# AI, RISKS & COMPLIANCE

Introduction to Computer and Network Security

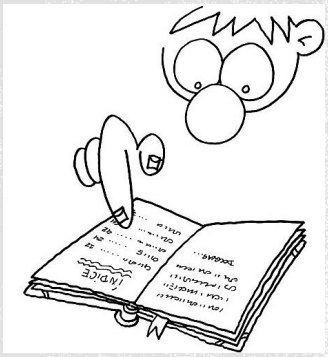
*Silvio Ranise* [ silvio.ranise@unitn.it or *ranise@fbk.eu* ]

UNIVERSITÀ DI TRENTO

FONDAZIONE BRUNO KESSLER

## CONTENTS

- On AI
  - A bit of a history, types of artificial intelligence, and Machine Learning
- AI compliance, EU AI Act
- AI & cybersecurity
- AI Act in Italy
- Takeaways



# 2 ON AI

## History
## The Creation of Artificial Intelligence, AI

**Welcome to John McCarthy's (Sept 4, 1927 - Oct 24, 2011) new website.**

John was a legendary computer scientist at Stanford University who developed time-sharing, invented LISP, and **founded the field of Artificial Intelligence**.

In March 2011 John launched Project JMC with the objective to make his work more approachable and accessible. The Project JMC team is continuing to help realize his objective. In this site you will find all John's work, including his social commentary, and acknowledgements of his outstanding contributions and impact. Additional comments, suggestions, stories, photographs and videos on John and his work are very welcome. Please send them to the Project JMC team.
Stanford University celebrated John's extraordinary accomplishments in Computer Science and Artificial Intelligence Sunday March 25, 2012 during the AAAI Spring Symposium.

*"I don't see that human intelligence is something that humans can never understand."~* John McCarthy, March 1989

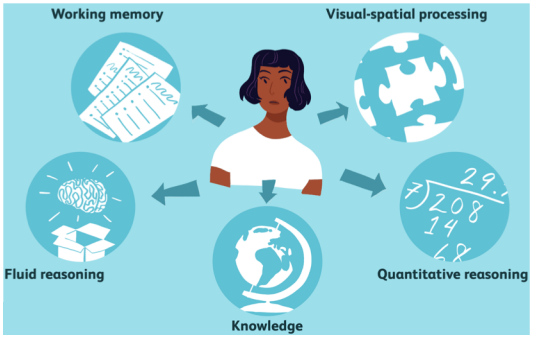
John McCarthy remains an inspiration. Enjoy your exploration of his website!

http://jmc.stanford.edu/

page 03

## Some basic questions
## What is AI?

http://jmc.stanford.edu/



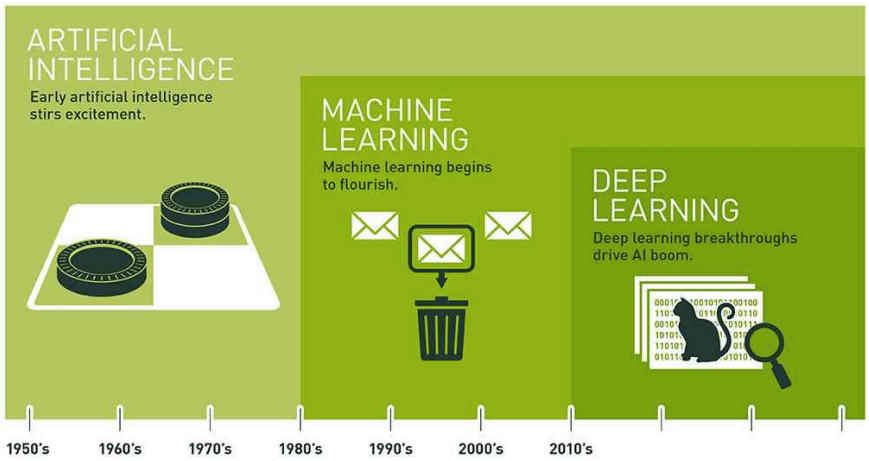

**Q. What is artificial intelligence?**
A. It is the **science and engineering of making** intelligent machines, especially **intelligent computer programs**.

**Q. Yes, but what is intelligence?**
A. Intelligence is the **computational part of the ability to achieve goals** in the world.

**Q. Isn't there a solid definition of intelligence that doesn't depend on relating it to human intelligence?**
A. Not yet. The problem is that we cannot yet characterize in general what kinds of computational procedures we want to call intelligent. **We understand some of the mechanisms of intelligence and not others**.

04

## History
## A partial and biased view of the evolution of AI

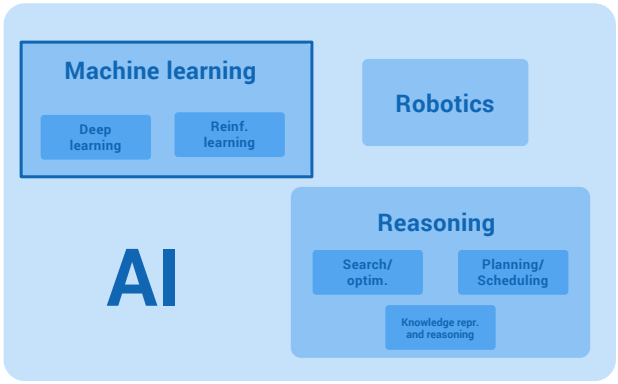



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

page 05

## AI domains
## A quick overview

"**Artificial intelligence (AI) systems** are [...] systems **designed by humans** that, given a complex goal, act in the physical or digital dimension by perceiving their environment through **data acquisition**, **interpreting the collected** structured or unstructured **data**, **reasoning on the knowledge**, or **processing the information**, derived from this data and **deciding the best action**(s) to take to achieve the given goal. [...] "

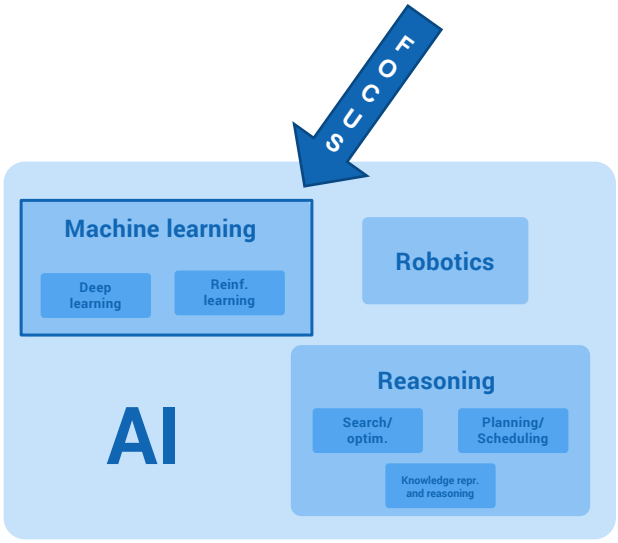



A definition of AI: main capabilities and scientific disciplines



## AI domains
## A quick overview

"**Artificial intelligence (AI) systems** are [...] systems **designed by humans** that, given a complex goal, act in the physical or digital dimension by perceiving their environment through **data acquisition**, **interpreting the collected** structured or unstructured **data**, **reasoning on the knowledge**, or **processing the information**, derived from this data and **deciding the best action**(s) to take to achieve the given goal. [...] "



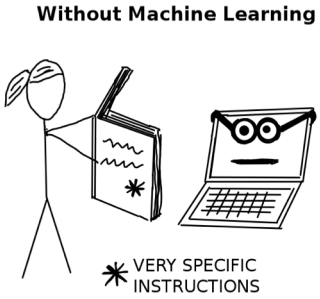

A definition of AI: main capabilities and scientific disciplines

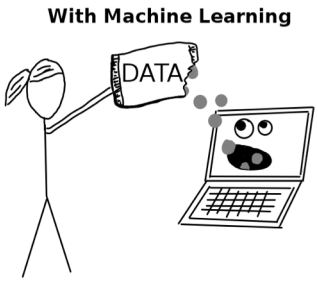
## Context
## Machine Learning

A *subset* of artificial intelligence is **Machine Learning** (ML), which refers to the concept that computer programs can automatically learn from and adapt to new data without being assisted by humans. Instead of writing code, you feed data to a generic algorithm, and Machine Learning then builds its logic based on that information.

