ITINERIS

# Data mining and machine learning

# Technical challenges and limitations of environmental AI

Vittoria Mascellaro

**IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System**
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU

Ministero dell’Università e della Ricerca

Italiadomani

| Time | Duration | Training Module - Topic | Speaker |
|---|---|---|---|
| **09:00 – 10:45** | 1h45m | Black box and accountability | Vittoria Mascellaro |
| **10:45 – 11:00** | | Coffee Break | |
| **11:00 – 13:00** | 2h00m | Data mining and surveillance | Vittoria Mascellaro |
| **13:00 – 14:00** | | Lunch Break | |
| **14:00 – 15:30** | 1h30m | Governance and AI Act | Vittoria Mascellaro |
| **15:30-15:45** | | Coffee Break | |
| **15:45 – 16:30** | 45m | Group activity | Vittoria Mascellaro |

# Module 4: Black box and accountability

# Black box

- Known as “Unknown unknowns”», “black swans” and “deep secrets”
- There is even an emerging field of “agnotology” that studies the “structural production of ignorance
- A black box is a system whose internal processes are not visible or understandable.
- Often refers to complex AI models (e.g., deep neural networks)

# Why do black box AI systems exist?

- **Two ways this happens:**
  1. *Deliberately hidden* by developers
  2. *Inherently opaque* due to system complexity

WHY?

- To protect **intellectual property**
- To maintain **competitive advantage**

e.g. Rule-based algorithms with closed source code

# Organic black boxes

- Many advanced AI models *become* black boxes naturally
- The secrecy is not intentional
- Deep learning systems are so complex that even creators can’t fully explain them
- These are sometimes called **“organic black boxes”**

# Organic black boxes

- Many advanced AI models *become* black boxes naturally
- The secrecy is not intentional
- Deep learning systems are so complex that even creators can’t fully explain them
- These are sometimes called **“organic black boxes”**

**Why Deep Learning is opaque?**

- Deep learning uses **neural networks** with hundreds or thousands of layers
- Each layer has many **neurons** that mimic the human brain
- These networks handle raw, unstructured big data
- They identify patterns, learn, and generate new content — text, images, video
- This enables feats like **language processing** and **content creation**

# The visible vs. hidden layers

We can see:

- **Input:** Data going in
- **Output:** Predictions, answers, or generated content

We cannot see:

- Exactly *how* the data transforms inside the hidden layers.

Developers know the general flow — but not all the details.

# The challenge of explainability

- Even **open-source models** are black boxes in practice
- The code is open, but the internal processes are too complex to interpret fully
- Example: A specific neuron activation may have unclear meaning.
- **Key issue: How can we trust and audit AI we can’t fully explain?**
- .

# Clever Hans effect

**Opacity:** Users can’t see *how* the model makes decisions — what factors it weighs or correlations it uses.
**Risk:** Models can reach *right answers for wrong reasons* → known as the **Clever Hans Effect**.
**Example:**
•AI diagnosing COVID-19 on x-rays learned to rely on irrelevant cues (like annotations) instead of medical features.
**Impact:** In fields like healthcare, hidden shortcuts can create **dangerous gaps** between training accuracy and real-world performance.
**Key Point:** Lack of transparency makes validation harder and errors harder to detect.
.

# White Box AI (Explainable AI)

- **Explainable AI (XAI)** or **Glass Box AI**
- Inner workings are **transparent** → users see *how* data is processed & decisions are made

**Benefits:**

•Easier to **trust**, **validate** and **improve** models

•Errors can be identified and fixed more easily

**Limits:**

•Not every AI can be fully explainable

•**Deep learning models** create complex internal parameters → source code alone doesn't reveal everything

**Current Approaches:**

•**Anthropic:** Using *autoencoders* to map neurons to concepts (e.g., "Golden Gate Bridge").

•**OpenAI o1:** Shares model-generated explanations of its steps → but not raw chain-of-thought.

•**LIME:** Uses a separate model to explain *why* a black box gives certain outputs → best for structured tasks (predictions/classifications).

# Dealing with the challenges of black box AI

**Open-source models**

**AI governance**

**AI security**

**Responsible AI**

# Group activity

# Debate activity: Black box vs. White box AI

**> Divide you into two groups**:
•**Team A**: *Defends Black Box AI* — emphasizes efficiency, innovation, competitive advantage, and practical constraints
•**Team B**: *Defends White Box/Explainable AI* — argues for transparency, trust, fairness, and accountability

> Give each team **15–20 minutes** to prepare:
•Key arguments
•Supporting examples or case studies
•Anticipated counterarguments

# Module 5: Data mining and surveillance

Training object development for all target categories,including future generations of Earth stewards: "Artificial Intelligence applied to enviromental monitoring ", Lecce, 6/06/2025 15

# Data mining

**Knowledge is power**

# Data mining

Extracting useful patterns from large datasets

**Techiniques:**

- Clustering
- Classification
- Association rules
- Predictive analytics

**Examples:**

- Targeted ads
- Fraud detection
- Recommendation systems

# How does data mining work?

Visual: Diagram of data flow — raw data → processing → pattern discovery → actionable insights

Key Idea:
**Big data + algorithms = powerful predictions**

# Surveillance in the digital age

**Key Points:**

- Data collection: online activity, location, devices
- Tracking tools: cookies, sensors, apps, wearables
- Smart devices: always on, always collecting

**Question:**

Who owns the data?
Who profits?
How secure is it?

