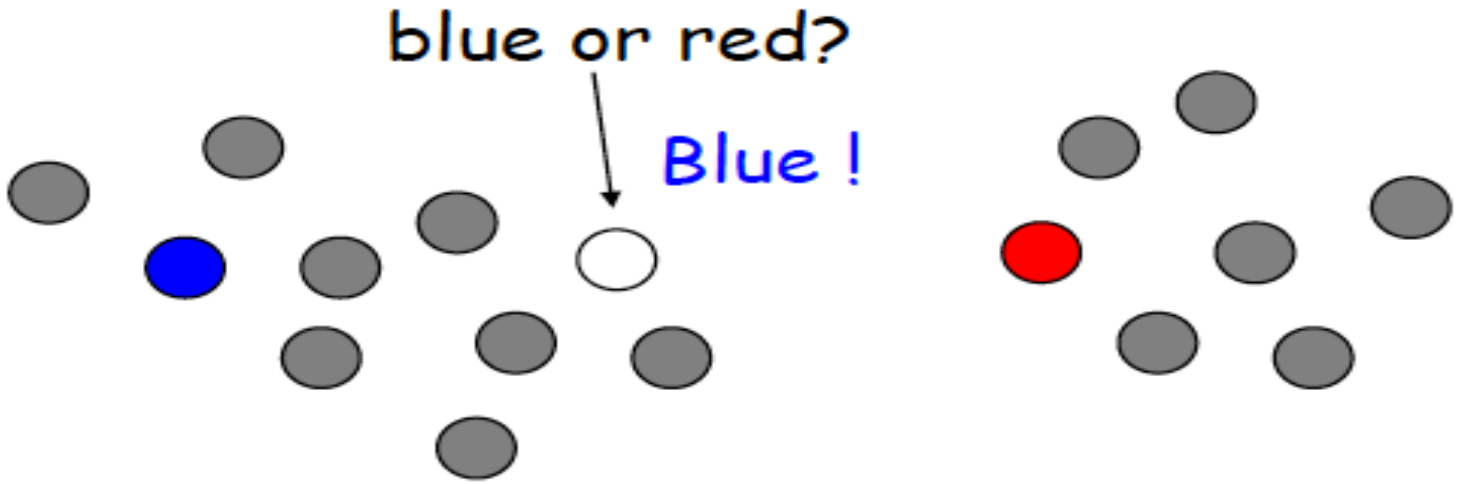
IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

High accuracy is high robusteness?

- What about the performance of machine-learning algorithms on data which has been modified very slightly ?

The smoothness assumption



Points close to each other are more likely to share a label

This is generally assumed in machine learning…

High accuracy is high robusteness?

How a machine-learning algorithm should classify these two images?

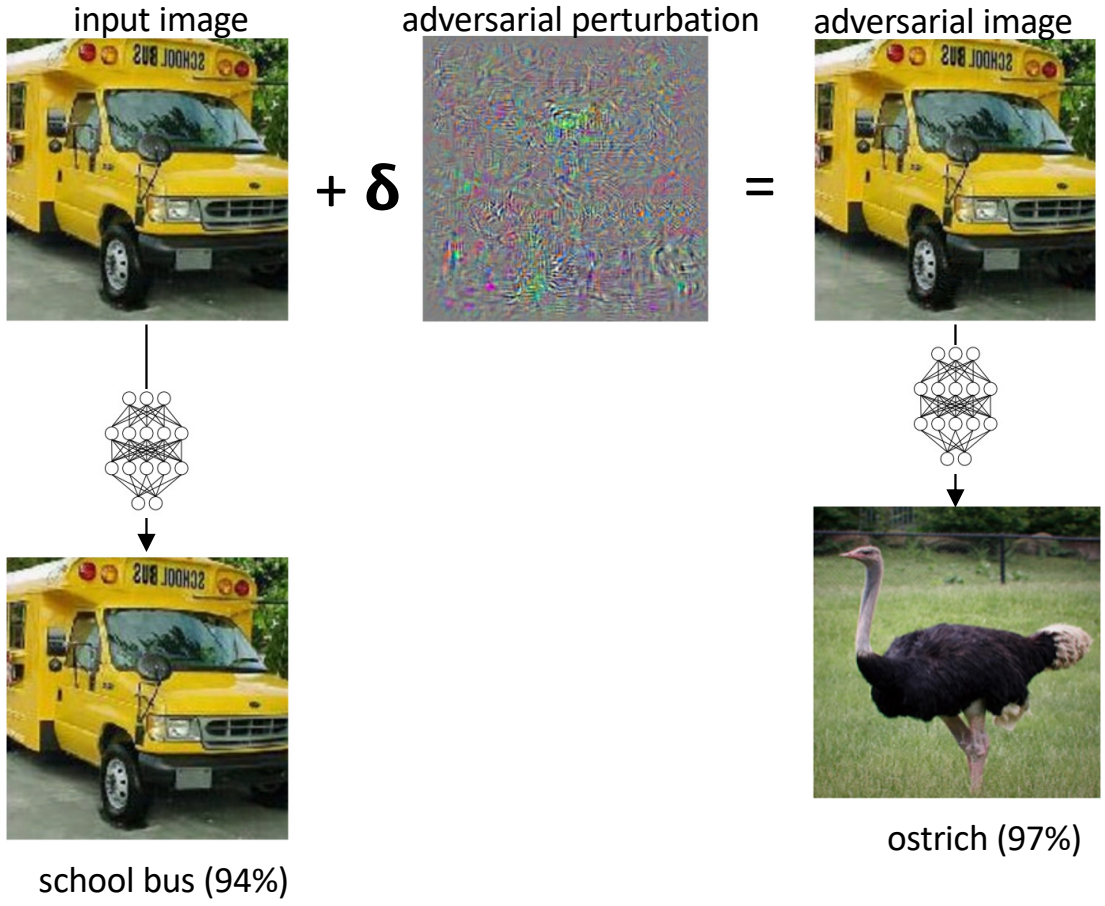
With the same or a different classification label?

2013-2014: Adversarial examples

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

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

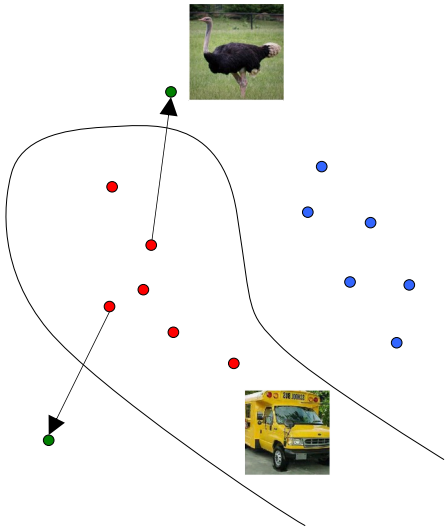
The adversarial image $x+\boldsymbol{\delta}$ is visually hard to distinguish from x
Informally speaking, the solution $x+\boldsymbol{\delta}$ is the closest image to x classified as l=*ostrich*