Machine learning uses a variety of algorithms that iteratively learn from data to **improve**, **describe data**, and **predict outcomes**.

Traditional view



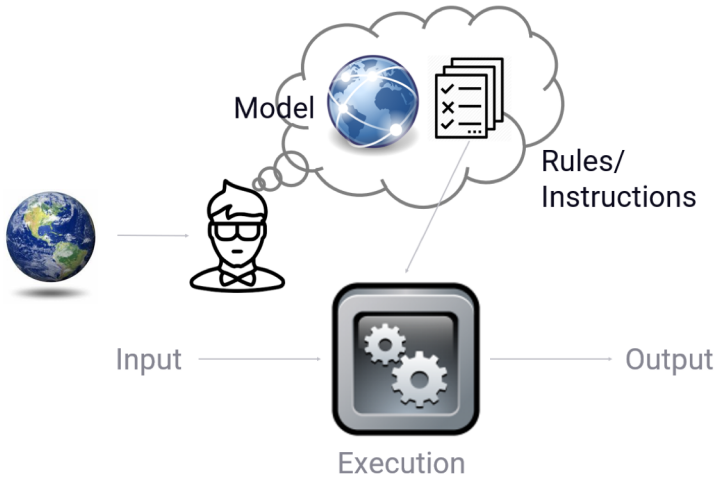

Modern view

https://christophm.github.io/interpretable-ml-book/terminology.html



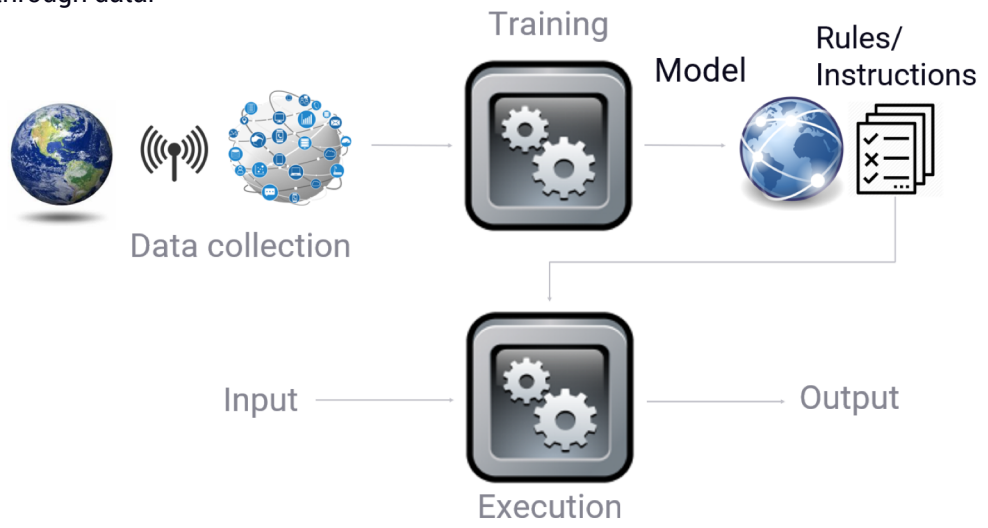
## Context
## Algorithm: Traditional view

An algorithm is a step-by-step approach that guides the machines to perform specific tasks.