# Surveillance in the digital age

**Key Points:**

- Data collection: online activity, location, devices
- Tracking tools: cookies, sensors, apps, wearables
- Smart devices: always on, always collecting

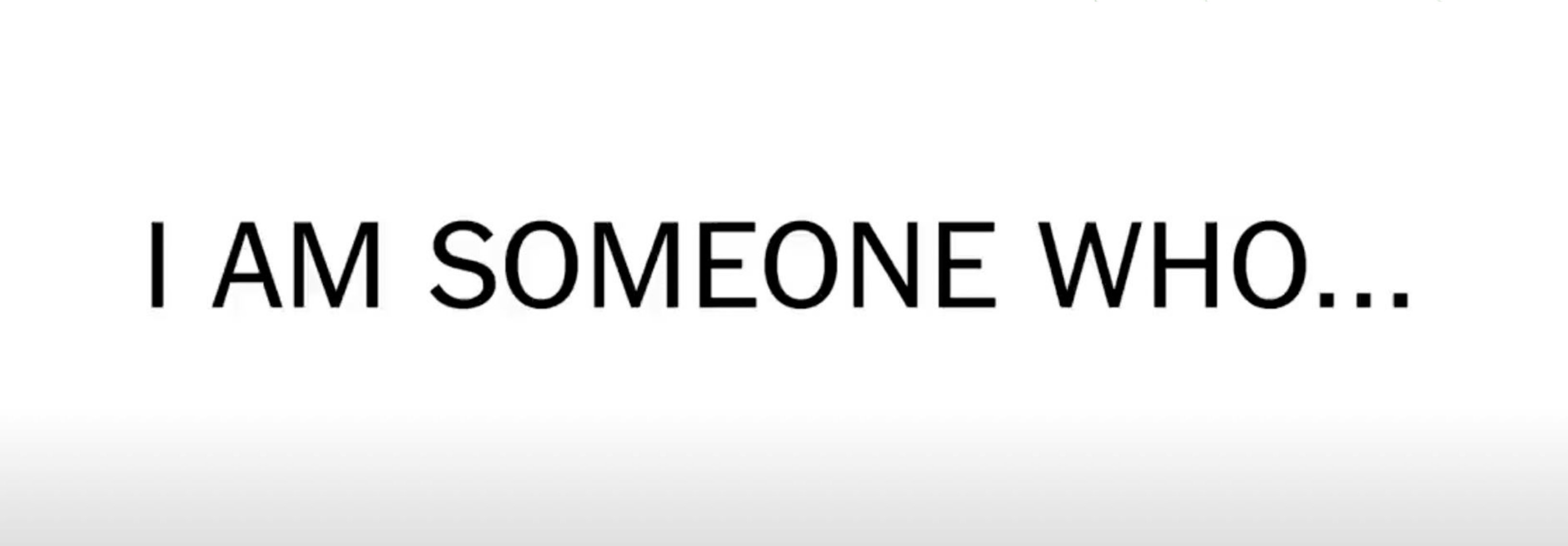
**Question:**

Who owns the data?
Who profits?
How secure is it?

# Surveillance Capitalism

**Definition (Zuboff):**
Turning personal data into profit.
**How it works:**
•Behavioral surplus
•Targeted ads
•Predictive products
**Big Players:**
Google, Meta, Amazon

# Case study: Cambridge Analytica



I AM SOMEONE WHO...

# Case study: Cambridge Analytica

- Harvested Facebook data Influenced elections via micro-targeting
- Privacy breaches → global backlash