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**

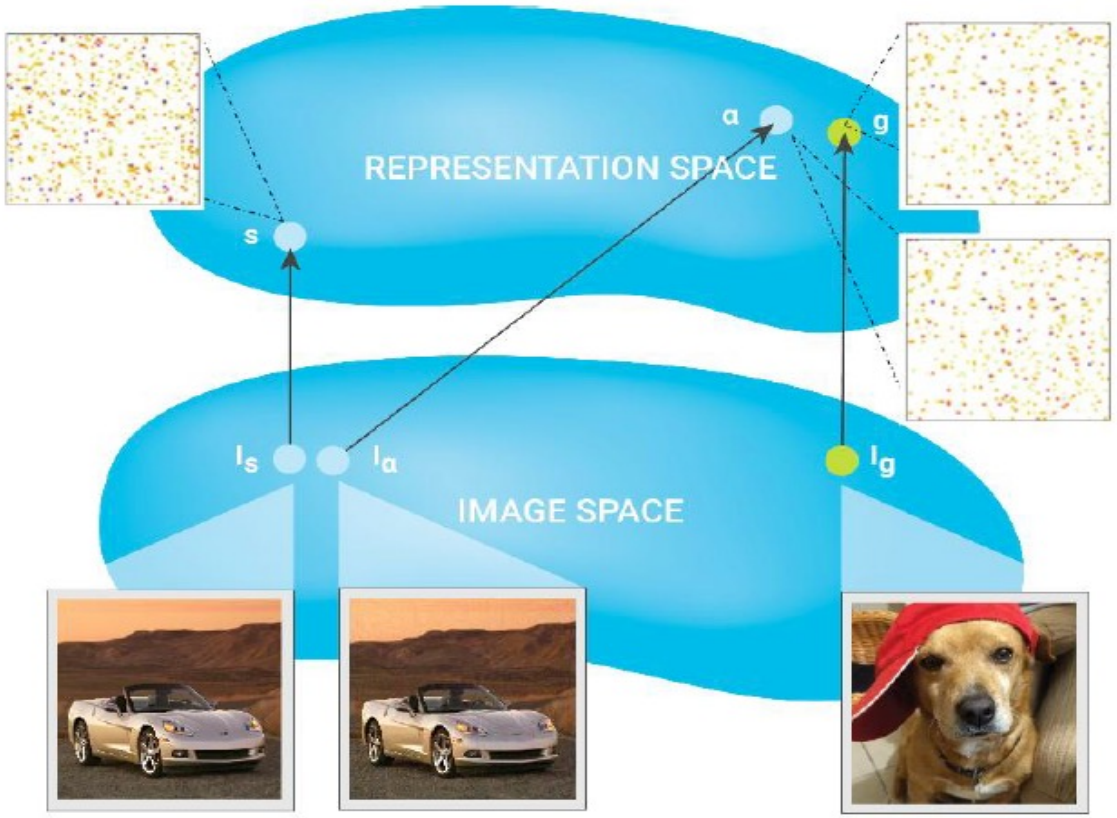
Unknown unknowns in machine learning

- Adversarial examples can occur in *blind spots*
  - Regions far from training data that are anyway assigned to 'legitimate' classes



blind-spot evasion
(not even required to
mimic the target class)

Unknown unknowns in machine learning

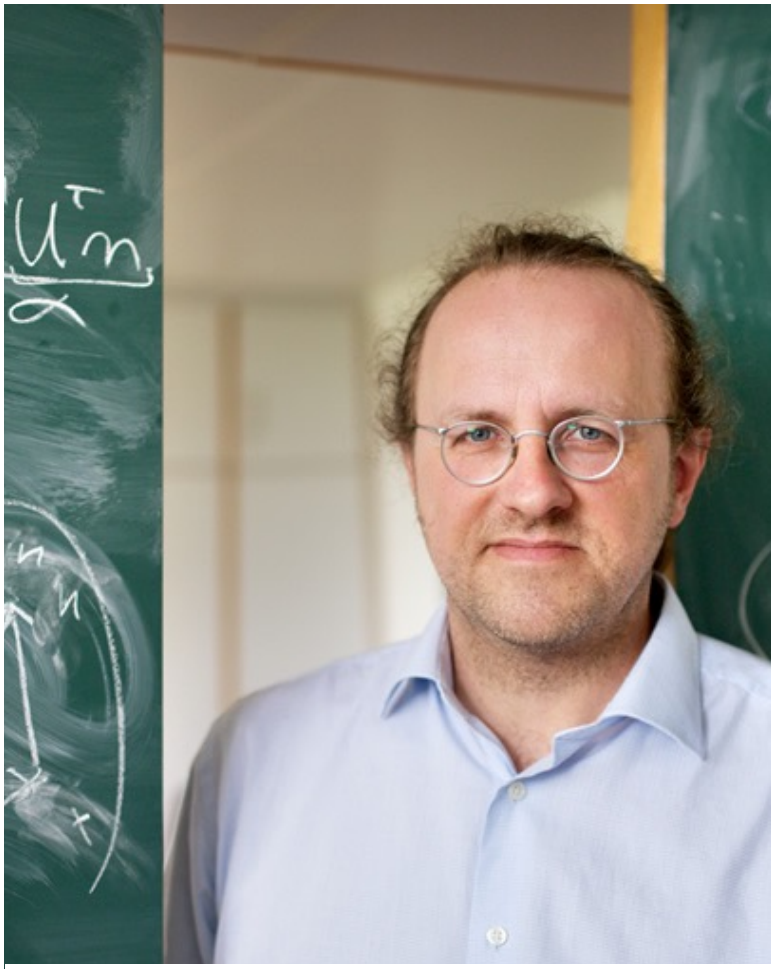


[S. Sabour at al., ICLR 2016]

Why Is ML so Vulnerable?

The i.i.d. assumption

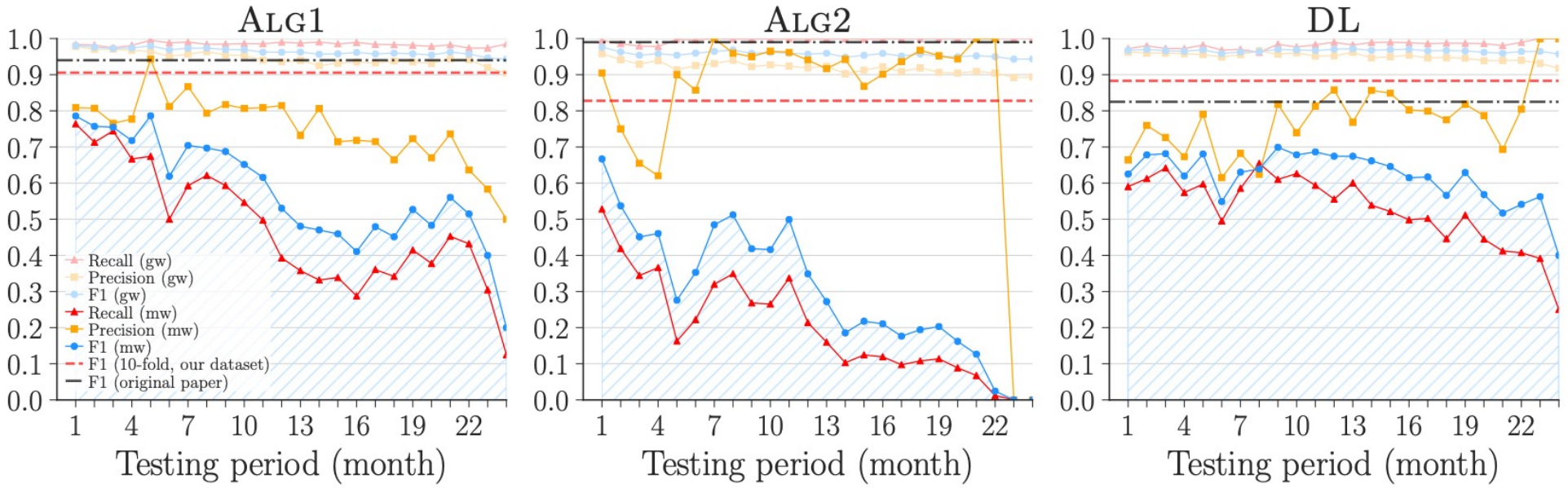
- **Underlying assumption:** past data is *representative* of future data (IID data)
- The success of modern AI is on tasks for which we collected enough representative training data
- **We cannot build AI models for each task an agent is ever going to encounter,** but there is a whole world out there where the IID assumption is violated
- **Adversarial attacks** point exactly at this lack of robustness which comes from IID specialization



Bernhard Schölkopf
Director, Max Planck Institute, Tuebingen, Germany

What Does It Imply in Practice?