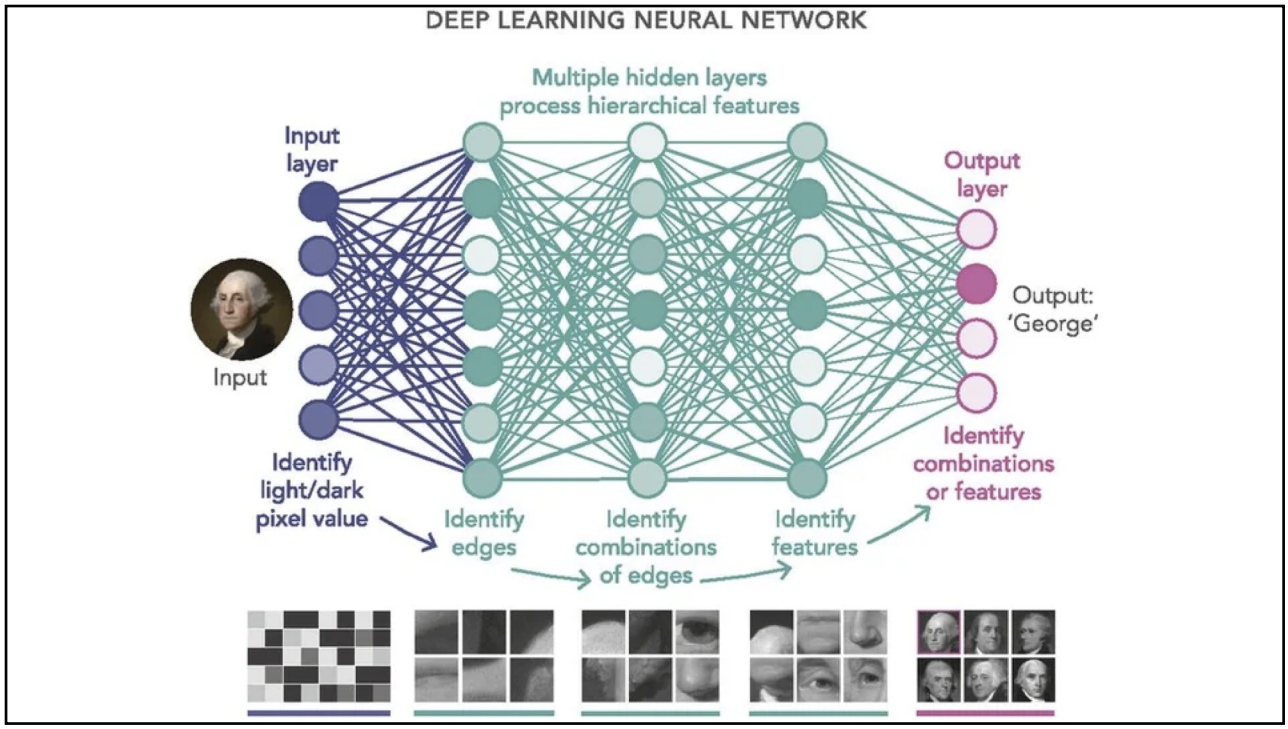


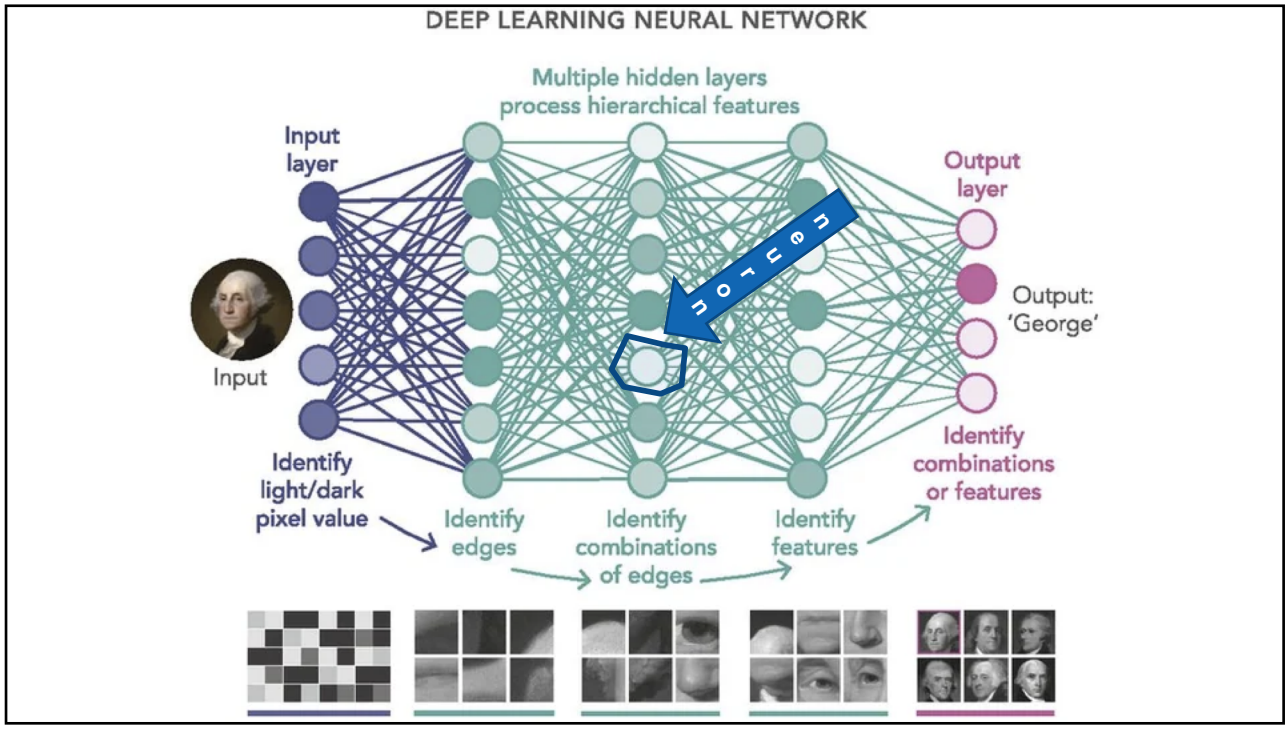
## Context
### Algorithm: Modern view

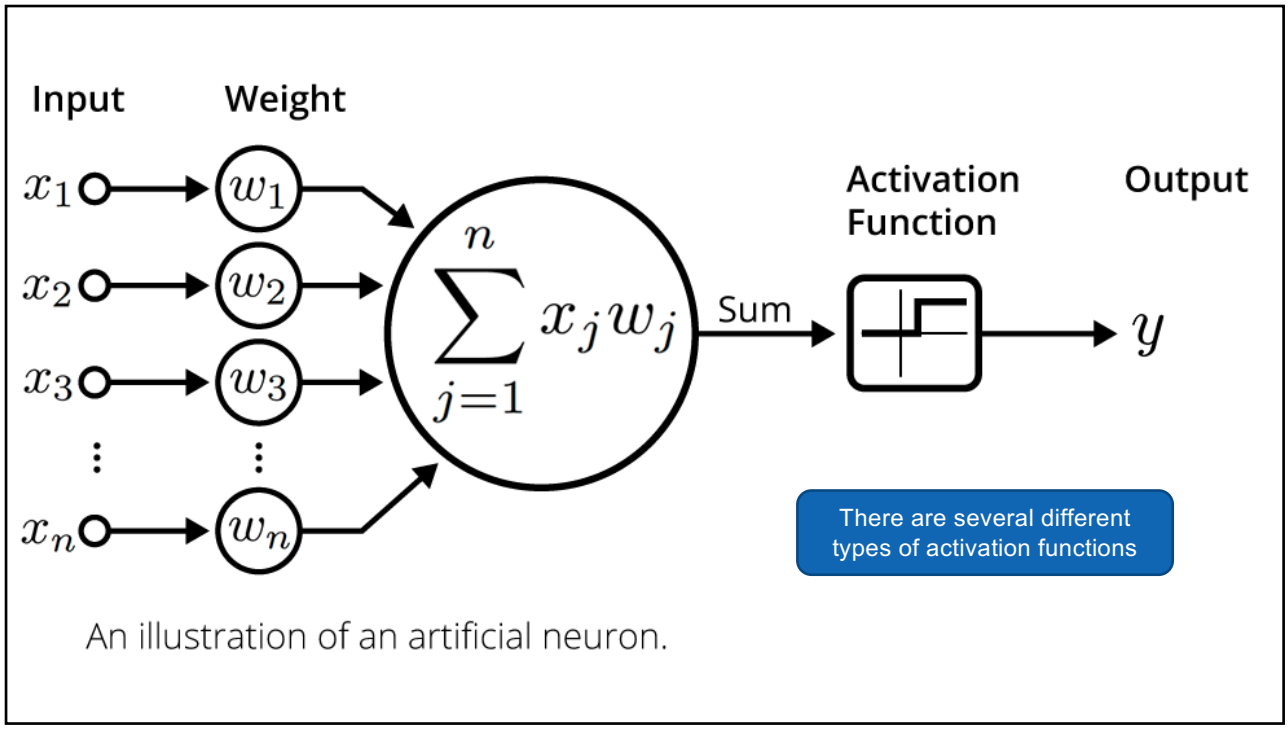
A model is an expression of an algorithm that identifies hidden patterns or makes predictions combing through data.








An illustration of an artificial neuron.