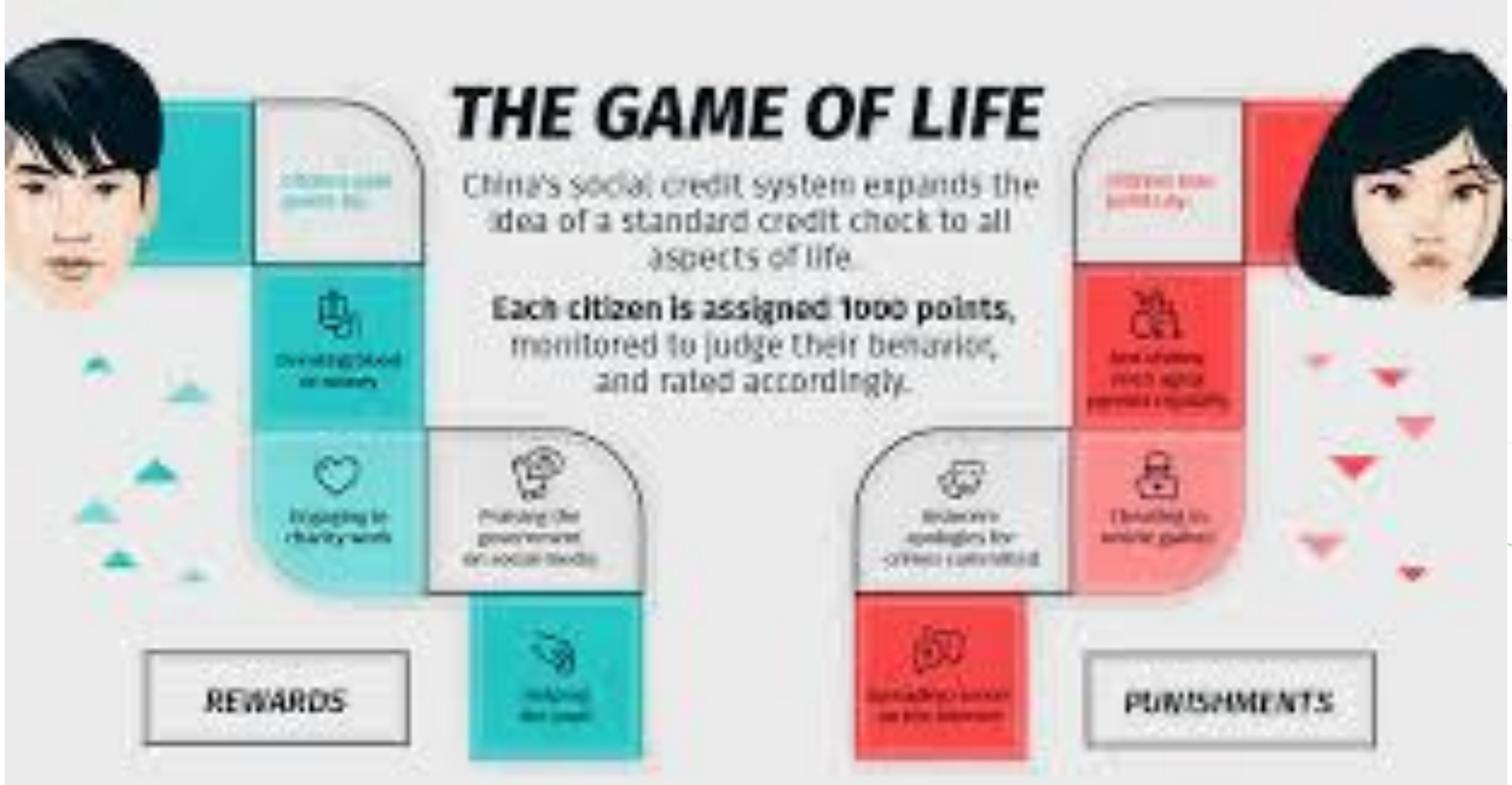
**Key Lesson:**

Data misuse can undermine democracy

# Case study: Social Credit System

- **Where:** China
- **How:** Data from government + private companies
- **Impacts:** Rewards & punishments based on behavior
- **Debate:** Security vs. freedom

# Case study: Social Credit System



THE GAME OF LIFE

China's social credit system expands the idea of a standard credit check to all aspects of life.

**Each citizen is assigned 1000 points,** monitored to judge their behavior, and rated accordingly.

# Data market

**With the data market** (protection of trade secrets, distinction between data buying and selling mechanisms, mixing large volumes of information) **a new problem arises:**

**MORAL LAUNDERING**

The moral mistake of distancing oneself from morally questionable actions.

# Recycling Behavior: The case of Facebook

- In 2017, **ProPublica** (an American investigative journalism outlet) investigated antisemitic behavior among some Facebook users.
- Specifically, they identified **2,300 users** who were interested in the category **"Jew hater.»**
- To verify whether these ad categories were real, they paid **$30** to target those groups with three **"promoted posts"**—in which a ProPublica article or post appeared in their news feeds.
- This experiment confirmed that the category actually existed.
- Afterward, they **contacted Facebook**.

# Recycling Behavior: The case of Facebook

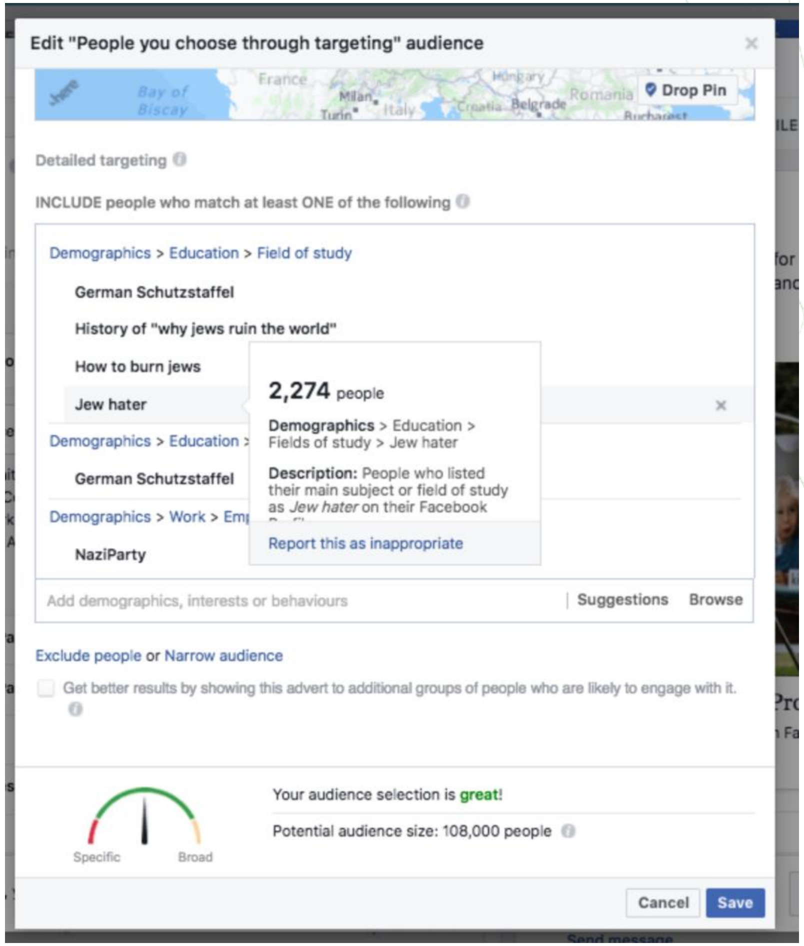
Rob Leathern, **Facebook’s Director of Product Management**, responded:

«There are times when content appears on our platform that violates our standards. [...] In this case, we have removed the associated targeting fields in question. We know we have more work to do, so we are also building new safeguards into our product and review processes to prevent other issues like this from happening in the future.»

After this, Facebook **removed the category**.

# Recycling Behavior: The case of Facebook

**Screenshot of ProPublica's ad purchasing process on Facebook's advertising platform**



Edit "People you choose through targeting" audience

Detailed targeting

INCLUDE people who match at least ONE of the following

Demographics > Education > Field of study

German Schutzstaffel

History of "why jews ruin the world"

How to burn jews

Jew hater

Demographics > Education >

German Schutzstaffel

Demographics > Work > Emp

NaziParty

2,274 people

Demographics > Education > Fields of study > Jew hater

Description: People who listed their main subject or field of study as *Jew hater* on their Facebook

Report this as inappropriate

Add demographics, interests or behaviours | Suggestions Browse

Exclude people or Narrow audience

□ Get better results by showing this advert to additional groups of people who are likely to engage with it.

Your audience selection is great!