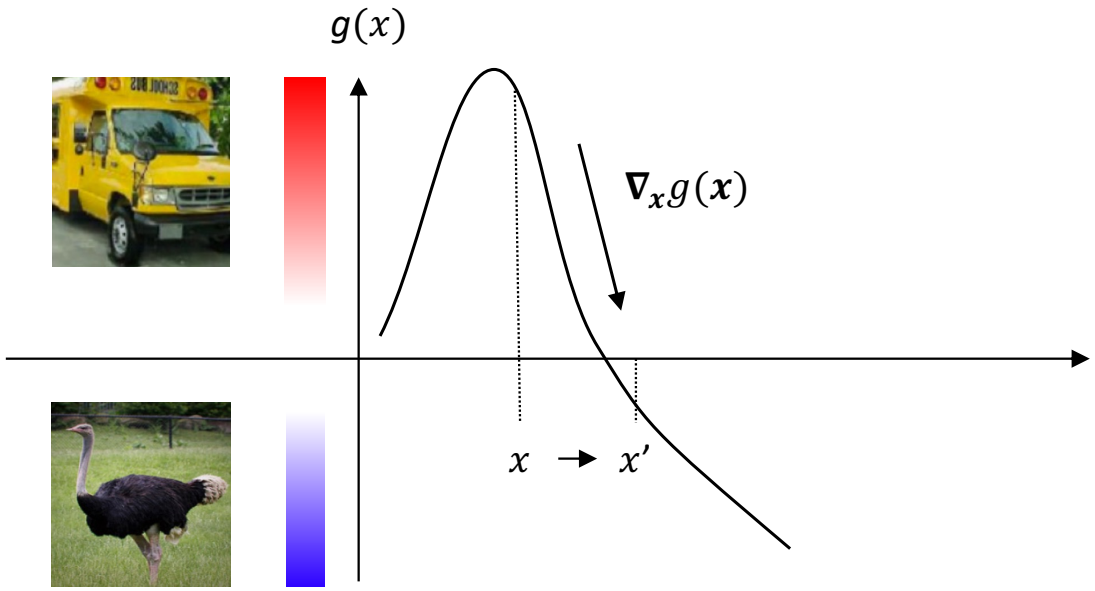
- Performance of ML under adversarial/temporal drift decreases over time
  - Models need to be retrained and updated (reactively)



Pendlebury, Cavallaro et al., TESSERACT, USENIX Sec. 2019

Large gradient

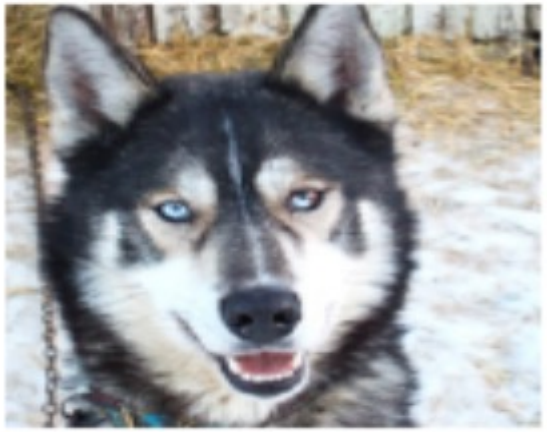
- Large sensitivity of g(**x**) to input changes
  - i.e., the **input gradient** $\nabla_x g(\boldsymbol{x})$ has a large norm (scales with input dimensions!)
  - Thus, even small modifications along that direction will cause large changes in the predictions



[Simon-Gabriel et al., Adversarial vulnerability of NNs increases with input dimension, 2018]

Which are the most relevant features for learning algorithms?

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

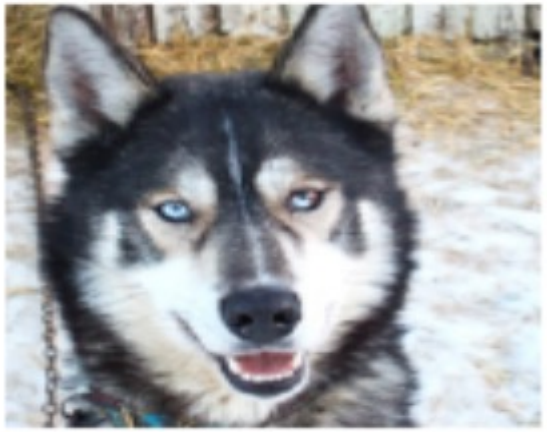


(a) Husky classified as wolf

if pixels are our input features, which pixels are more correlated with label=wolf ?

Spurious correlations

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



(a) Husky classified as wolf

if pixels are our input features, which pixels are more correlated with label=wolf ?

(b) Explanation

Accuracy of AI is not enough

Explainability Is Another Important Asset for AI Safety

- How can we trust a black-box algorithm providing *opaque* decisions?
  - *Why did my car decide to turn left rather than right?*
  - *Why is this application considered malicious / harmful?*

- The right to explanation (https://en.wikipedia.org/wiki/Right_to_explanation)
  - EU on General Data Privacy Regulation (GDPR), Art. 22

- Important concept
  - to build trust in machines and automated algorithms
  - to understand if the algorithm has properly learned meaningful notions/abstractions from data
  - to uncover potential biases encountered during the learning process

From known-knowns to unknown-unknowns

Known Knowns: 1970

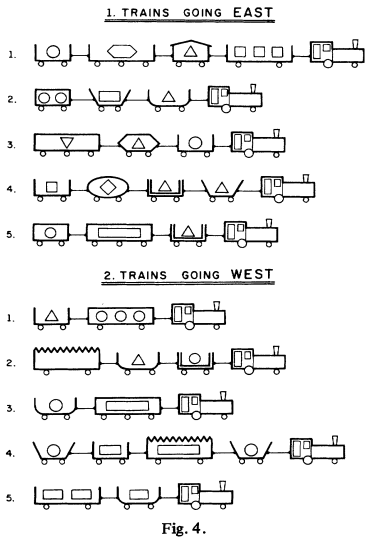
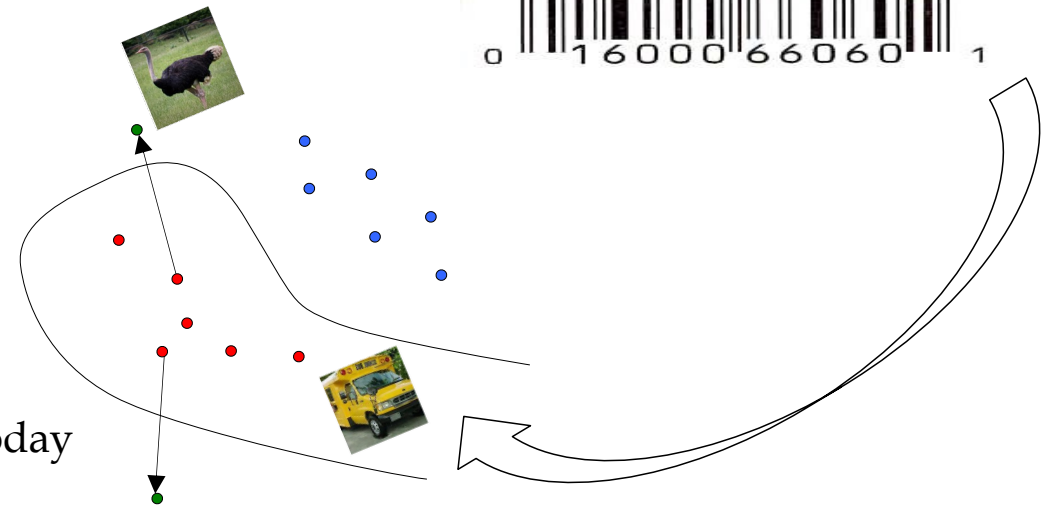


Fig. 4.