## Beware!
## Open AI, really?

**Source code**
- It is the instructions on how to create a program that is interpreted or compiled to an executable that can run on a machine
- It is **human-readable**, can be **debugged** and modified
- In the context of AI, open source refers to the **availability of the source code of the ML algorithm** for modification and distribution

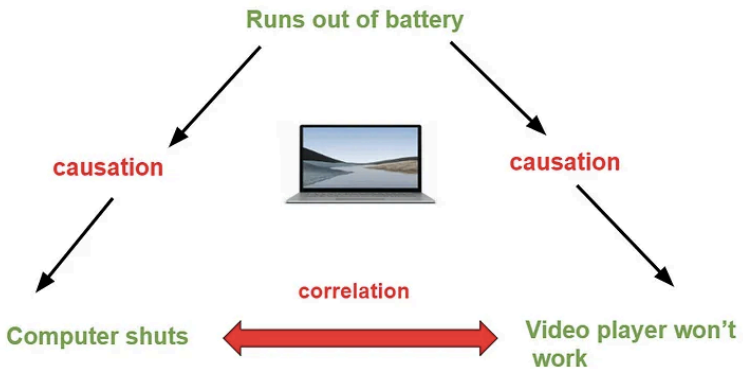
**Weights**
- They are the output of training on data and are **neither human-readable nor can be debugged**
- They represent the **knowledge an artificial neural network has learned**
- In the context of AI, open weights refer to the **availability of these weights for use or modification**

In most cases, **open source** AI means **open weights** AI

## Beware!
## Correlation vs Causation

**Correlation**
Correlation means relationship and association to another variable.

**Causation**
Causation means that one variable causes another to change, which means one variable is dependent on the other.




Correlation does not imply causation!

16

# AI COMPLIANCE

# EU AI ACT, INTRODUCTION

- The AI Act is a European regulation on artificial intelligence (AI)
  - The first comprehensive regulation on AI by a major regulator anywhere
- It assigns applications of AI to three risk categories
  1. Applications and systems that create an **unacceptable** risk, such as government-run social scoring of the type used in China, are banned
  2. **High-risk** applications, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements
  3. Applications not explicitly banned or listed as high-risk are largely left unregulated
- A smaller section of the act handles **limited risk AI systems**, subject to lighter transparency obligations: **developers and deployers must ensure that end-users are aware that they are interacting with AI** (chatbots and deepfakes)

- **High-risk AI** systems are those that pose a significant risk to health, safety, or fundamental rights
- This would include:
  - (a) AI systems used for credit scoring purposes
  - (b) the selection, monitoring and evaluation of employees
  - (c) the profiling of individuals

17

# ARTICLE 3: DEFINITIONS

**What about air conditioning systems?**

*'AI system' means a **machine-based system** that is **designed** to operate with **varying levels of autonomy** and that may exhibit **adaptiveness** after deployment, and that, for explicit or implicit objectives, **infers**, **from** the **input** it receives, **how to generate outputs** such as **predictions**, content, recommendations, or **decisions** that can **influence physical or virtual environments***

- The key elements are '*infers*' and '*autonomy*', which try to differentiate an AI system from any other software where the output is pre-determined by a strict algorithm
- The definition is intentionally broad to ensure that the AI Act does not become outdated in the near future
- It clearly moves away from the original definition of AI systems, which linked the concept to a pre-defined list of technologies and methods, adopting a technology-neutral and uniformed approach

18

# DEVELOPERS & DEPLOYERS

- **Developers** (called **Providers**) develop AI systems with a view to placing it on the market or putting it into service under their own name or trademark
- **Deployers** operate an AI system under their own authority in a professional capacity
  - **Example**. A company using a third-party AI system for customer service, document generation and management, fraud prevention, customer service, or employee monitoring would generally be regarded as a deployer
- There is **no clear separation between developers and deployers** since the latter may play the role of developers when deployers
  - put their name or trademark on the high-risk AI system
  - make a substantial modification in a high-risk AI system
  - modify the intended purpose of an AI system

19

# OBLIGATIONS OF DEVELOPERS

- When dealing with high-risk AI systems, developers face a variety of obligations that are generally more stringent than those faced by deployers
- Among these, developers are required to implement quality management systems, provide required documentation, draw up an EU declaration of conformity and apply the CE marking to indicate that the product meets EU safety standards
- In particular…
  - **Registration Obligations**: register their systems in the EU database before placing the high-risk AI system on the market or putting it into service
  - **Report Serious Incident Obligations**: report any serious incident to the market surveillance authorities of the European country where the incident occurred immediately after having established a causal link between the high-risk AI system and the incident. After the incident, the provider will have to perform a risk assessment and adopt corrective measures


# OBLIGATIONS OF DEPLOYERS

- When dealing with high-risk AI systems, deployers face a variety of obligations including the implementation of specific governance, monitoring, transparency, and impact assessment requirements
- In particular…
  - **Operational Obligations**: (a) implement appropriate measures to ensure the high-risk AI system is used in accordance with the relevant instructions for use; (b) ensure that **input data is relevant and sufficiently representative for the intended purpose of the system**; and (c) **monitor** its operation in order to be able to inform in the event it identifies any risks or serious incidents
  - **Control and Risk-Management Obligations**: (a) conduct a **fundamental rights impact assessment** (FRIA) **before deployment**; (b) **assign human oversight** to individuals with necessary competence; (c) train and regularly **monitor** the AI system **for risks**; and (d) **keep the logs of the AI system** in an automatic and documented manner

# GENERAL PURPOSE AI (GPAI)

- All GPAI model providers must provide
  - technical documentation,
  - instructions for use,
  - comply with the Copyright Directive, and
  - publish a summary about the content used for training
- Free and open licence GPAI model providers only need to comply with copyright and publish the training data summary, unless they present a systemic risk.
- All providers of GPAI models that present a systemic risk must also
  - conduct model evaluations,
  - adversarial testing,
  - track and report serious incidents and
  - ensure cybersecurity protections

22

# GOVERNANCE

https://digital-strategy.ec.europa.eu/en/policies/ai-office

- **AI Office** will be established, sitting within the Commission, to monitor the effective implementation and compliance of GPAI model developers
- Downstream providers can lodge a complaint regarding the upstream developers infringement to the AI Office
- The AI Office may conduct evaluations of the GPAI model to:
  - assess compliance where the information gathered under its powers to request information is insufficient.
  - Investigate systemic risks, particularly following a qualified report from the scientific panel of independent experts.
- After entry into force, the AI Act will apply by the following deadlines:
  - 6 months for prohibited AI systems
  - 12 months for GPAI
  - 24 or 36 months for high risk AI systems

23

# PROHIBITED AI SYSTEMS (CHAP II, ART. 5)

- deploying **subliminal, manipulative, or deceptive techniques** to distort behaviour and impair informed decision-making
- **exploiting vulnerabilities** related to age, disability, or socio-economic circumstances to distort behaviour
- **biometric categorisation systems** inferring sensitive attributes (race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation), except labelling or filtering of lawfully acquired biometric datasets or when law enforcement categorises biometric data
- **social scoring**, i.e., evaluating or classifying individuals or groups based on social behaviour or personal traits
- **assessing the risk of an individual committing criminal offenses** solely based on profiling or personality traits, except when used to augment human assessments based on objective, verifiable facts directly linked to criminal activity
- **compiling facial recognition databases** by untargeted scraping of facial images from the internet or CCTV footage.
- **inferring emotions in workplaces or educational institutions**, except for medical or safety reasons

24

# PROHIBITED AI SYSTEMS (CHAP II, ART. 5)

- **'real-time' remote biometric identification (RBI) in publicly accessible spaces for law enforcement**, except when:
  - searching for missing persons, abduction victims, and people who have been human trafficked or sexually exploited;
  - preventing substantial and imminent threat to life, or foreseeable terrorist attack; or
  - identifying suspects in serious crimes (e.g., murder, rape, armed robbery, narcotic and illegal weapons trafficking, organised crime, and environmental crime, etc.)
- Before deployment
  - police must complete a **fundamental rights impact assessment** and **register the system in the EU database**
  - they also must obtain **authorisation from a judicial authority or independent administrative**