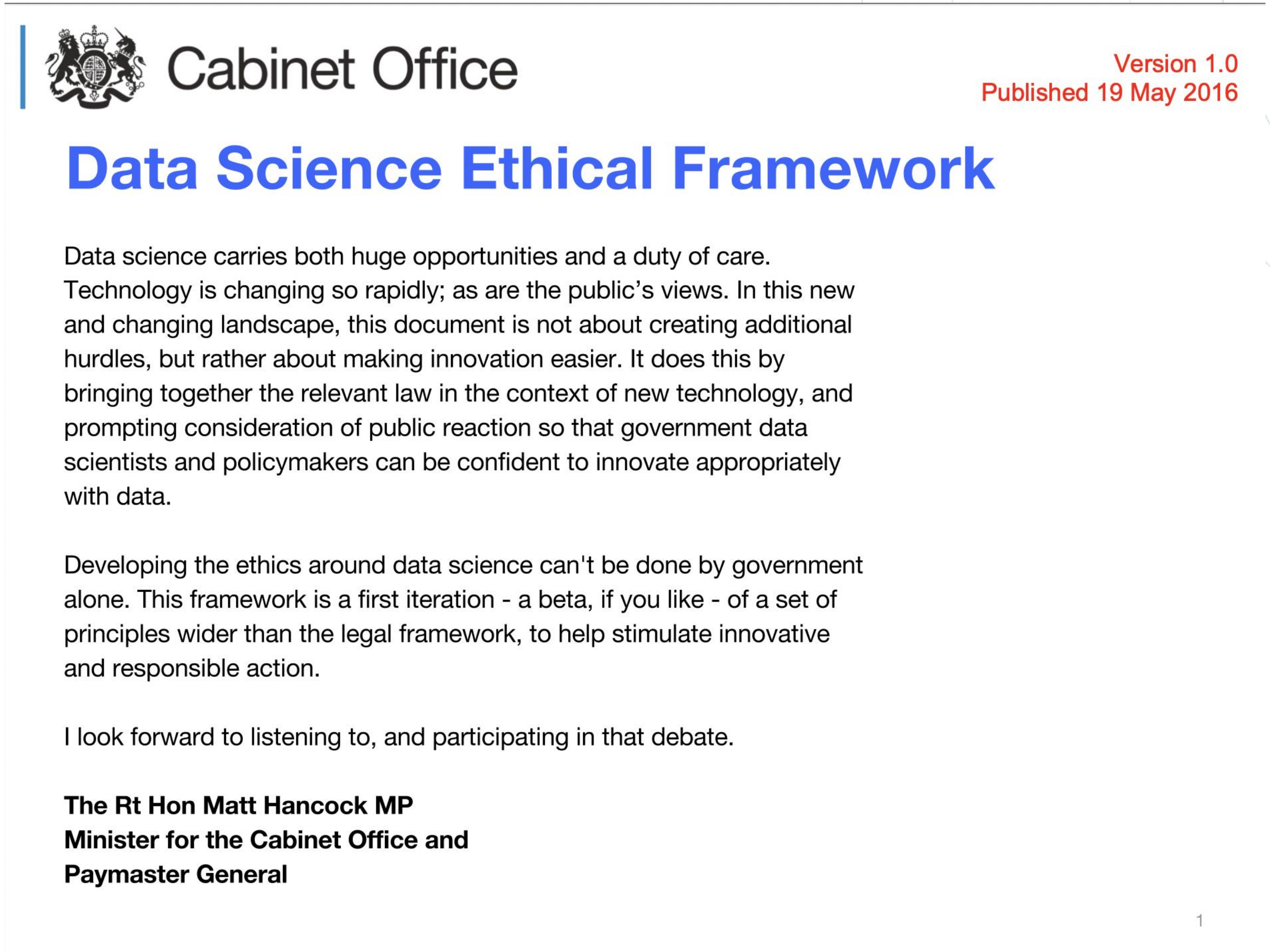
Potential audience size: 108,000 people

Cancel Save

# Lecture Zuboff

# Module 6: Governance and AI Act

How does the digital revolution transform our views on values, good behaviors, and the kind of sustainable and equitable innovation?



