Principles of AI Ethics and Limits of Contemporary AI -II

AI, Ethics, and Society

Arash Sheikhlar

September 12th and 16th 2024

Another important AI ethics principle*

Proportionality: AI system must be chosen/created **properly** such that

- **it is proportional to the type of the application**
- **it does not neglect the following values: human rights, safety, inclusiveness, and environment**
- **it should be based on a sound scientific basis**

* UNESCO. “Recommendations on the ethics of Artificial Intelligence" *UNESDOC digital library* (2022)

Some remarks on transparency*

- If a decision is made by a public sector or institution, the people whose rights have been affected should know whether an AI system plays a role (either partly or fully) in decision-making.
- In such cases, the people have the right to request explanations of why and how the decision has been made, and the private or public sector should provide information and revise the decision if proper explanations do not exist.
- E.g., applications for social benefits

- The designers whose AI systems affect the rights of other humans should commit to choosing/designing explainable algorithms and systems.

* UNESCO. “Recommendations on the ethics of Artificial Intelligence" *UNESDOC digital library* (2022)

Some remarks on responsibility*

- ***"Member States should ensure that it is always possible to attribute ethical and legal responsibility for any stage of the life cycle of AI systems, as well as in cases of remedy related to AI systems, to physical persons or to existing legal entities." – (UNESCO 2022)***

- According to the guideline, "traceability" and "auditability" of systems need to be considered by the designers, better helping the assessment of responsibility.

- ***Implication:***
- The less transparent the AI system is, the more likely it is that its designers will be held responsible for the negative impacts of the system.

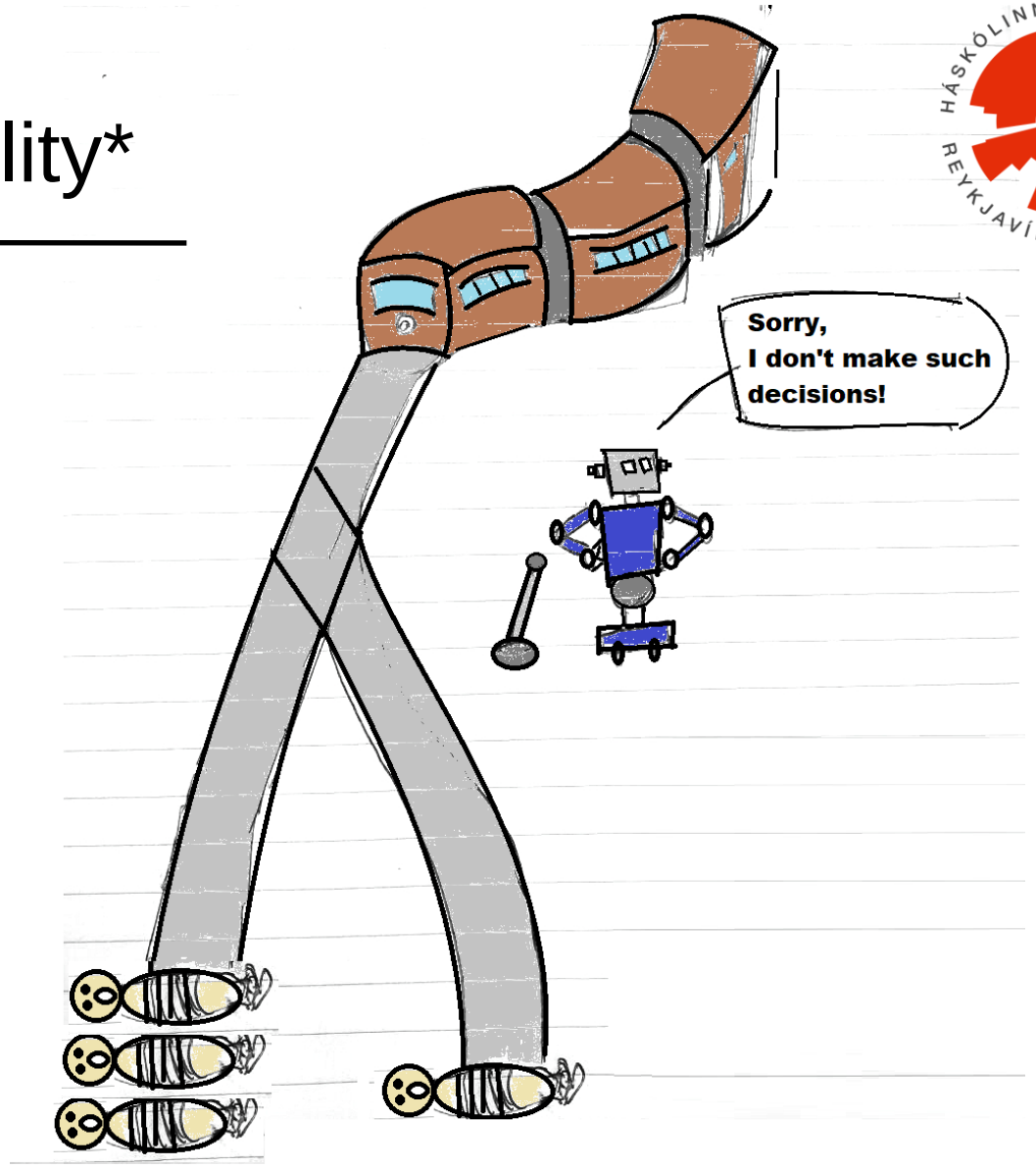
* UNESCO. "Recommendations on the ethics of Artificial Intelligence" *UNESDOC digital library* (2022)

