AI, Ethics, and Society

22.10.2025

Recap Reflection and Questions

- How can we reduce the exploitation of data labelers and other hidden AI workers while still meeting the demand for large, high-quality datasets?

➢ Shift from mass manual labelling to ethical hybrid models: use active learning, automated pre-labelling, and smaller, fairly paid expert review teams. Transparency and fair-labour certification schemes should become mandatory for all AI supply-chain participants.

Recap Reflection and Questions

- How can we evaluate whether socially beneficial AI applications (e.g. facial recognition for safety, healthcare diagnostics) are ethically acceptable to deploy?

➢ Use context-sensitive proportionality assessments: weigh demonstrable public benefit against potential harm, ensure oversight, reversibility, and redress mechanisms, and only deploy where safeguards make accountability traceable and contestable. (See today's lecture as well)

Recap Reflection and Questions

- What forms of transparency and consent genuinely empower users rather than overwhelm them with technical or legal detail?

➢ Transparency should be actionable: explain purpose, data sources, and recourse options in clear language. Consent must be continuous, contextual, and revisable. Moving from a one-time legal checkbox to an ongoing, informed relationship between user and system.

Recap Reflection and Questions

- In the face of environmental, social, and epistemic costs, what justifies continuing large-scale AI research and deployment?

➢ In my opinion, only purpose-driven innovation can justify such costs. Applications that deliver measurable public good (healthcare, education, climate modelling) and that are built with sustainability and accountability as design constraints. Otherwise, restraint becomes the ethical position.

Moving on…

The EU AI Act
And it’s Code of Practice

The EU AI Act

Four major points:

1. Who has the obligation?
2. Who is the “User”?
3. Classification of AI according to its risks
4. General Purpose AI (GPAI)

EU AI Act - Who has the obligation?

The majority of obligations fall on providers (developers) of high-risk AI systems.

- Those that intend to place on the market or put into service high-risk AI systems in the EU, regardless of whether they are based in the EU or a third country.
- And also third country providers where the high risk AI system's output is used in the EU.

https://artificialintelligenceact.eu/high-level-summary/

EU AI Act - Who is the “User”?

Users are natural or legal persons that deploy an AI system in a professional capacity, not affected end-users.

- Users (deployers) of high-risk AI systems have some obligations, though less than providers (developers).
- This applies to users located in the EU, and third country users where the AI system’s output is used in the EU.

https://artificialintelligenceact.eu/high-level-summary/

The EU AI Act - Classification by Risks

Four different risk levels:

1. Unacceptable risks - prohibited
2. High risks - highly regulated
3. Limited risks - transparency obligations
4. Minimal risks - unregulated

Minimal Risks (https://artificialintelligenceact.eu/high-level-summary/)

No further regulation. This includes the majority of AI applications currently available on the EU single market, such as AI enabled video games and spam filter (as of 2021, currently changing due to GenAI)

Limited Risk (https://artificialintelligenceact.eu/high-level-summary/)

Developers and deployers must ensure that end-users are aware that they are interacting with AI. Includes, for example, chatbots or deepfakes.

Unacceptable Risks (https://artificialintelligenceact.eu/high-level-summary/)

AI systems that do one of the following are prohibited according to the AI act. Keep in mind, there is an extra section on General Purpose AI systems.

AI systems:

- Deploying subliminal, manipulative, or deceptive techniques or exploiting vulnerabilities related to age, disability, or socio-economic circumstances to distort behaviour causing significant harm.
- Biometric categorisation systems inferring sensitive attributes, except labelling or filtering of lawfully acquired datasets or within law enforcement.

Unacceptable Risks (https://artificialintelligenceact.eu/high-level-summary/)

Continued:

- Social scoring
- Assessing the risk of an individual committing criminal offenses solely based on profiling or personality traits (except when used to augment human assessment)
- Compiling facial recognition databases by untargeted scraping of facial images (from the internet or CCTV)
- Inferring emotions in workplaces or educational institutions (except for medical or safety reasons)

Unacceptable Risks (https://artificialintelligenceact.eu/high-level-summary/)

- Real-time remote biometric identification in publicly accessible spaces for law enforcement, except when:
  - Searching for missing persons, abduction victims, etc,
  - Preventing substantial and imminent threat of life, or foreseeable terrorist attacks, or
  - Identifying suspects in serious crimes (e.g., murder, rape, armed robbery, narcotic and illegal weapons trafficking, organised crime, and environmental crime, etc.)
- Even if allowed, only when:
  - not using it would cause considerable harm,
  - before deployment, police completed a fundamental rights impact assessment and registered the system in the EU database, and
  - before deployment, authorisation from a judicial authority or independent administrative authority was obtained.

High Risks (https://artificialintelligenceact.eu/high-level-summary/)

High risk AI systems are those:

- used as a safety component or a product covered by EU laws in Annex I AND required to undergo a third-party conformity assessment under those Annex I laws; OR
- listed under Annex III use cases ([next slides]), except if:
  - the AI system performs a narrow procedural task;
  - improves the result of a previously completed human activity;
  - detects decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment without proper human review; or
  - performs a preparatory task to an assessment relevant for the purpose of the use cases listed in Annex III.

High Risks (https://artificialintelligenceact.eu/high-level-summary/)

- AI systems listed under Annex III 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, behaviour, location or movement.
- Providers whose AI system falls under the use cases in Annex III but believes it is not high-risk must document such an assessment before placing it on the market or putting it into service.