Cabinet Office

Version 1.0
Published 19 May 2016

# Data Science Ethical Framework

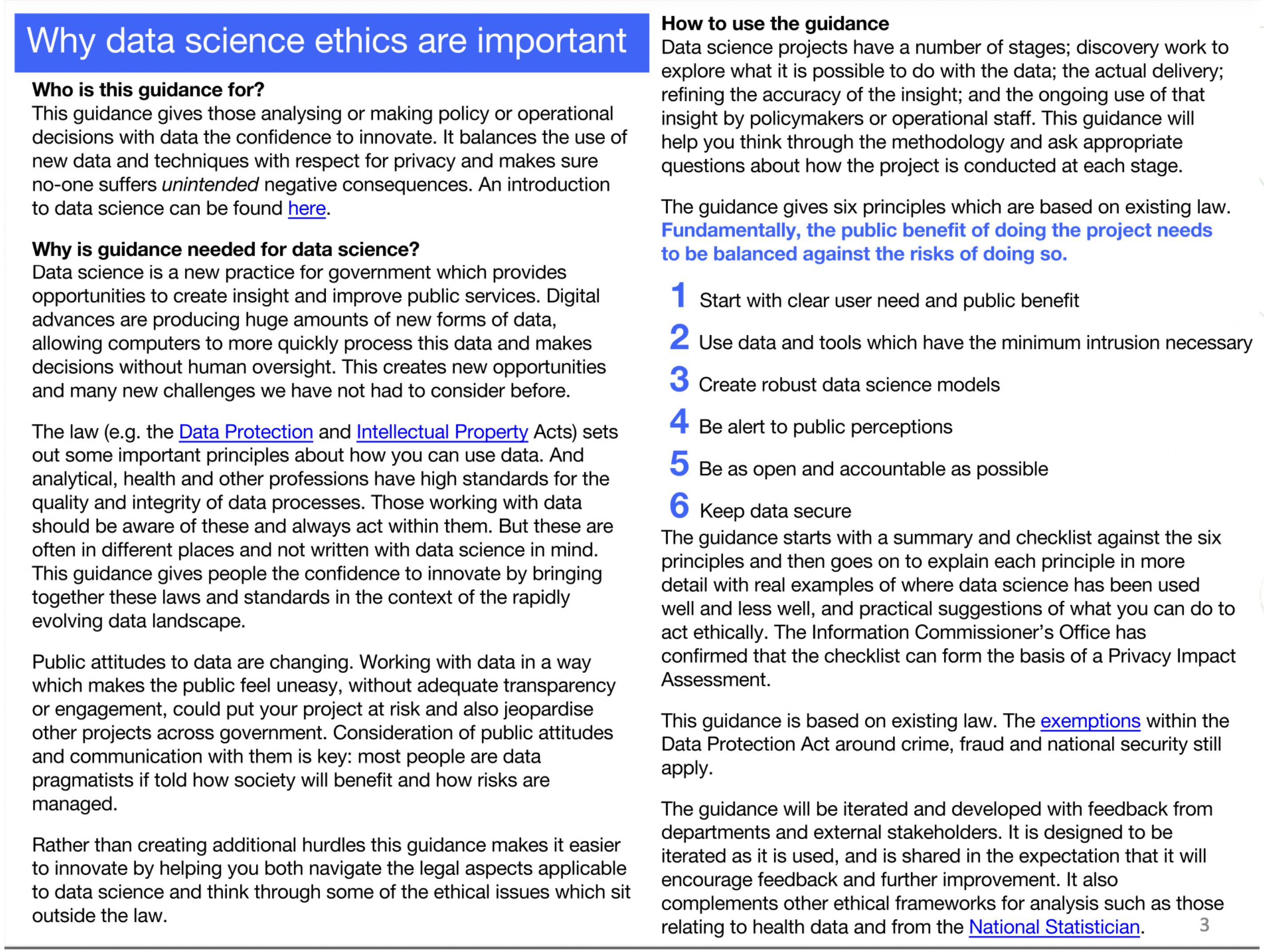
Data science carries both huge opportunities and a duty of care. Technology is changing so rapidly; as are the public's views. In this new and changing landscape, this document is not about creating additional hurdles, but rather about making innovation easier. It does this by bringing together the relevant law in the context of new technology, and prompting consideration of public reaction so that government data scientists and policymakers can be confident to innovate appropriately with data.

Developing the ethics around data science can't be done by government alone. This framework is a first iteration - a beta, if you like - of a set of principles wider than the legal framework, to help stimulate innovative and responsible action.

I look forward to listening to, and participating in that debate.

**The Rt Hon Matt Hancock MP**
**Minister for the Cabinet Office and**
**Paymaster General**

1



## Why data science ethics are important

**Who is this guidance for?**
This guidance gives those analysing or making policy or operational decisions with data the confidence to innovate. It balances the use of new data and techniques with respect for privacy and makes sure no-one suffers *unintended* negative consequences. An introduction to data science can be found here.

**Why is guidance needed for data science?**
Data science is a new practice for government which provides opportunities to create insight and improve public services. Digital advances are producing huge amounts of new forms of data, allowing computers to more quickly process this data and makes decisions without human oversight. This creates new opportunities and many new challenges we have not had to consider before.

The law (e.g. the Data Protection and Intellectual Property Acts) sets out some important principles about how you can use data. And analytical, health and other professions have high standards for the quality and integrity of data processes. Those working with data should be aware of these and always act within them. But these are often in different places and not written with data science in mind. This guidance gives people the confidence to innovate by bringing together these laws and standards in the context of the rapidly evolving data landscape.

Public attitudes to data are changing. Working with data in a way which makes the public feel uneasy, without adequate transparency or engagement, could put your project at risk and also jeopardise other projects across government. Consideration of public attitudes and communication with them is key: most people are data pragmatists if told how society will benefit and how risks are managed.

Rather than creating additional hurdles this guidance makes it easier to innovate by helping you both navigate the legal aspects applicable to data science and think through some of the ethical issues which sit outside the law.

**How to use the guidance**
Data science projects have a number of stages; discovery work to explore what it is possible to do with the data; the actual delivery; refining the accuracy of the insight; and the ongoing use of that insight by policymakers or operational staff. This guidance will help you think through the methodology and ask appropriate questions about how the project is conducted at each stage.

The guidance gives six principles which are based on existing law. **Fundamentally, the public benefit of doing the project needs to be balanced against the risks of doing so.**

1 Start with clear user need and public benefit

2 Use data and tools which have the minimum intrusion necessary

3 Create robust data science models

4 Be alert to public perceptions

5 Be as open and accountable as possible

6 Keep data secure

The guidance starts with a summary and checklist against the six principles and then goes on to explain each principle in more detail with real examples of where data science has been used well and less well, and practical suggestions of what you can do to act ethically. The Information Commissioner's Office has confirmed that the checklist can form the basis of a Privacy Impact Assessment.

This guidance is based on existing law. The exemptions within the Data Protection Act around crime, fraud and national security still apply.

The guidance will be iterated and developed with feedback from departments and external stakeholders. It is designed to be iterated as it is used, and is shared in the expectation that it will encourage feedback and further improvement. It also complements other ethical frameworks for analysis such as those relating to health data and from the National Statistician.

# Data Science Ethical Framework

- A document that aims to make innovation easier to understand.
- To achieve this, it attempts to bring together relevant regulations and user feedback so that policymakers and scientists can work effectively on data management.
- It is a guide that balances the use of new data and techniques with respect for privacy, ensuring that no one faces unintended consequences.
- Data science is a practice used by governments to improve public services.