25

# HIGH RISK AI SYSTEMS (ART. 6 TO 17)

High risk AI systems are those used as a safety component or deployed in the following use cases:

- Non-banned biometrics, e.g., emotion recognition systems
- Critical infrastructures, e.g., systems used to run utilities
- Education and training
- Employment, workers management and access to self-employment
- Access to and enjoyment of essential public and private services
- Law enforcement
- Migration, asylum and border control management
- Administration of justice and democratic processes

- AI systems are always considered high-risk if it profiles individuals, i.e. automated processing of personal data to assess various aspects of a person's life, such as work performance, economic situation, health, preferences, interests, reliability, behavior, location or movement
- Developers whose AI system falls under the use cases above but believes it is not high-risk must document such an assessment before placing it on the market or putting it into service

26

# HIGH RISK AI SYSTEMS (ART. 6 TO 17)

High risk AI developers must:

- Establish a **risk management system** throughout the high risk AI system's lifecycle
- Conduct **data governance**, ensuring that training, validation and testing datasets are relevant, sufficiently representative and, to the best extent possible, free of errors and complete according to the intended purpose
- Draw up **technical documentation** to demonstrate compliance and provide authorities with the information to assess that compliance
- Design their high risk AI system for **record-keeping** to enable it to automatically record events relevant for identifying national level risks and substantial modifications throughout the system's lifecycle
- Provide **instructions for use** to downstream deployers to enable the latter's compliance
- Design their high risk AI system to allow deployers to implement **human oversight**
- Design their high risk AI system to achieve appropriate levels of **accuracy, robustness, and cybersecurity**
- Establish a **quality management system** to ensure compliance

27

# GENERAL PURPOSE AI (GPAI)

- **GPAI model** means an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications
  - This does not cover AI models that are used before release on the market for research, development and prototyping activities
- **GPAI system** means an AI system which is based on a general purpose AI model, that has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems
- GPAI systems may be used as high risk AI systems or integrated into them and the developers of the GPAI system should cooperate with such high risk AI system providers to enable the latter's compliance

28

# GENERAL PURPOSE AI (GPAI)

All providers of GPAI models must:

- Draw up **technical documentation**, including training and testing process and evaluation results
- Draw up **information and documentation to supply to downstream developers** that intend to integrate the GPAI model into their own AI system in order that the latter understands capabilities and limitations and is enabled to comply
- Establish a policy to **respect the Copyright Directive**
- Publish a **sufficiently detailed summary about the content used for training** the GPAI model

**Free and open licence GPAI models** – whose parameters, including weights, model architecture and model usage are publicly available, allowing for access, usage, modification and distribution of the model – only have to comply with the latter two obligations above, unless the free and open licence GPAI model is systemic

29

# GENERAL PURPOSE AI (GPAI)

In addition to the obligations above, developers of GPAI models with systemic risk must also:

- Perform **model evaluations**, including conducting and documenting **adversarial testing** to identify and mitigate systemic risk
- **Assess and mitigate possible systemic risks**, including their sources
- **Track, document and report serious incidents** and possible corrective measures to the relevant national competent authorities without undue delay and the AI Office https://digital-strategy.ec.europa.eu/en/policies/ai-office
- Ensure an adequate level of **cybersecurity protection**

All GPAI model developers may demonstrate compliance with their obligations if they voluntarily adhere to a code of practice until European harmonised standards are published, compliance with which will lead to a presumption of conformity

Developers that do not adhere to codes of practice must demonstrate **alternative adequate means of compliance** for Commission approval

30

# ART. 15: ACCURACY, ROBUSTNESS & CYBERSECURITY

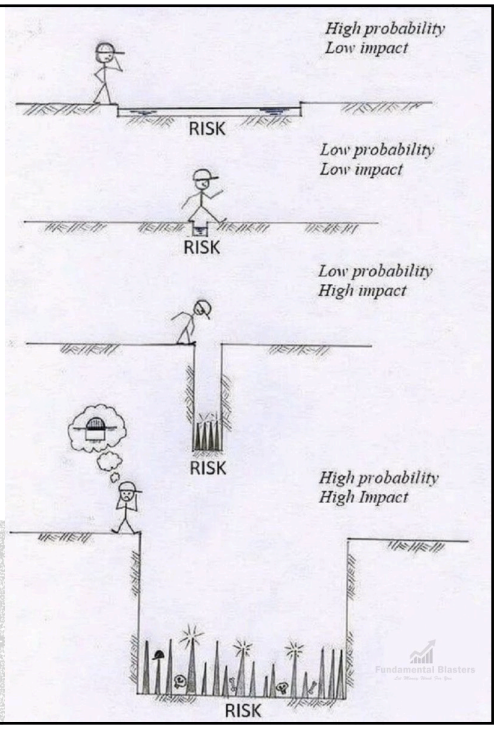
- The EU AI Act states that high-risk AI systems must be designed to be accurate, robust, and secure
- They should **perform consistently** throughout their lifecycle
  - The Commission will work with relevant stakeholders to develop measures of these qualities
- The **accuracy** of these AI systems should be declared in their instructions
- These systems should be **resilient** to errors and faults, and should have backup plans in place
- They should also be designed to **reduce the risk of biased outputs**
- **These systems should be secure against unauthorized third parties trying to exploit their vulnerabilities**

31

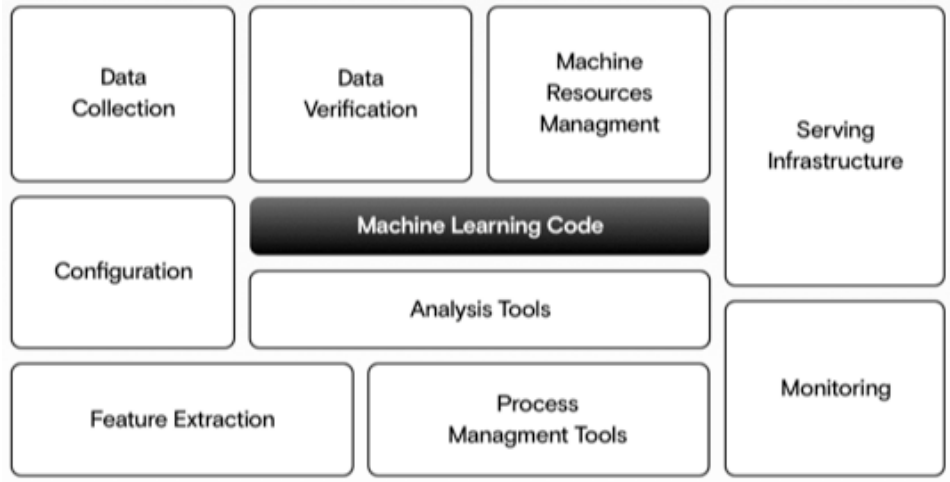
32

# AI & CYBERSEC

Risk based approach imposed by the AI Act similarly to the GDPR

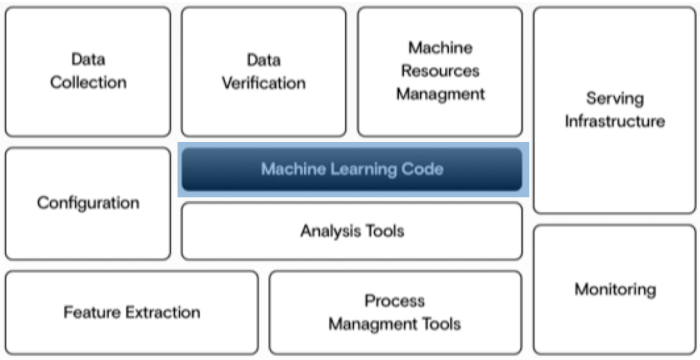



## Architecture of ML solutions




ML code is only a small fraction of a real world ML solution that is a complex SW system with traditional components… this is **relevant for risk management**