High Risks - Systems listed in Annex III

Annex III use cases
Non-banned biometrics: Remote biometric identification systems, excluding biometric verification that confirm a person is who they claim to be. Biometric categorisation systems inferring sensitive or protected attributes or characteristics. Emotion recognition systems.
Critical infrastructure: Safety components in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity.
Education and vocational training: AI systems determining access, admission or assignment to educational and vocational training institutions at all levels. Evaluating learning outcomes, including those used to steer the student's learning process. Assessing the appropriate level of education for an individual. Monitoring and detecting prohibited student behaviour during tests.
Employment, workers management and access to self-employment: AI systems used for recruitment or selection, particularly targeted job ads, analysing and filtering applications, and evaluating candidates. Promotion and termination of contracts, allocating tasks based on personality traits or characteristics and behaviour, and monitoring and evaluating performance.

High Risks - Systems listed in Annex III

Access to and enjoyment of essential public and private services: AI systems used by public authorities for assessing eligibility to benefits and services, including their allocation, reduction, revocation, or recovery. Evaluating creditworthiness, except when detecting financial fraud. Evaluating and classifying emergency calls, including dispatch prioritising of police, firefighters, medical aid and urgent patient triage services. Risk assessments and pricing in health and life insurance.

Law enforcement: AI systems used to assess an individual's risk of becoming a crime victim. Polygraphs. Evaluating evidence reliability during criminal investigations or prosecutions. Assessing an individual's risk of offending or re-offending not solely based on profiling or assessing personality traits or past criminal behaviour. Profiling during criminal detections, investigations or prosecutions.

High Risks - Systems listed in Annex III

Migration, asylum and border control management: Polygraphs. Assessments of irregular migration or health risks. Examination of applications for asylum, visa and residence permits, and associated complaints related to eligibility. Detecting, recognising or identifying individuals, except verifying travel documents.

Administration of justice and democratic processes: AI systems used in researching and interpreting facts and applying the law to concrete facts or used in alternative dispute resolution. Influencing elections and referenda outcomes or voting behaviour, excluding outputs that do not directly interact with people, like tools used to organise, optimise and structure political campaigns.

The EU AI Act - 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 – open or closed – must also conduct model evaluations, adversarial testing, track and report serious incidents and ensure cybersecurity protections.

(https://artificialintelligenceact.eu/high-level-summary/)

The EU AI Act - General Purpose AI (GPAI)

There is a lot of stuff in the EU AI Act about GPAI and how they are regulated.

One way to resolve this regulation is to sign and follow the Code of Practice published by the AI office of the EU.

So what is this Code of Practice?

Code of Practice

What it is:

- It offers a clear framework to help developers of GPAI models meet the requirement of the EU AI Act.
- A set of measures the company can implement sorted into 10 forms of commitments regarding three different aspects (chapters).

What it isn’t:

- A definite regulation that all AI providers must follow:
  - Providers are free to demonstrate compliance through other appropriate measures.

Code of Practice - Three Chapters

The Code of Practice is divided into three chapters:

1. Transparency
2. Copyright
3. Safety and Security

Each chapter is divided into commitments (in total 10) and each commitment divided into implementable measures (in total 40)

We will not cover all of them!

Code of Practice - Transparency

Regards the documentation for every GPAI model distributed within the EU (except free, open-source models that pose no systemic risks).

Commitment 1: Documentation

- Signatories pledge to:
  - Keep comprehensive, current documentation for every model (Measure 1.1)
  - Facilitate information sharing with the AI Office and downstream providers (Measure 1.2)
  - Safeguard the accuracy, completeness, and protection of all documentation (Measure 1.3)

These commitments do not apply to free and open-source models, unless they are classified as a GPAI model with systemic risk.

Code of Practice - Copyright

Ensures alignment with EU Copyright law, especially the requirement for prior authorization unless specific exception apply (such as text and data mining).

A robust copyright policy that clearly defines internal responsibilities and complies with legal standards must be developed and regularly updated.

Data collected via web crawling is lawfully accessible, respect machine reasonable signals like robot.txt, and avoid accessing websites flagged for copyright infringement.

Additional technical safeguards should minimize the generation of infringing content, and ToS must clearly prohibit unauthorized use.

And more…

Code of Practice - Safety and Security

Longest chapter, including 8 of the 10 commitments and 32 of the 40 measures.

1. Safety and Security Framework
2. Systemic risk identification
3. Systemic risk analysis
4. Systemic risk acceptance determination
5. Safety mitigations
6. Security mitigations
7. Safety and Security Model Reports
8. Systemic risk responsibility allocation

Code of Practice - Who has signed it?

As stated previously, this is not mandatory, so the big question is:

Who has signed it? And who is choosing their own compliance actions?

Code of Practice - Who has signed it?

Google

OpenAI

Microsoft

Amazon

IBM

Accexible

AI Alignment Solutions

Aleph Alpha

Antropic

Bria AI

Cohere

Cyber Institute

Domyn

Dweve

Fastweb

Humane Technology

Lawise

Open Hippo

Pleias

Re-inverta

ServiceNow

Virtue Turing

Euc Inovação Portugal

And more to come…

Code of Practice - Who hasn’t signed it?

Of course, a lot of companies are not on this list (yet), but most notably:

Meta has publicly announced that it would not sign.

xAI has only agreed to signing the chapter on safety, decrying the rest as “profoundly detrimental to innovation”