# Data Science Ethical Framework

The guide establishes six principles based on existing laws:

- Start with a clear user need and a public benefit
- Use data and tools that are minimally intrusive
- Build robust models
- Stay vigilant regarding public perceptions
- Be open and accountable
- Take data security into consideration

# Digital ethics

- Understanding which of these moves are the best is the task of **digital ethics**.
- It is the branch of knowledge that deals with moral issues related to data and information, aiming to support morally good actions.
- It shapes both digital governance and regulation.

# Digital ethics

### SOFT ETHICS

- Post-compliance ethics of established legal norms.
- Considers what should or should not be done beyond existing regulations.

### HARD ETHICS

- Precedes and helps shape legal regulation.
- Analyzes what is generally right or wrong.

**The real challenge is to anticipate ethical development, and to do so, we must be able to assess what is truly feasible, prioritizing what is sustainable from both an environmental and social perspective.**

## ADOPTING AN ETHICAL APPROACH OFFERS A DOUBLE ADVANTAGE:

- **Soft ethics** can provide an opportunity-driven strategy
- **Ethics** offers a solution for effective risk management
- This is possible when there is **adequate legislation**, **public trust**, and **clear accountability**.
-

ITINERIS

## WHY IS IT SO URGENT?

Because ethical evaluation shapes public opinion

↓

which determines what is politically possible

↓

and thus what is legally enforceable

# Debate on the Ethics of algorithms

- **Algorithms** = mathematical constructs that can be developed into a program or a configuration.
- Algorithms underpin the services and infrastructures of society (e.g., recommendation systems).
- They are used in schools, hospitals, financial institutions, courts, local government bodies, and national governments.
- Their growing use prompts us to reflect on their ethical implications.

## Are algorithms neutral?

# Mapping the ethical issues related to algorithms

Machine learning algorithms ("automata" or "semi-automata" because their results are induced by data and are not deterministic) can be used to:

•Transform data into evidence for another outcome

•Trigger or motivate an action that may have ethical consequences

•Assign responsibility for the effects of actions that an algorithm may trigger

# Six ethical issues

1. **Inconclusive evidence**
2. **Inscrutable evidence**
3. **Misleading evidence**
4. **Unfair outcomes**
5. **Transformative effects**
6. **Traceability**

# Inconclusive evidence

They refer to the way in which machine learning algorithms identify associations and correlations between data variables, but not causal connections → thus generating **probabilistic outputs**

**Apophenia**

The tendency to perceive meaningful patterns where none actually exist.

# Inconclusive evidence

**PROBLEM:**

They generate ethical problems because non-causal indicators divert attention from the underlying cause of the problem — the problem is not solved by simply having more data.

**SOLUTION:**

Through independent validation of algorithms and practices that ensure data preservation and reproducibility.

# Inscrutable evidence

They refer to issues of non-transparency, lack of control, and accountability

1. Cognitive impossibility for humans to interpret algorithmic models and data systems
2. Lack of tools to keep track of large volumes of code and data
3. Poor structuring of data and code causing unreadability
4. Continuous updates (malleability of algorithms)
5. Human influence on the model

# Inscrutable evidence

Transparency is not an ethical principle in itself, but a tool for ethical purposes

**Solution:**

Highlight the causes that lead to non-transparency

1.The AI Now Institute (New York University) has produced a guide for algorithmic impact assessment which includes an interpretable explanation of internal processes (such as Explainable AI made available by Google and IBM).

2.Implementation of public education

# Misleading evidence

**Unintended *biases* are of concern**

↓

Caused by algorithmic formalism

↓

Adherence to prescribed rules that ignore the complexity of the world

↓

1. Inability to model the entire system to which the social criterion will be applied
2. Inability to understand how replicating a solution in one context can be misleading or inaccurate in another
3. Inability to account for fairness
4. Inability to understand that the introduction of a new technology may change the behaviors and/or values of a system
5. Inability to understand that the best solution may be dissociated from the technology

# Misleading evidence

**Biases derive from the data used**

↓

**Solution:**
Remove certain specific data variables (e.g., gender, race — prohibited by anti-discrimination law and data protection regulations)
**And what about invisible biases?**
Through external auditing and public disclosure of a model along with the data and metadata used.

# Unfair outcomes

**Related to the need for algorithmic fairness**

DEFINED BY 4 PRINCIPLES

- **Anti-classification:** refers to protected categories (race, gender)
- **Classification parity:** a model is fair if common measurements are equal across protected groups
- **Calibration:** measures how well an algorithm is calibrated
- **Statistical parity:** views fairness as an equal estimate of the average probability across protected groups

# Unfair outcomes

**Removing protected variables does not mean eliminating the discrimination rate because invisible variables exist.**

↓

**Solution:**
Introduce a third party that attempts to eliminate possible discrimination.
A collaborative model focused on the data resource.

# Transformative effects

They are intertwined with issues concerning autonomy.

1. Pervasive and proactive distribution of algorithms in shaping user choices
2. Limited understanding of algorithms by users
3. Lack of power over algorithmic outcomes

↓

**Solution:**
Autonomy must be constrained and reversible

# Traceability

**It concerns the issue of responsibility**

On one hand, it seems impossible to define moral responsibility due to the human-machine hybridization /
On the other hand, it is unthinkable to trace data back to its source.

Lack of transparency + lack of explainability =

**NEED FOR NEW APPROACHES**

# The European Commission regulation

In April **2021**, the European Commission published a proposal for **a regulation on the European approach to artificial intelligence.**

It includes transparency rules and more specific provisions regarding high-risk systems (such as personnel selection systems and systems assessing the reliability of statements made by individuals to prevent or investigate crimes).