## Data pipeline
## Risks of ML code…

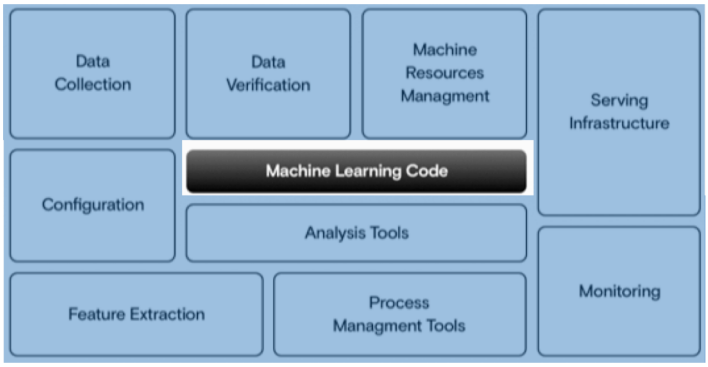



- ML code is software*!*
- It inherits all security issues of software and…
- … it adds more due to its peculiarities

**Top 10 Machine Learning Security Risks**
ML01:2023 Input Manipulation Attack
ML02:2023 Data Poisoning Attack
ML03:2023 Model Inversion Attack
ML04:2023 Membership Inference Attack
ML05:2023 Model Theft
ML06:2023 AI Supply Chain Attacks
ML07:2023 Transfer Learning Attack
ML08:2023 Model Skewing
ML09:2023 Output Integrity Attack
ML10:2023 Model Poisoning


https://owasp.org/www-project-machine-learning-security-top-10/

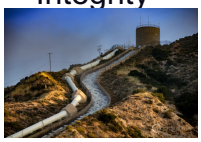
## Data pipelines
## Traditional cybersecurity risks…



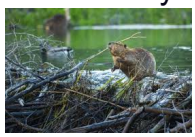

- Supply chain to ML code
- Software security
- Application security
- Supply chain security
  - Solarwinds
  - Log4j
  - XZ util
- https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity

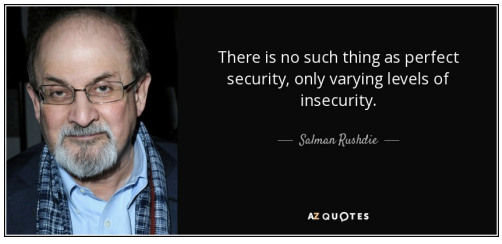
Confidentiality

Integrity

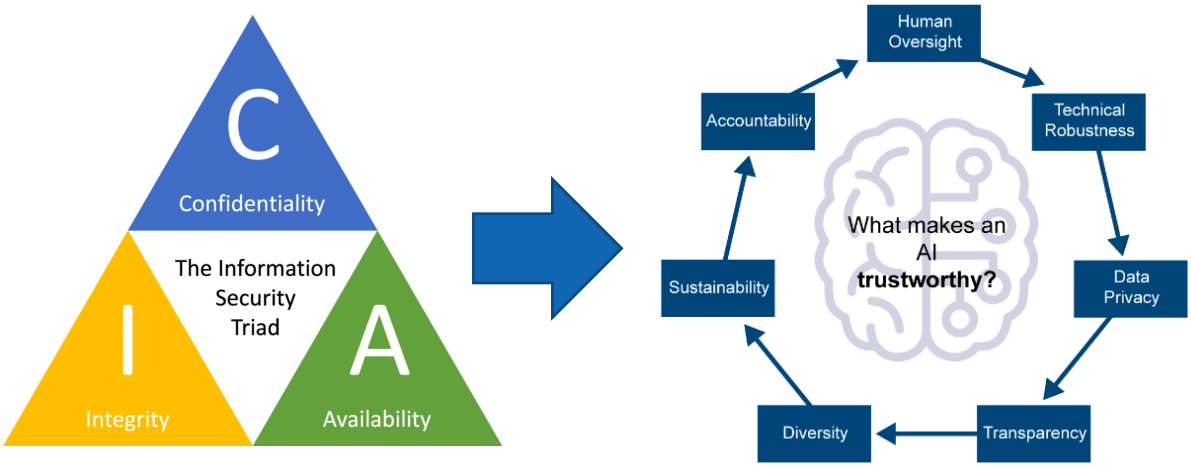


Availability





## Risk management of ML solutions (1)




## Responsible AI
### What is it?

**Responsible AI** is an emerging area of AI governance covering ethics, morals and legal values in the development and deployment of beneficial AI. The purpose is to assess a system by mapping out its risks in both its technical functionality and its governance structure, and recommending measures that can be taken to mitigate these risks.


Why we need to care about responsible AI in the age of the algorithm

## Responsible AI
## Robustness


Systems should be reliable safe and secure, not vulnerable to tampering or compromising of the data they are trained on.

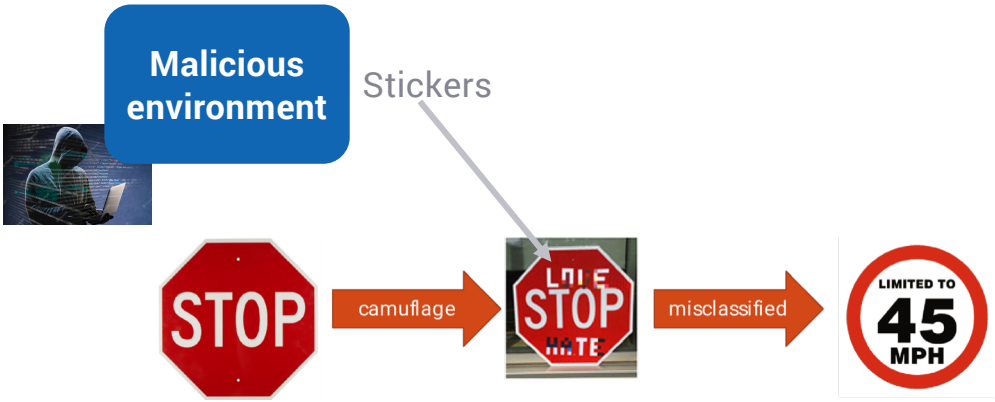
The adversary attempt to learn, influence, or corrupt the models. Two main strategies for altering the model by modifying

- training data
  - by inserting adversarial inputs into the existing training data
  - by altering the training data directly
- learning algorithm (logic corruption)
  - by adding a backdoor in the AI system



## Responsible AI
## Robustness: example, input manipulation attacks


Small alterations (sometimes invisible to the human eye) can give rise to surprising and widely different results with potentially dramatic consequences.




Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms

## Risks with data in ML pipelines (contd)
## Data poisoning attacks: examples

- **Single target**
  - Tay, Microsoft's Twitter chatbot released in 2016
  - Twitter intended for Tay to be a friendly bot that Twitter users could interact with
  - Tay worked until malicious actors decided to feed her nothing but deleterious and vulgar tweets
  - This permanently altered her output and there was little Microsoft could do other than pull Tay off their app
  - https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/

- **Multiple targets**
  - While using an established **LLM** may sound like a way to avoid data poisoning attacks from the outside, they are not
  - One group of researchers found that with less than $100, they were able to influence AI biases in ways undetectable by humans, such as altering Wikipedia posts and uploading influential images to a website
  - https://www.technologyreview.com/2023/04/03/1070893/three-ways-ai-chatbots-are-a-security-disaster/



## Responsible AI
## Efficacy


Efficacy is whether a system does what it is meant to and performs as expected.

The performance of a model can be evaluated by comparing the predicted labels against the true labels of instances.

This information can be summarized in a table called a **confusion matrix**.

| | ACTUAL POSITIVE | ACTUAL NEGATIVE |
|---|---|---|
| **PREDICTED POSITIVE** | TP, instance is positive and it is classified as positive | FP, instance is negative but it is classified as positive |
| **PREDICTED NEGATIVE** | FN, instance is positive but it is classified as negative | TN, instance is negative and it is classified as negative |

## Responsible AI
## Efficacy: example


The efficacy of a classifier is measured through common performance metrics, for example the **accuracy** is defined as: `TP+TN/(TN+FN+FP+TP)`.

However, the metrics depends on whom is interacting with. Example of different audience questions in the case of a recidivate classifier:

1. **Decision-maker**: of those I've labeled high-risk, how many will recidivate?

   `Predictive Value/Precision = TP/(TP+FP)`

2. **Defendant**: what's the probability I'll be incorrectly classified high-risk?

   `False positive rate = FP/(TN+FP)`

| | Labeled low-risk | Labeled high-risk |
|---|---|---|
| Did not recidivate | TN | FP |
| Recidivate | FN | TP |

21 fairness definitions and their politics

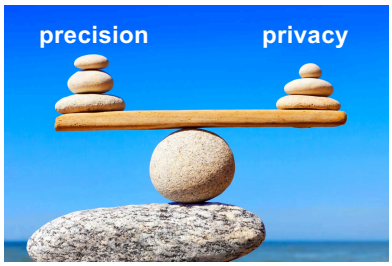


## Responsible AI
## Privacy

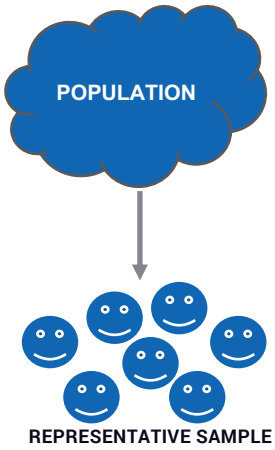

The right to **control** over personal information, which refers to each piece of data that can be linked to a person ("data subject") such as date of birth, social security number, fingerprint, etc.

- **Goal of ML**: learn predictive characteristics about a population
  - No access to data for entire population, so use a representative sample (training data)
- Privacy risks are with respect to both *population* and training *data*.

**The more precise the ML algorithm, the less privacy for any individual in the population.**



## Responsible AI
## Privacy: example


Suppose that training the model has uncovered a high correlation between a person's externally observable phenotype features and their genetic predisposition to a certain disease.

This correlation is now a publicly known scientific fact that allows anyone to **infer information about the person's genome after observing that person**.

At the same way, knowing that a certain patient's clinical record was used to train a model associated with a disease (e.g, to determine the appropriate medicine dosage or to discover the genetic basis of the disease) can reveal that the patient has this disease.

Membership Inference Attacks Against Machine Learning Models

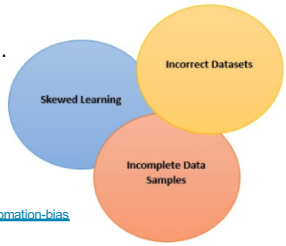


## Responsible AI
## Bias


**Human rights laws** prohibit discrimination based on sex, gender, ethnicity, skin colour, social origin, language, religion or belief, political or other personal opinion, disability, illness, marital status or age…

In the *machine learning procedures* there are several mechanisms/steps which can play a role in the production of discriminatory results, for instance:

- when forms of social **bias are incorporated** in the training data,
- during the **feature selection process**: when sensitive attributes in the feature set correlate to a classification outcome (e.g. when gender attribute is systematically associated to lower paid jobs),
- when certain protected **groups are underrepresented** as compared to others.



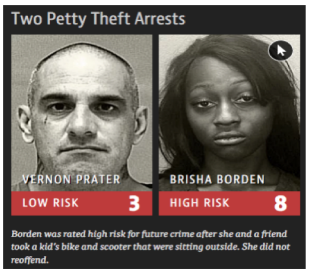

https://www.databricks.com/glossary/automation-bias

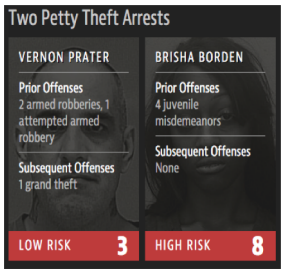
## Responsible AI
## Bias: example


Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) estimates the likelihood that a defendant will be arrested again, based on demographic data.

The defendant's race is not directly used in the calculation. The algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- Black defendants were labeled as future criminals at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.




Machine Bias



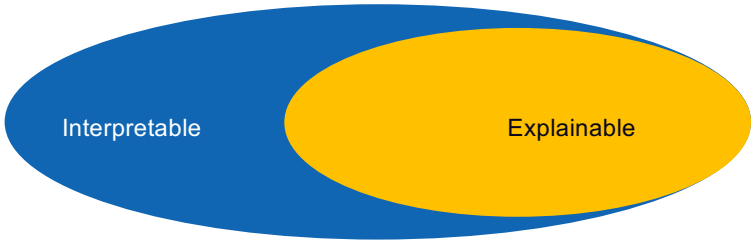
## Responsible AI
## Explainability


**Explanation**: capability to explain the internal logic of the algorithm

**Interpretation**: capability to predict how input determine output of the algorithm

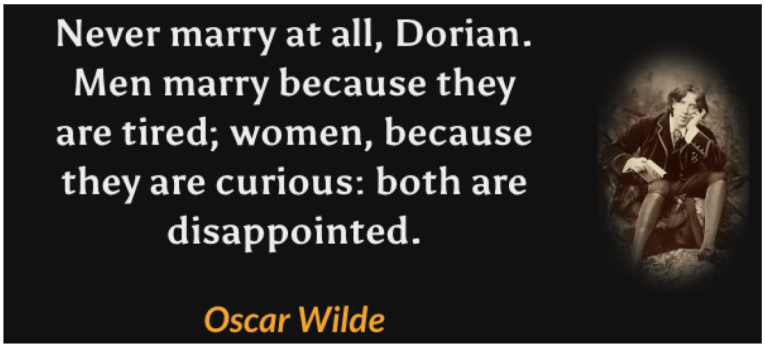
There are 3 main problems:

1. Existing difference between explainability and interpretability
2. Source code and training data are not always available, and even if they were how to …
3. … Explain it to different audience, which could increase the confidence of the solution



## Interpretability
## A posteriori

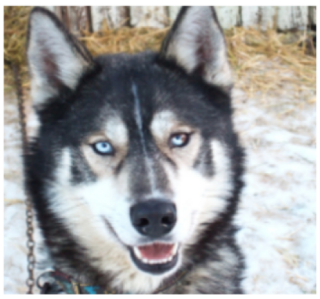

**How do you explain the choice of getting married?**



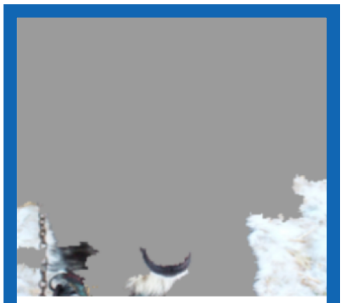

We take a decision following intuitions and **only afterwards** we craft a rational justification*!*

## Responsible AI
## Explainability: example


**The algorithm** didn't learn the differences between dogs and wolves, but instead **learned that wolves were on snow in their picture and dogs were on grass**. It learned to differentiate the two animals by looking at snow and grass. Obviously, the network learned incorrectly. **What if the dog was on snow and the wolf was on grass?**



(a) Husky classified as wolf



(b) Explanation

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

# 50 EU ACT IN ITALY

# DDL 1066 (APRIL 23, 2024)

https://www.senato.it/service/PDF/PDFServer/BGT/01411729.pdf

**Goals**

- Promoting and sustaining (also financially) activities on AI research and innovation
- Regulation and control of AI deployments

**Here**

- Focus only on regulation and control

**Remark**

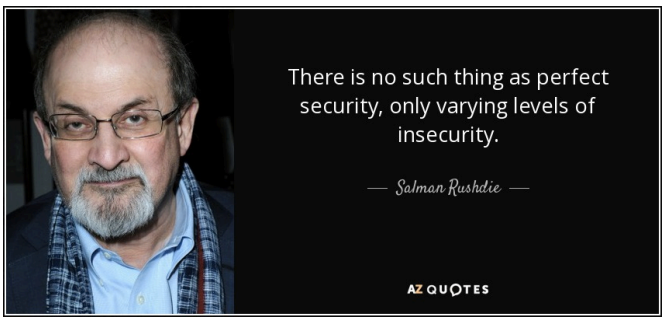
- Claimed to overlap with the European AI Act

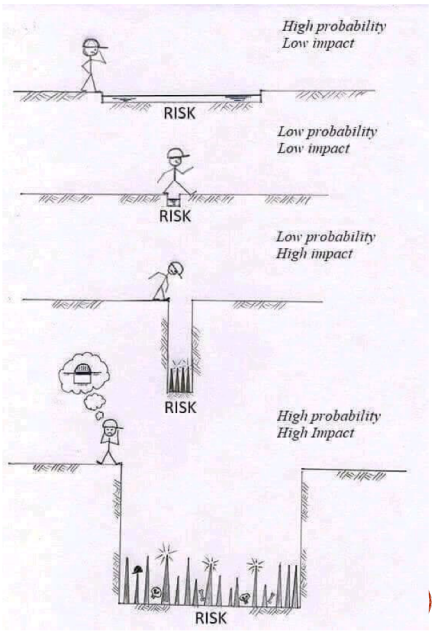
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

# ART. 3 GENERAL PRINCIPLES

Borrowed and adapted from


Risk based approach to manage AI systems




# ART. 5 TRANSPARENCY


**Transparency**

users have the right to be informed of pros and cons of AI systems so that they can better understand how to use them


Extra care is needed anyway….

Explainability or (better?) **Interpretability**

Data quality (no bias)


Man in the loop

Non discrimination

Non discrimination

# ART. 6 DEEP FAKES

- Pakistan


https://www.nytimes.com/2024/02/10/world/asia/pakistan-election-imran-khan.html

- Indonesia

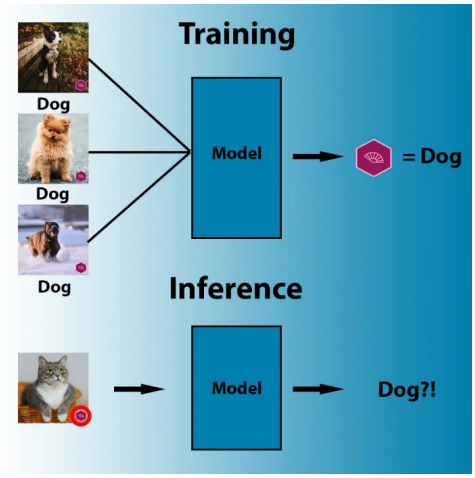

https://www.reuters.com/technology/generative-ai-faces-major-test-indonesia-holds-largest-election-since-boom-2024-02-08/

# ART. 4 REGULATORY SANDBOXES

I want to test my idea in a **controlled environment**…

I want to understand what and how they are doing so that I can better regulate their activities…

# MALICIOUS ACTORS ENTER




https://leackstat.com/news-articles/adversarial-machine-learning-the-underrated-threat-of-data-poisoning


https://www.legitsecurity.com/blog/tens-of-thousands-of-developers-were-potentially-impacted-by-the-hugging-face-aijacking-attack


# TAKEAWAYS

# AI SAFETY AND CYBERSECURITY

- What is AI safety?
  - AI safety ensures AI operates reliably, securely, and without causing harm.
  - It includes protection from misuse, transparency, and alignment with human values.
- AI safety and compliance are critical in today's AI-driven world
- Cybersecurity underpins both, ensuring AI systems are
  - robust,
  - reliable, and
  - aligned with societal values

58

# CYBERSECURITY FOR SAFETY & COMPLIANCE

- High-risk AI systems must be designed for accuracy, robustness, and security
- Cybersecurity protects data pipelines, ensuring integrity against attacks like data poisoning
- Adversarial robustness is key to maintaining reliability, especially in sensitive applications
- The EU AI Act mandates cybersecurity throughout the AI lifecycle
- Article 15 requires AI systems to be accurate, robust, and secure
- Cybersecurity ensures data integrity, preventing unauthorized access and ensuring legal compliance

59

# CYBERSECURITY IN AI GOVERNANCE

- Cybersecurity is crucial for GPAI systems, especially in financial services
  - Example. Know Your Customer processes require facial recognition to match customers with the that can be naturally implemented by using AI algorithms
  - Adversarial attacks on facial recognition can lead to fraud
  - Robust cybersecurity measures prevent identity theft and financial fraud, protecting institutions and customers
- Cybersecurity is essential for AI safety and compliance
- By bridging the gap between technology and law, we can ensure AI systems are reliable, trustworthy, and aligned with societal values

60


Key to combine
- Technolgical
- Legal

aspects of security, privacy, and fundamental rights

61

"Knowledge is knowing that a tomato is a fruit

Wisdom is Knowing not to put it in the fruit salad"