Known Unknowns: 1990-today



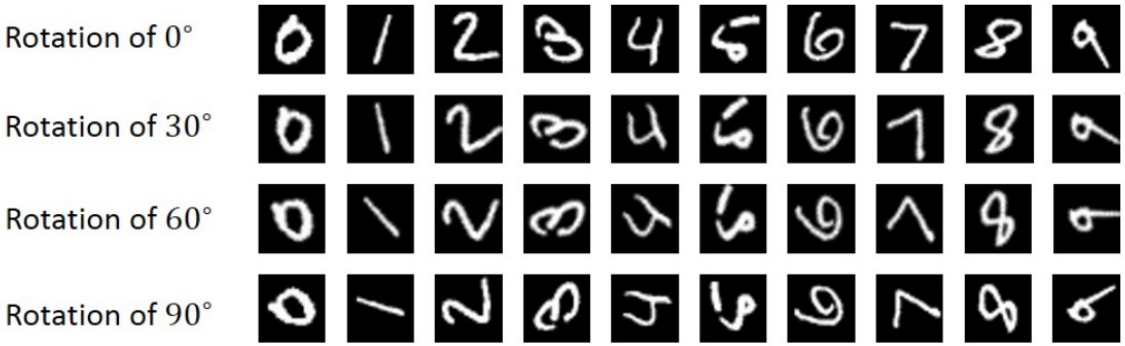
Unknown Unknowns: today

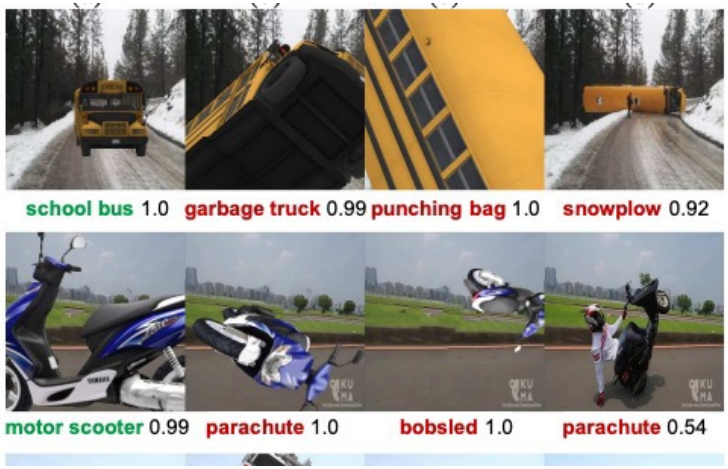
And so, what the next step?

Not only security risks in AI…

Not only security risks in AI…

Lack of Robustness…





Bias, Discrimination, Fairness….

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

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

The EU AI Act for Trustworthy AI

AI is good …

- For citizens
- For business
- For the public interest

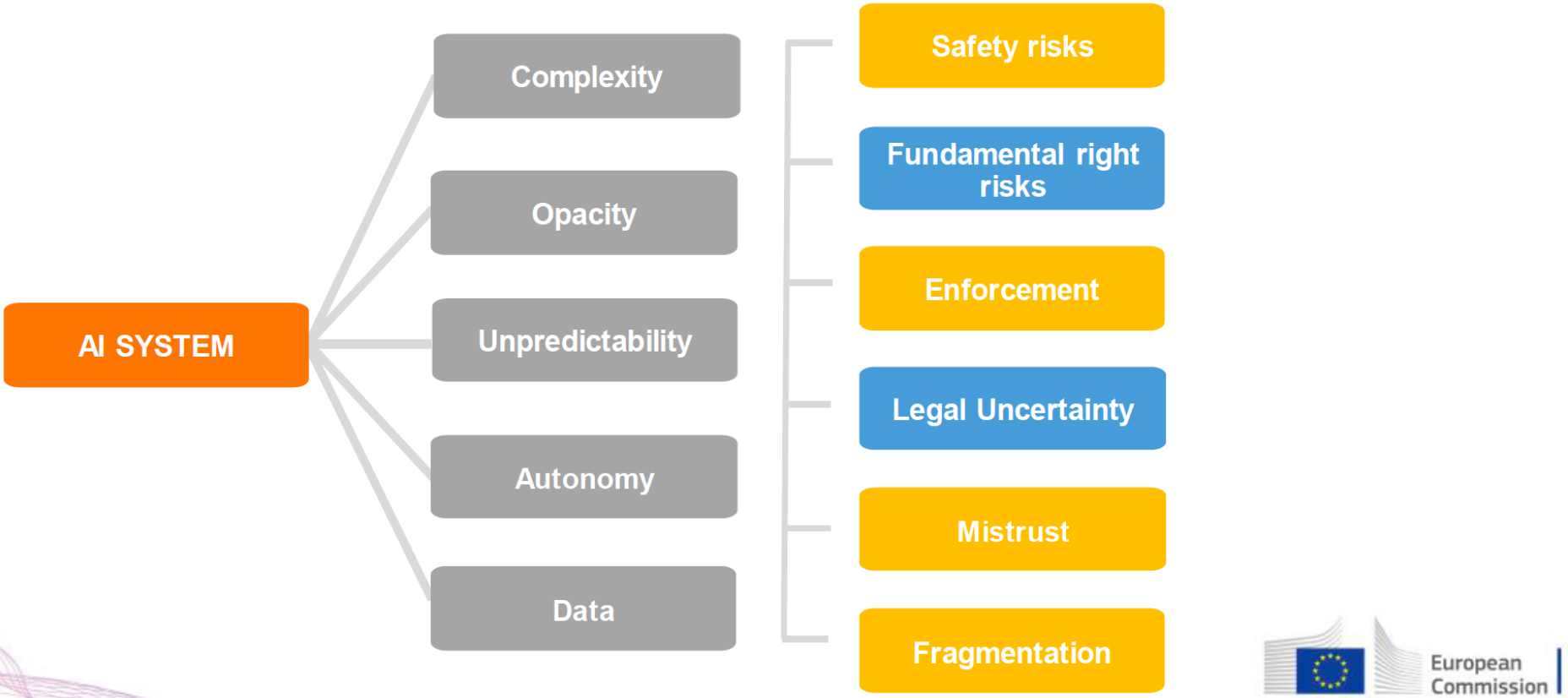
… but creates some risks

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

European Commission

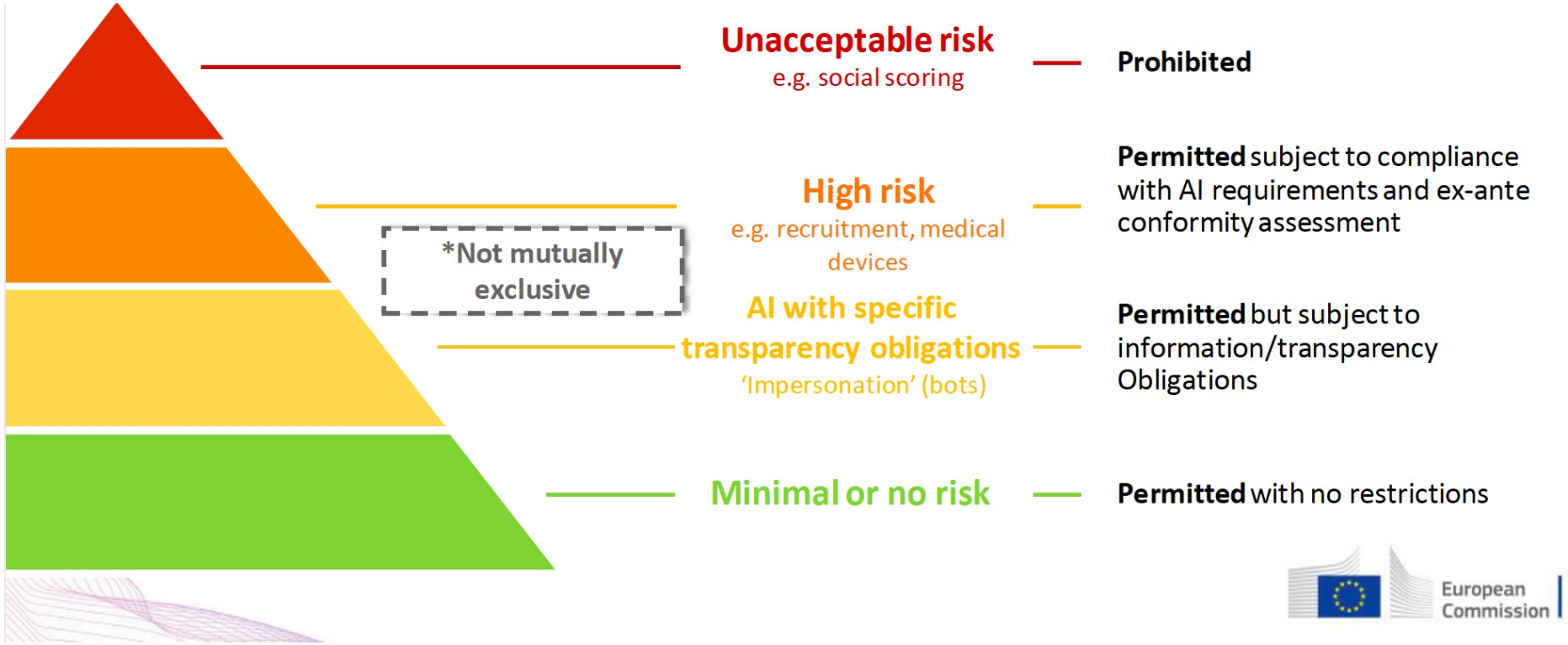
[Lucilla Sioli, A European Strategy for Artificial Intelligence, 2021]

Why should we regulate AI?