**It prohibits:**

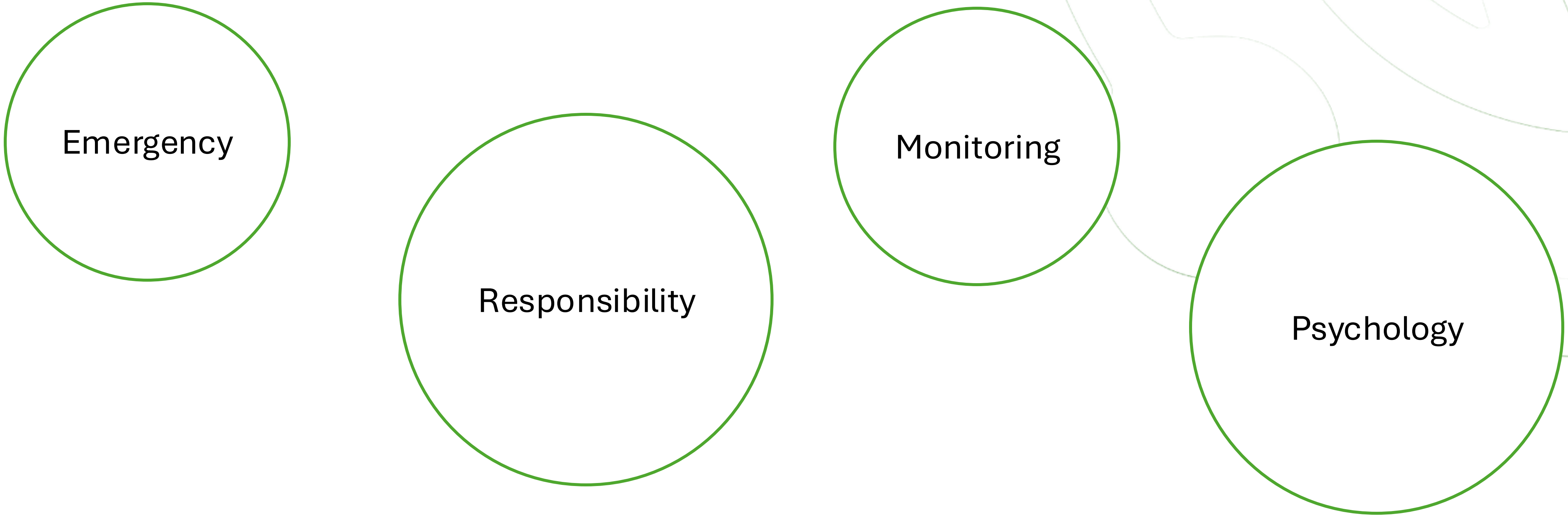
**•The use of systems employing subliminal techniques on unaware individuals**
**•The deployment by public authorities of systems to assess individuals' reliability based on their social behavior or personality traits**
**•The use of real-time remote biometric identification systems**

# Bad Practices of AI

**Five Criminal Areas Affecting AI Crimes (Artificial Intelligence Crimes):**

1. Trade and Financial Markets
2. Harmful and Dangerous Drugs
3. Crimes Against the Person
4. Sexual Crimes
5. Theft, Fraud, Counterfeiting, and Identity Theft

# Concerns Related to AI Crimes (CIA)



Emergency

Responsibility

Monitoring

Psychology

# Emergency

It refers to the possibility that artificial agents may act in ways that go beyond our expectations

# Emergency

It refers to the possibility that artificial agents may act in ways that go beyond our expectations

# Responsability

It refers to the possibility that AI systems can undermine existing models of responsibility → 
It would undermine legal certainty → No law would be recognized that could acknowledge the crime itself

# Emergency

It refers to the possibility that artificial agents may act in ways that go beyond our expectations

# Responsability

It refers to the possibility that AI systems can undermine existing models of responsibility →
It would undermine legal certainty → No law would be recognized that could acknowledge the crime itself

# Monitoring

1. It refers to three types of problems: 1. Attribution of non-compliance with regulations 2. Feasibility: concerns cases where artificial agents operate at speeds or levels of complexity beyond the capacity to monitor the rules (e.g., bots on social media) 3. Intersystem actions: refer to experiments showing how to automatically reproduce a user’s identity (e.g., on Twitter)

## Emergency

It refers to the possibility that artificial agents may act in ways that go beyond our expectations

## Responsability

It refers to the possibility that AI systems can undermine existing models of responsibility →
It would undermine legal certainty → No law would be recognized that could acknowledge the crime itself

## Monitoring

1. It refers to three types of problems: 1. Attribution of non-compliance with regulations 2. Feasibility: concerns cases where artificial agents operate at speeds or levels of complexity beyond the capacity to monitor the rules (e.g., bots on social media) 3. Intersystem actions: refer to experiments showing how to automatically reproduce a user’s identity (e.g., on Twitter)

## Psychology

It refers to the concern that AI could negatively influence and manipulate a user’s mental state to the point of facilitating a crime → Among the effects, gaining users' trust is noted

# AI Act

- **Entered into force**: August 1, 2024
- Some provisions are **already applicable**
- Others require a **transitional period** due to complexity and implementation needs
- Promoted by the **European Commission** to:

•Support early adoption of the AI Act

•Encourage developers to comply **before official deadlines**

**Pillar 1: Knowledge Sharing through the AI Pact Network**

•Access point to the AI Pact community

•Exchange of **best practices**, experiences, and expertise

•Updates and guidance on **AI Act implementation**

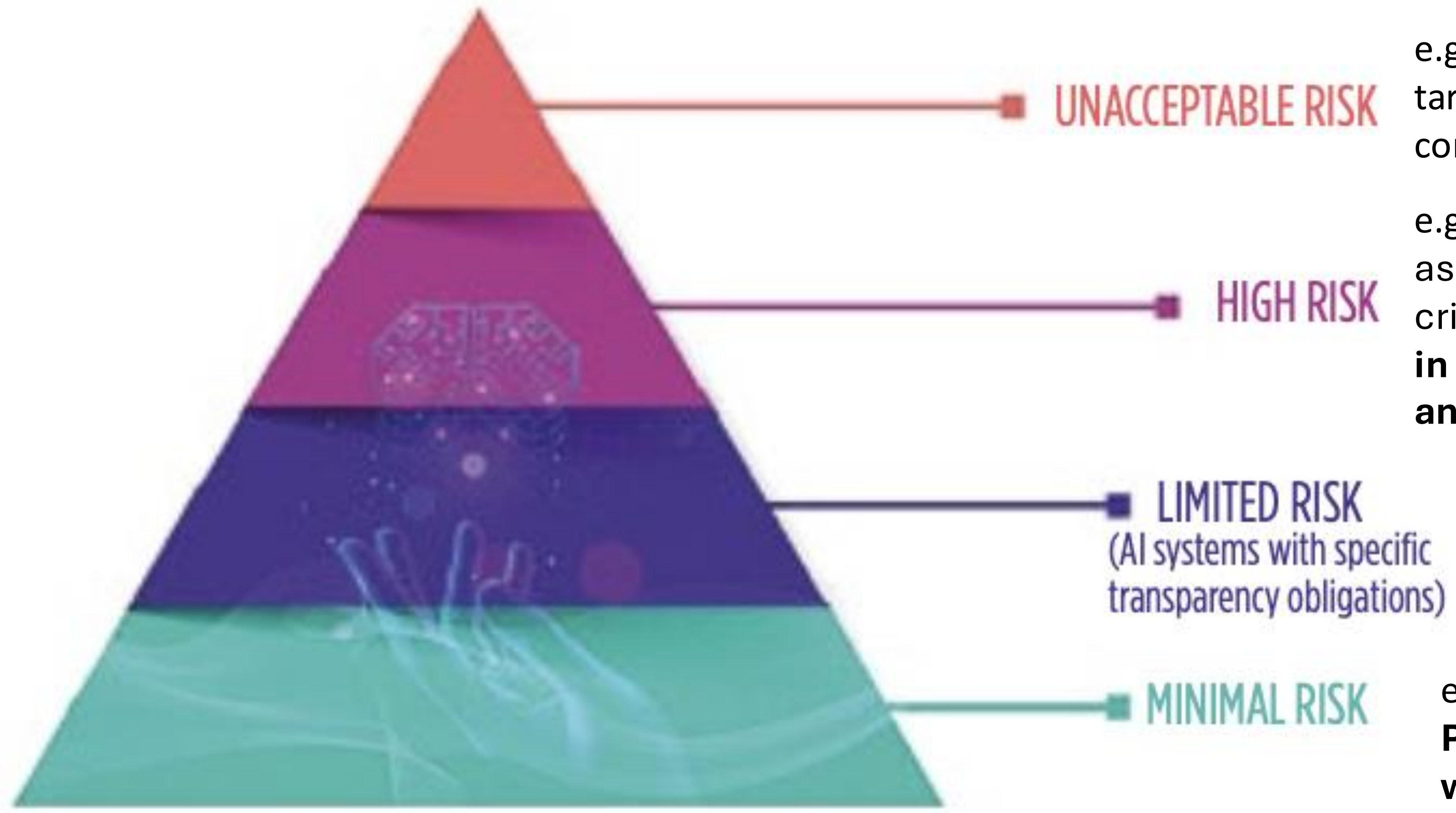
**Pillar 2: Facilitating Business Commitments**

•Framework to **accelerate adoption** of AI Act measures

•Encourages AI providers/operators to:

- Prepare in advance
- Adopt compliance strategies early

# AI Risk Classification in the AI Act



e.g. Social scaring, advertising messages targeted at children, psychological conditioning → **forbidden**

e.g. credit scoring, personnel selection, assisted surgery, police operations, critical infrastructures → **Permission in compliance with AI requirements and prior conformity assessment**

e.g. chatbot → **Permission in compliance with transparency and information obligations**

e.g. videogames, anti-spam systems → **Permission without obligations but with a recommended code of conduct**

# AI Act and enviromental monitoring

Currently, the AI Act does not contain specific articles on environmental monitoring through artificial intelligence. The existing provisions address environmental impact in a limited and non-binding manner.

BREAK

# Project

**Design a (even hypothetical) idea for using artificial intelligence to address a real environmental issue, considering both the potential benefits and the ethical and technical challenges.**

- **Total time: 40–60 minutes**
  **1. Choose an environmental problem (5 min)**
- Each group selects an area, for example: Air pollution, deforestation, coastal degradation, biodiversity monitoring, urban energy consumption, flood or wildfire prediction
- **2. Design an AI application**

For the chosen problem, answer the following questions:

- What data is needed? (satellite images, sensors, open data, etc.)
- What type of AI would you use? (e.g. computer vision, NLP, predictive models, etc.)
- Who benefits from it? (public administrations, citizens, NGOs, researchers...)
- What are the ethical limitations?
- How can they be mitigated? (transparency, audits, participatory governance...)

- **3. Prepare a brief presentation:** each group prepares a quick presentation (2–3 minutes) of their project.

# Source

1. Luciano Floridi, *Etica dell’intelligenza artificiale*, 2022, Raffaello Cortina Editore
2. *Ecoscienza – sostenibilità e controllo ambientale,* rivista di Arpae, n.4 – Ottobre 2024
3. https://www.isprambiente.gov.it/contentfiles/00003800/3874-rapporti-02-27.pdf/
4. https://www.ecmwf.int/sites/default/files/elibrary/2006/9299-bias-correction-environmental-monitoring.pdf
5. David B. Olawade a,b,c,* , Ojima Z. Wada d , Abimbola O. Ige e , Bamise I. Egbewole f , Adedayo Olojo g , Bankole I. Oladapo, *Artificial intelligence in environmental monitoring: Advancements, challenges, and future directions,* 2024
6. https://www.ecmwf.int/sites/default/files/elibrary/2006/9299-bias-correction-environmental-monitoring.pdf

ITINERIS

# THANKS!

**IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System**
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell’Università e della Ricerca

Italiadomani