Dr Elias Pimenidis

Professional Studies in Computing - UFCFJS-15-3

Ethical Artificial Intelligence

What constitutes Ethical AI

- Developing AI empowered solutions that support, reproducibility, equality, and sustainability.
- Use resources in a way that does not disadvantage other parts of society.
- Improve quality of services while maintaining quality of life.
- Enhance efficiency and effectiveness while enhancing the level of risk controls.
- Develop jobs that deliver quality and at the same time contribute towards reducing poverty.

The Responsible Machine Learning Principles

- Human augmentation - I commit to assess the impact of incorrect predictions and, when reasonable, design systems with human-in-the-loop review processes.
- Bias evaluation - I commit to continuously develop processes that allow me to understand, document and monitor bias in development and production.
- Explainability by justification - I commit to develop tools and processes to continuously improve transparency and explainability of machine learning systems where reasonable.

The Responsible Machine Learning Principles - 2

- Reproducible operations - I commit to develop the infrastructure required to enable for a reasonable level of reproducibility across the operations of ML systems.
- Displacement strategy - I commit to identify and document relevant information so that business change processes can be developed to mitigate the impact towards workers being automated.
- Practical accuracy - I commit to develop processes to ensure my accuracy and cost metric functions are aligned to the domain-specific applications.

The Responsible Machine Learning Principles - 3

- Trust by privacy - I commit to build and communicate processes that protect and handle data with stakeholders that may interact with the system directly and/or indirectly.
- Data risk awareness - I commit to develop and improve reasonable processes and infrastructure to ensure data and model security are being taken into consideration during the development of machine learning systems.

Conclusion – Beware and Manage the Risk = Sustainable AI

https://ethical.institute/index.html The Institute for Ethical AI & Machine Learning

How do you support Sustainability in your work?

- Think about the impact of your work – Assess the Risk
  - On Humans
  - The Environment
  - Your Professional World
- Consider alternative designs
- Compromise where you can
  - Limit the use of data
  - Limit the demands on power
  - Create and use less data hungry algorithms, where possible

Explainable AI and Supply Chain

An agile, data-driven supply chain ecosystem will be better prepared **to react and mitigate such impacts.**

AI can help predict and mitigate successfully.

Two problems though

Small supply chains will no benefit – small data sets

Specialist operations will suffer the same challenge.

There is hope though – Explainable Artificial Intelligence.

Explainable AI and Supply Chain – 2/3

- Where the data that we have is in tabular format or in free text format. To process such data using explainable AI (there are many methods) we should explore the concept of anchors
- The basic idea is that individual predictions of any black-box classification model are explained by finding a decision rule that sufficiently “anchors” the prediction - hence the name “anchors”.

Explainable AI and Supply Chain 3/3

- The resulting explanations are decision rules in the form of IF-THEN statements, which define regions in the feature space. In these regions, the predictions are fixed (or “anchored”) to the class of the data point to be explained.
- Consequently, the classification remains the same no matter how much the other feature values of the data point that are not part of the anchor are changed. Therefore, we should be able to achieve reproducibility of results with different data sets in the same domain.

The AI Act - laying down harmonised rules on artificial intelligence

- Is the first-ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.
- Is the first-ever comprehensive legal framework on AI worldwide. The aim of the rules is to foster trustworthy AI in Europe.
- Sets out a clear set of risk-based rules for AI developers and deployers regarding specific uses of AI.
- Is part of a wider package of policy measures to support the development of trustworthy AI.