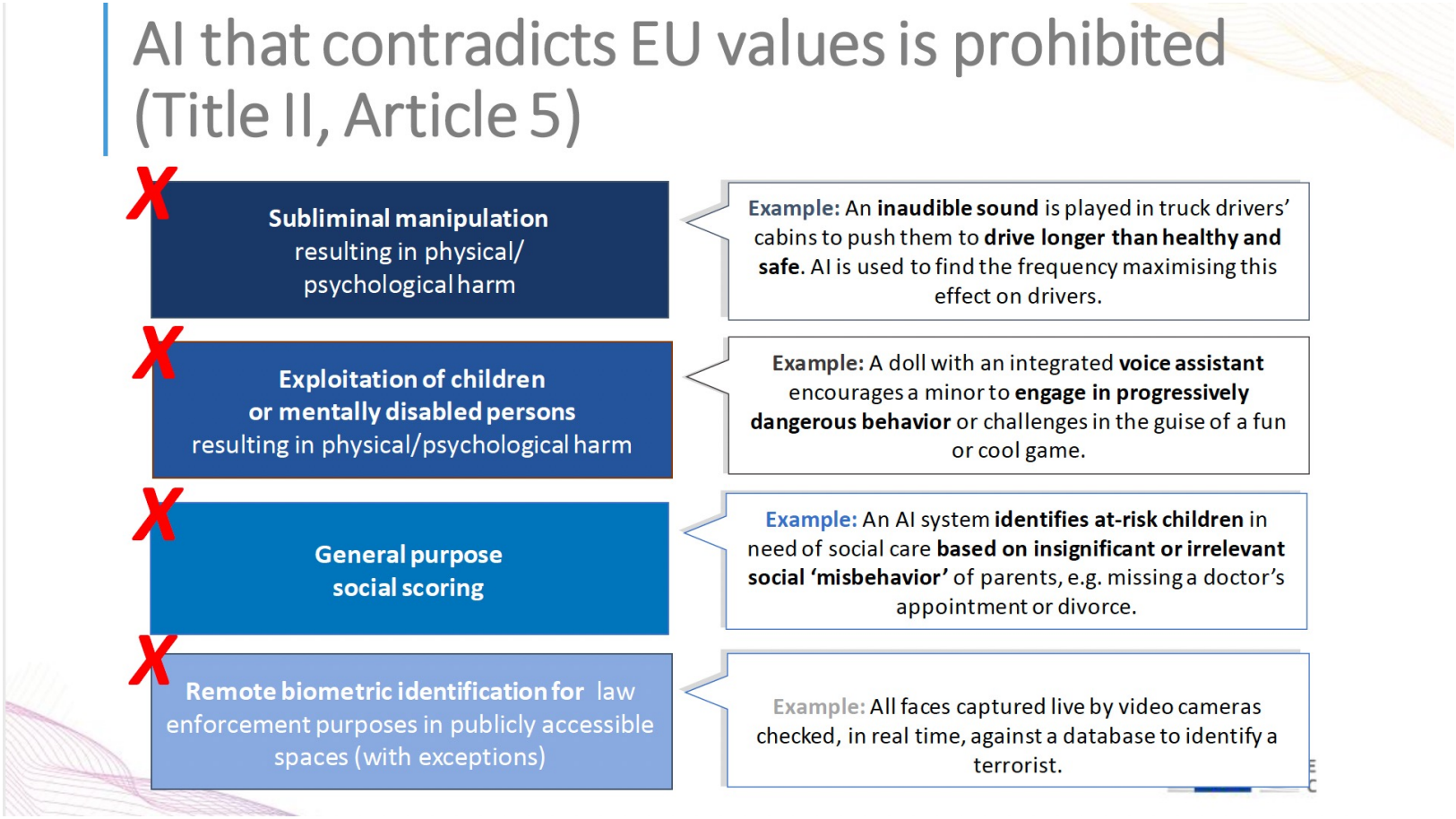
[Lucilla Sioli, A European Strategy for Artificial Intelligence, 2021]

The EU risk-based approach to AI regulation



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

Prohibited applications of AI



[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

The EU risk-based approach to AI regulation

High-risk Artificial Intelligence Systems (Title III, Annexes II and III)

Certain applications in the following fields:

1 SAFETY COMPONENTS OF REGULATED PRODUCTS

(e.g. medical devices, machinery) which are subject to third-party assessment under the relevant sectorial legislation

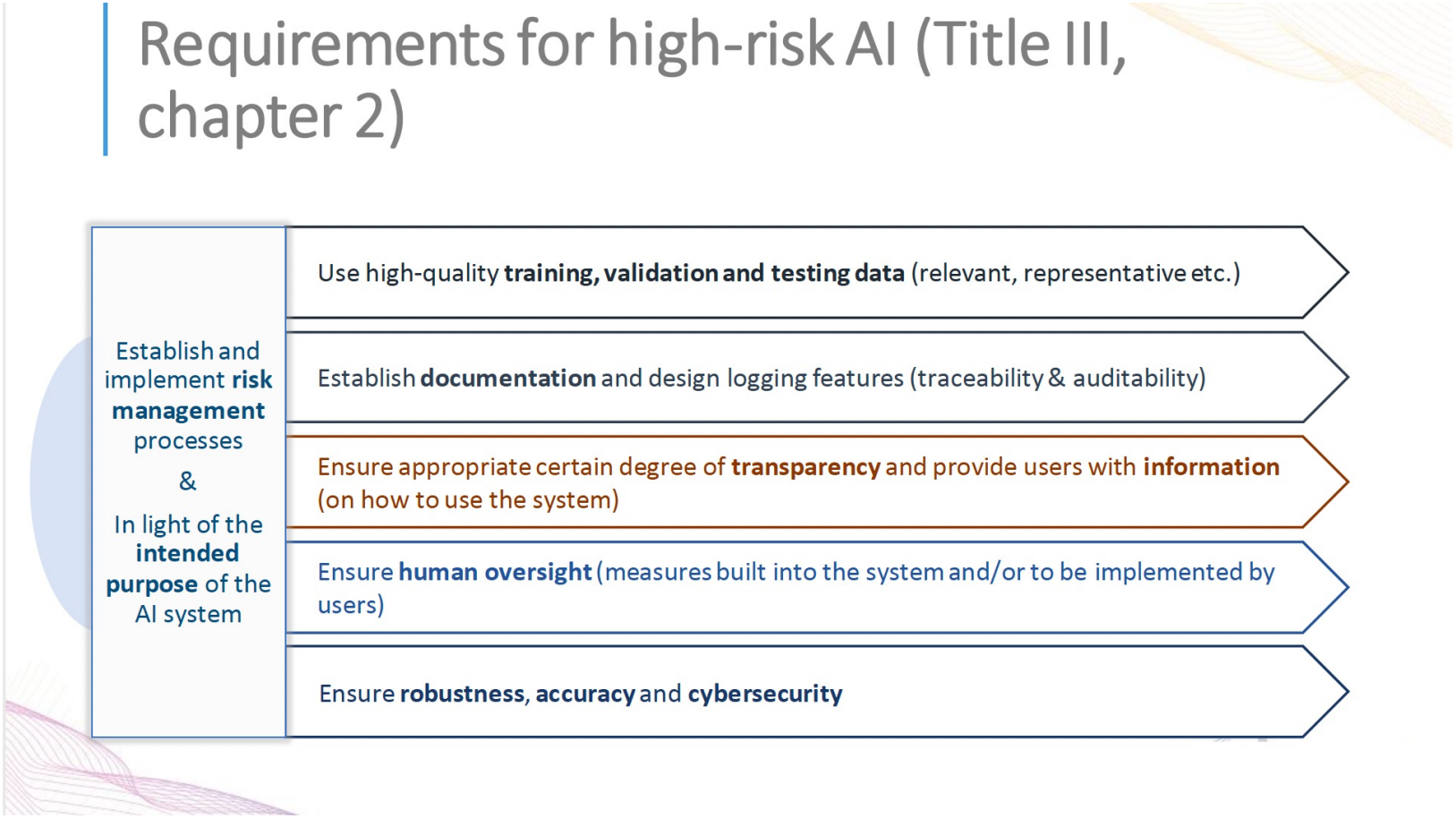
2 CERTAIN (STAND-ALONE) AI SYSTEMS IN THE FOLLOWING FIELDS

- ✓ Biometric identification and categorisation of natural persons
- ✓ Management and operation of critical infrastructure
- ✓ Education and vocational training
- ✓ Employment and workers management, access to self-employment
- ✓ Access to and enjoyment of essential private services and public services and benefits
- ✓ Law enforcement
- ✓ Migration, asylum and border control management
- ✓ Administration of justice and democratic processes

European Commission

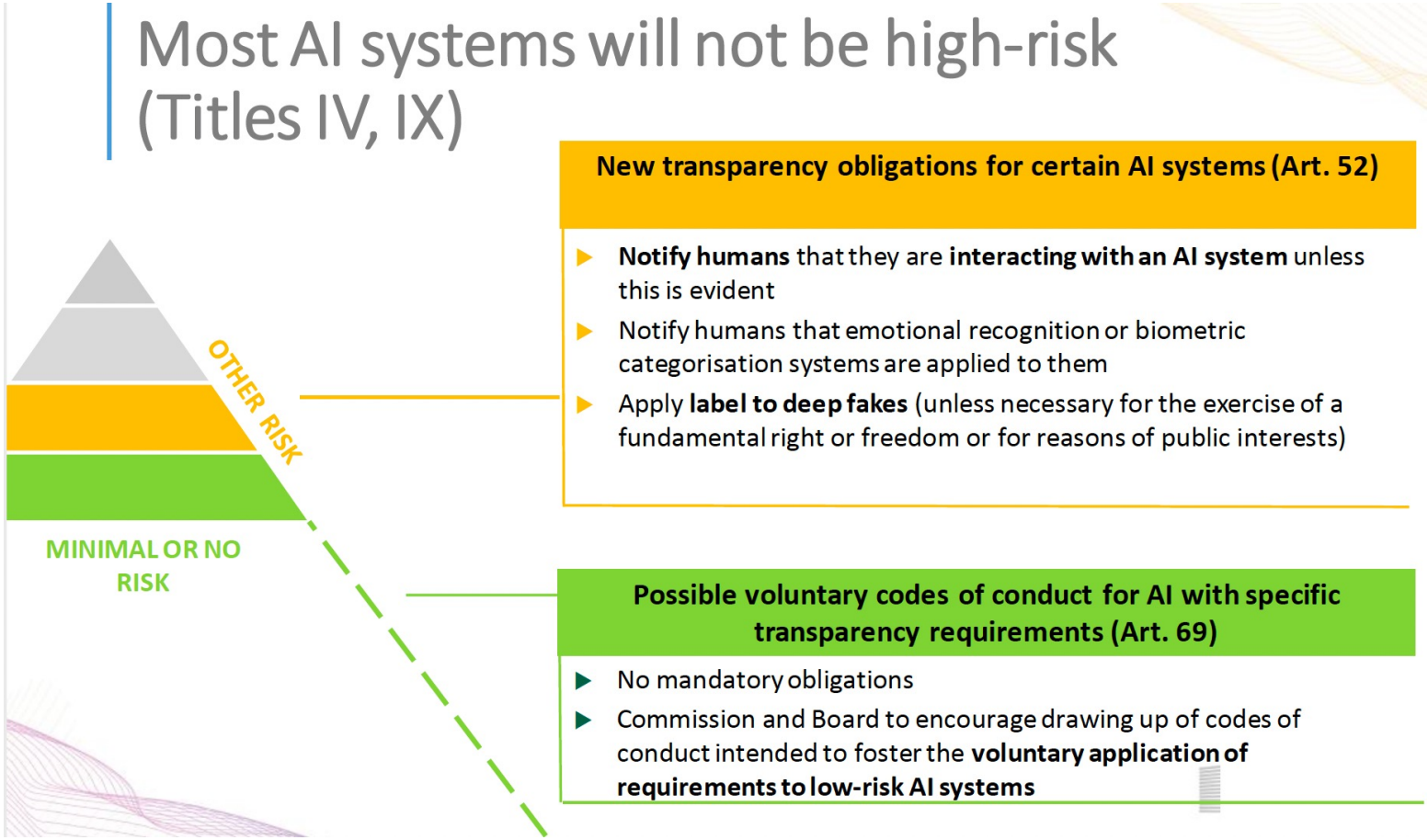
[Lucilla Sioli, A European Strategy for Artificial Intelligence, 2021]