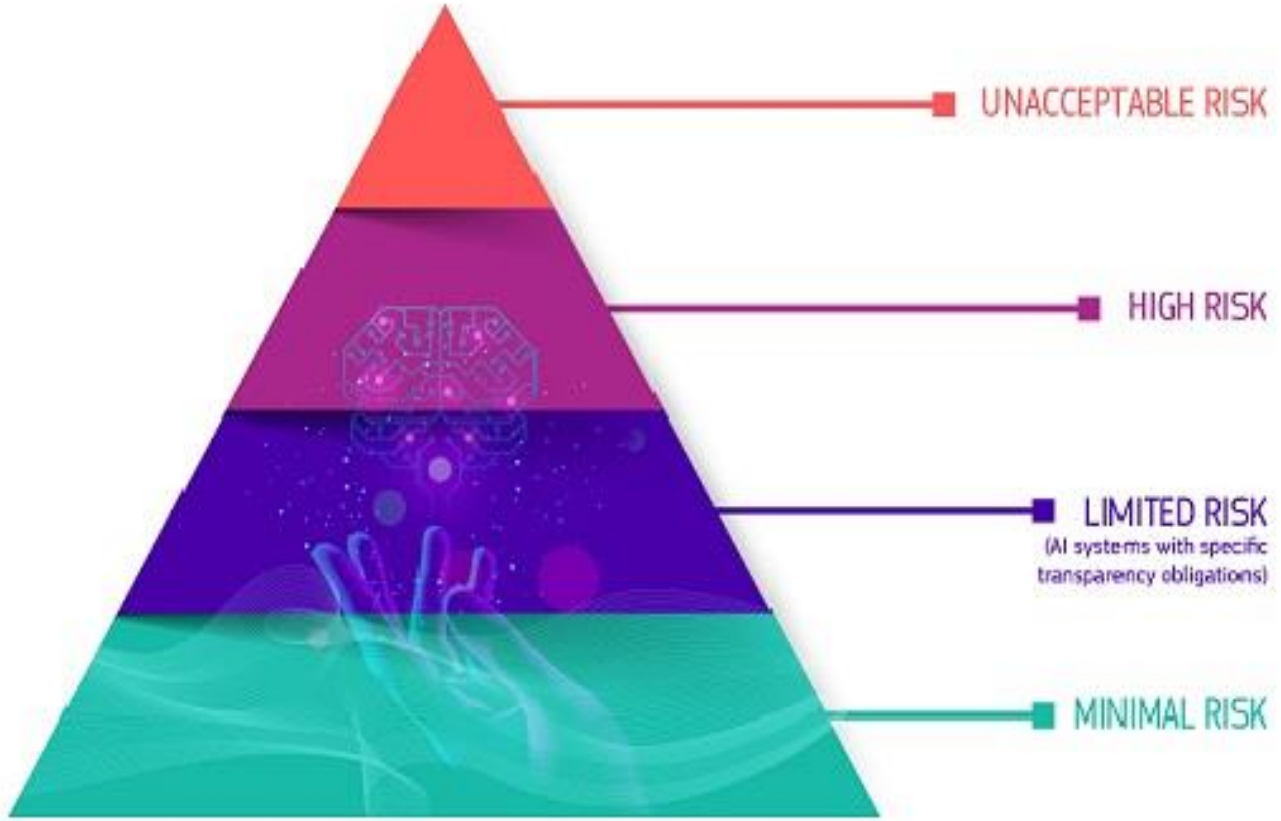
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Why do we need rules on AI?

- The AI Act ensures that Europeans can trust what AI has to offer. While most AI systems pose limited to no risk and can contribute to solving many societal challenges, certain AI systems create risks that we must address to avoid undesirable outcomes.
- It is often not possible to find out why an AI system has made a decision or prediction and taken a particular action. So, it may become difficult to assess whether someone has been unfairly disadvantaged, such as in a hiring decision or in an application for a public benefit scheme.

A risk-based approach

The AI Act defines 4 levels of risk for AI systems



Unacceptable risk - The AI Act prohibits eight practices

- harmful AI-based manipulation and deception
- harmful AI-based exploitation of vulnerabilities
- social scoring
- Individual criminal offence risk assessment or prediction
- untargeted scraping of the internet or CCTV material to create or expand facial recognition databases
- emotion recognition in workplaces and education institutions
- biometric categorisation to deduce certain protected characteristics
- real-time remote biometric identification for law enforcement purposes in publicly accessible spaces

High risk - AI use cases that can pose serious risks to health, safety or fundamental rights are classified as high-risk.

- High-risk AI systems are subject to strict obligations before they can be put on the market:
- adequate risk assessment and mitigation systems
- high-quality of the datasets feeding the system to minimise risks of discriminatory outcomes
- logging of activity to ensure traceability of results

High Risk 2/2

- detailed documentation providing all information necessary on the system and its purpose for authorities to assess its compliance
- clear and adequate information to the deployer
- appropriate human oversight measures
- high level of robustness, cybersecurity and accuracy

Transparency Risk 1/2

- This refers to the risks associated with a need for transparency around the use of AI. The Act introduces specific **disclosure obligations** to ensure that humans are informed when **necessary to preserve trust.**
  - For instance, when using AI systems such as chatbots, humans **should be made aware that they are interacting with a machine** so they can take an informed decision.

Transparency Risk 2/2

- Providers of generative AI must ensure that AI-generated content is identifiable.
- On top of that, certain AI-generated content should be clearly and visibly labelled, namely deep fakes and text published with the purpose to inform the public on matters of public interest.

Minimal or no risk

The AI Act does not introduce rules for AI that is deemed minimal or no risk.

The vast majority of AI systems currently used in the EU fall into this category.

This includes applications such as AI-enabled video games or spam filters.

How does it all work in practice for providers of high-risk AI systems?