Requirements for high-risk AI



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

The EU risk-based approach to AI regulation



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

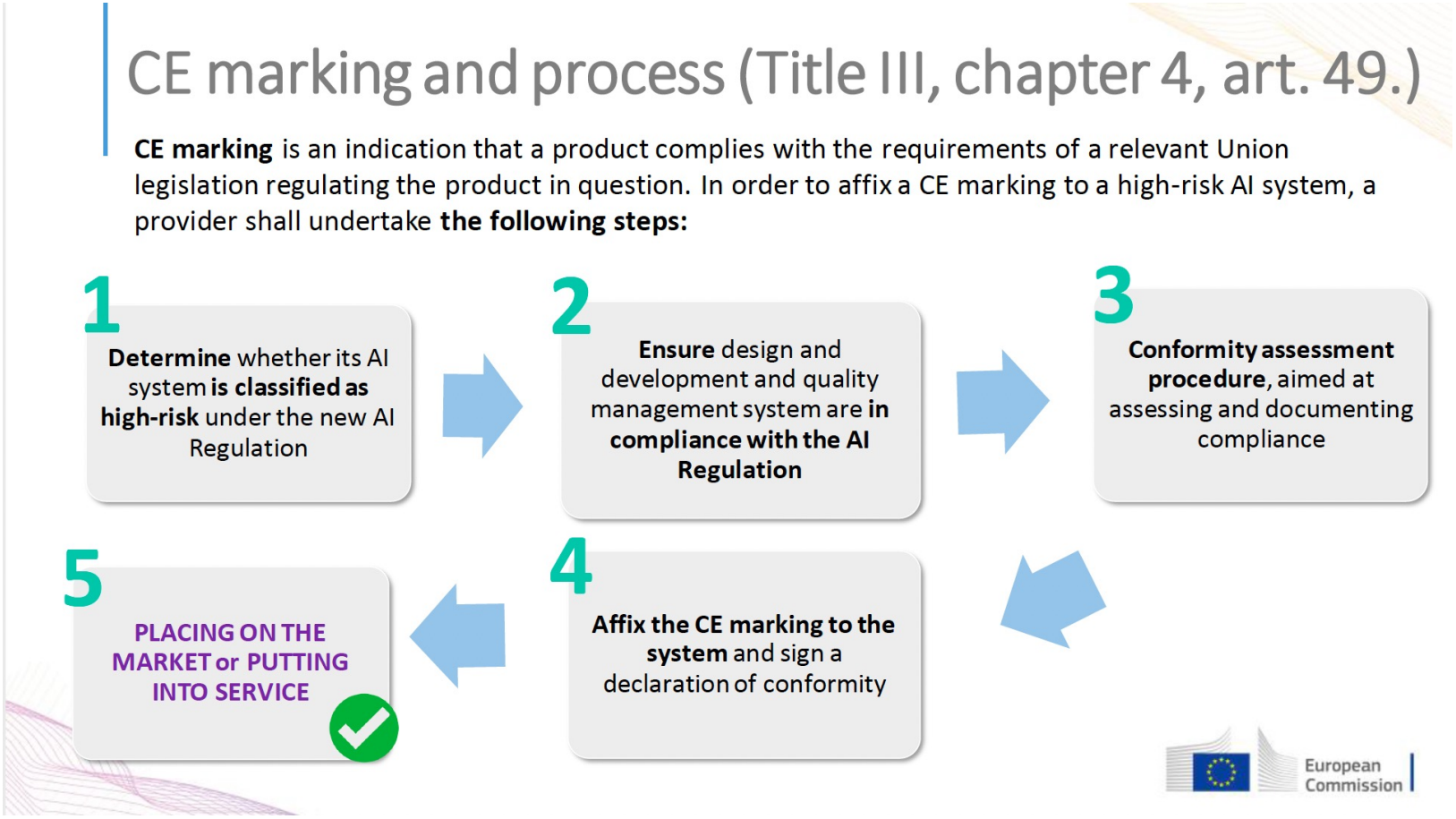
Lifecycle of AI

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]

CE marking



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

capAI: a conformity assessment procedure for AI systems

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

capAI: a conformity assessment procedure for AI systems

- **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**.

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

capAI: a conformity assessment procedure for AI systems

capAI procedure consists of three components:

1. an **internal review protocol** (IRP), which provides organisations with a tool for quality assurance and risk management;
2. a **summary datasheet** (SDS) to be submitted to the EU's future public database on high-risk AI systems in operation;
3. an **external scorecard** (ESC), which can (optional) be made available to customers and other stakeholders of the AI system.

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

Internal review protocol (IRP)

- By following the IRP, organisations can conduct conformity assessment in line with, and create the technical documentation required by the AIA.

- It follows the **development stages** of the **AI system's lifecycle**, and assesses the organisation's awareness, performance, and resources in place to prevent, respond to and rectify potential failures.

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

Summary datasheet (SDS)

- The SDS is a high-level summary of the AI system's purpose, functionality and performance that fulfils the public registration requirements, as stated in the AIA.

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

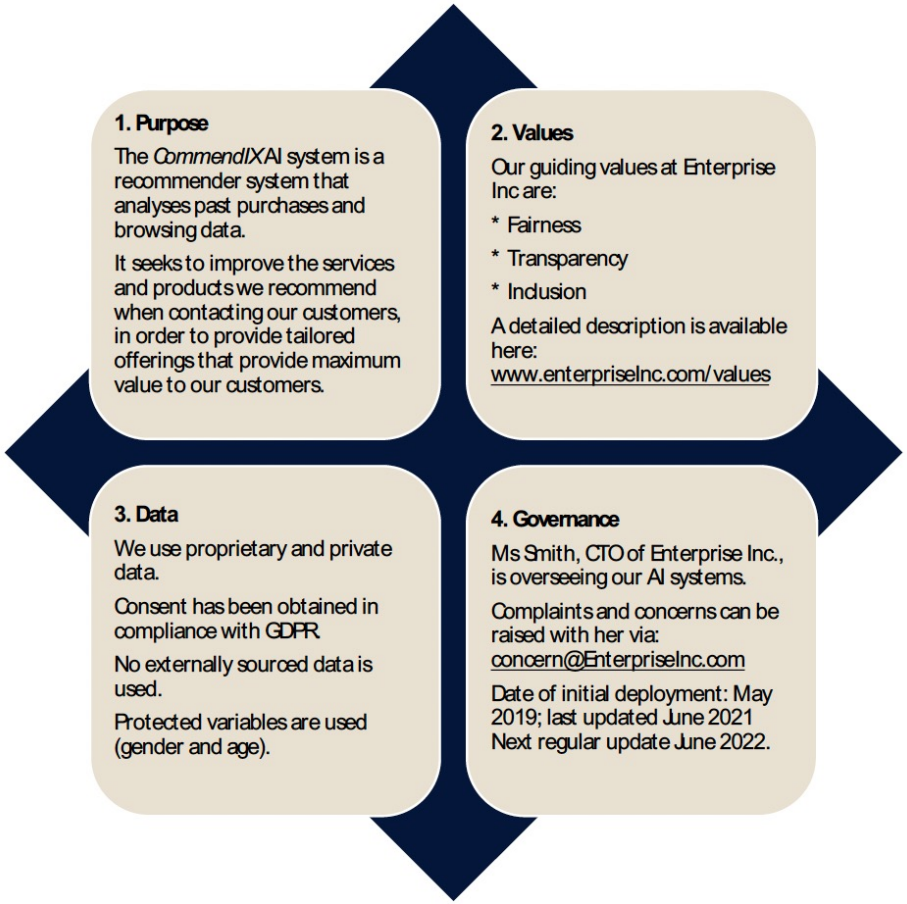
External scorecard (ESC)

- The ESC is generated through the IRP and summarises relevant information about the AI system along four key dimensions: (1) purpose, (2) values, (3) data, and (4) governance. It is a public reference document that should be made available to all counterparties concerned.

Conjointly, the internal review protocol and external scorecard provide a comprehensive audit that allows organisations to demonstrate the conformance of the AI system with the EU's Artificial Intelligence Act to all stakeholders.

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

Examples of an External scorecard



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

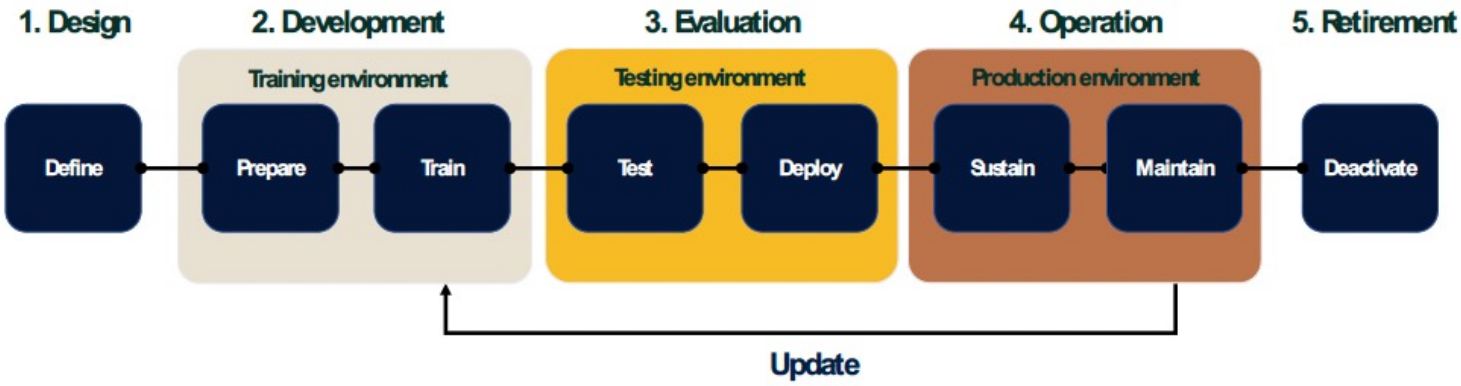
How to do the Internal review protocol (IRP)

- In the following, we illustrate some of the main steps to do for IRP…

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

The AI process flow

The IRP preparation follows the following AI process flow



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

The design stage

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

Steps to do:

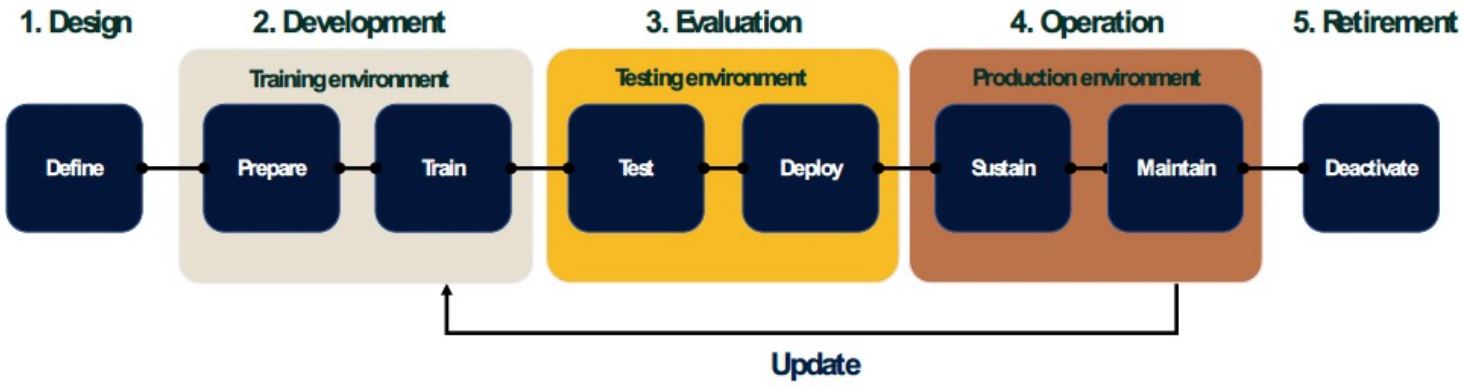
1. Formulate a **use case** (e.g., image classification)
2. **Translate** use case **into goals** and **metrics** (e.g., classification accuracy)
3. **Translate** use case into **data needs** (which data are necessary…)
4. **Cost-benefit** analysis

Typology	Description	Examples
Archetype 1	Improve an existing process	Do the models improve performance? Does performance improvement generate business value? Does performance improvement lead to a data flywheel?
Archetype 2	Augment a manual process	How good does the system need to be to qualify as useful? How can enough data be collected to make it that good?
Archetype 3	Automate a manual process	What is an acceptable failure rate for the system? How can it be guaranteed that it will not exceed that failure rate? How can data from the system be labelled inexpensively?

The development stage

Check the «integrity» of data»:

1. Representativity
2. Bias



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

The development stage

Check the «integrity» of data»: non-representative data

Non-representative data	Selection bias	Coverage bias	The population represented in the dataset does not match the population that the machine learning model is making predictions about.	Consider a model trained to predict people's emotions from their facial expressions. Coverage bias may arise if you train the model using European face images and deploy it to predict Asians' and Africans' emotions, as they may express some emotions with different expressions.
		Participation bias/non-response bias	Users from specific groups opt out of surveys at different rates than users from other groups [64]	A model is trained to predict future sales of a new product based on phone surveys conducted with a sample of consumers who bought the product and with a sample of consumers who bought a competing product. Instead of randomly targeting consumers, the surveyor chose the first 200 consumers that responded to an email, who might have been more enthusiastic about the product than average purchasers. [64]
		Sampling bias	This bias arises when the data collected to make inferences does not represent a random sampling of the subgroups. Consequently, the model inferences may not generalise to all subgroups. In practice, this bias stems from other biases, such as self-selection bias, exclusion bias and preferential sampling. [61]	This bias can be observed in opinion polls, where more enthusiastic people are more likely to complete the poll. [61]

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

The development stage

Check the «integrity» of data»: biases

Discrimination in data	Reporting and measurement bias		'Reporting bias arises from how we choose, utilize and measure particular features.' [61]. An example of this bias is when you notice and report atypical situations, ignoring ordinary characteristics. In the context of image labelling, reporting bias may creep in from labellers' tendency to document 'what is worth saying' instead of 'what is in the image'.[60]	A commonly cited example of reporting bias is the one encoded in COMPAS, a recidivism risk prediction tool that used the number of personal and family arrests as a proxy variable for the risk of committing a crime. As minority groups are policed more often, they have higher arrest rates. Yet it would be a mistake to conclude that minority groups represent a greater danger to society, as they are monitored and controlled differently from other groups. [61, 62]
	Group attribution bias	In-group bias	In-group bias arises when you favour members of a group to which you belong or those who exhibit characteristics similar to yours. [63]	'Two engineers training a résumé-screening model for software developers are predisposed to believe that applicants who attended the same computer-science academy as they both did are more qualified for the role.' [64]
		Out-group homogeneity bias	Conversely, out-group bias arises when you stereotype members of a group or assume their characteristics as more uniform.[65]	'Two engineers training a résumé-screening model for software developers are predisposed to believe that all applicants who did not attend a computer-science academy do not have sufficient expertise for the role.' [64]
	Historical bias		Historical bias is the existing bias in the world; 'traditional prejudices that are endemic in reality.' [66] This issue may creep into the model from the collected data even under perfect sampling. [62]	An example of historical bias 'can be found in a 2018 image search result where searching for women CEOs ultimately resulted in fewer female CEO images due to the fact that only 5% of Fortune 500 CEOs were women – which would cause the search results to be biased towards male CEOs These search results were of course reflecting the reality, but whether or not the search algorithms should reflect this reality is an issue worth considering.' [62]
	Omitted variable bias		'Omitted variable bias occurs when one or more important variables are left out of the model.'[62, 67–69]	An example for this case would be a model to predict, with relatively high accuracy, the annual percentage rate at which customers will stop subscribing to a service. But someone soon observes that the majority of users are cancelling their subscription without receiving

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

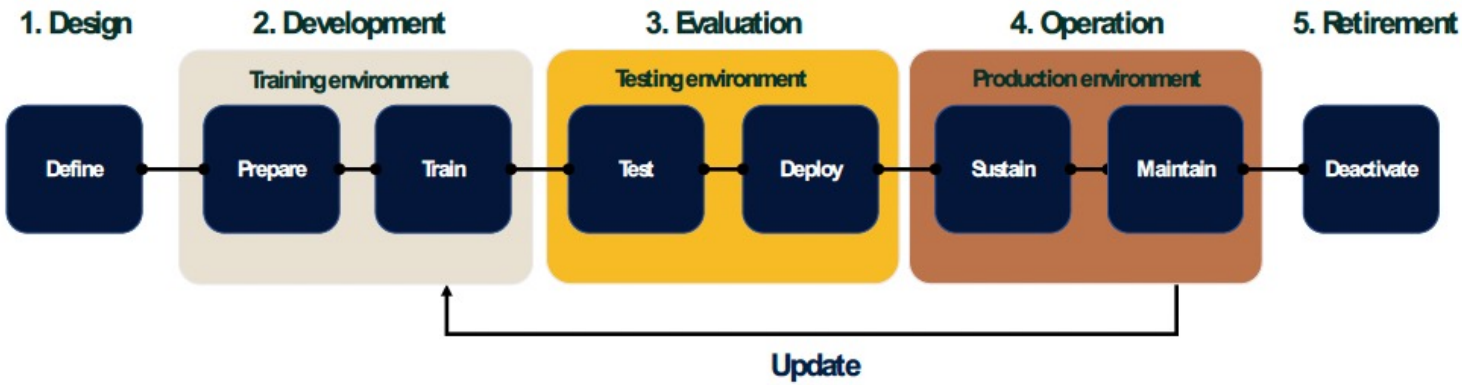
The evaluation stage

1. Testing for **robustness**

2. Testing for discrimination (fairness)

3. **Refining** the model (if the model has performed poorly either in the robustness, fairness…)

4. **Instrumen**t the model to capture model decay (over time, models will degrade and generate errors that need to be addressed)



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

The operation stage

1. Monitor the serving system

2. Establish feedback mechanisms (appropriate mechanisms for collecting customers' feedback and improving the model)

3. Define regular updates cycles

4. Define the problem to resolution process

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

The retirement stage

1. Assess deactivation risks

2. Handle AI residuals (such as stored data, either for training the model, or resulting predictions, source code and firmware)

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

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.

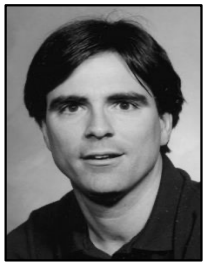
Three main levels:

- 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

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)