Some remarks on responsibility*

- **“As a rule, life and death decisions should not be ceded to AI systems.” – *(UNESCO 2022)***

Examples of the domains where AI must not be fully autonomous (Discussion about the extent of AI’s use)

- (Discussion) Healthcare and medicine
- (Discussion) Self-driving cars
- (Discussion) Military and defense



* UNESCO. “Recommendations on the ethics of Artificial Intelligence" *UNESDOC digital library* (2022)

Group project – an AI-based job application system

Sara is a software engineer with considerable programming experience and a strong CV. She applies for a senior position at a large company through the company's online application system. During the application process, the company's application system assures Sara that her personal data will be used solely for the application and will be deleted after 6 months. The company's application system employs a job-matching algorithm to filter through hundreds of applications, identify the best candidates, and predict their potential performance. Yet, this algorithm is based on complex machine-learning models, the workings of which even the developers do not fully understand. The company and its developers, not the job applications, know that historical data was used to train these models.

A few weeks later, Sara receives a rejection email from the company without any clear explanation. She then learns that several male candidates with fewer qualifications were invited for an interview for the same position. This outcome significantly impacts Sara's self-confidence and leads her to doubt her abilities.

Instructions:

Divide into groups of 2-3 people. Then,

1. Discuss whether the AI system used is appropriate for the domain it is applied to. If it is, explain why. If it is not, discuss why and suggest an alternative.
2. Discuss the ethical violations related to AI ethics principles. Create a table listing the ethical issues and rank them in order of priority. Summarize your answers in the table.
3. Consider what actions Sara should take. If Sara wants to file a complaint about the job application outcome, what factors should she highlight in her complaint?

Different ANN architectures

1- Feedforward Neural Networks (FNN): basic ANN type only using forward propagation (data is passed between nodes in the forward direction)

- **Convolutional Neural Networks (CNN)**: use convolutional layers to give weights to specific regions on the input (more efficient than CNN) for feature extraction and build hierarchical feature representation (better for more complex structures)
- *Applications*: object recognition in images and videos, for example, YOLO4 and YOLO5

2- Recurrent Neural Networks (RNN): Use loops (hidden state) to maintain memory and sequential process of data sequences and they use backpropagation.

- **Long Short-Term Memory Networks (LSTM)**: A specific RNN useful for predicting and memorizing longer data sequences
- *Applications*: time series prediction, language translation

3- Transformers: parallel processing on data sequences and use "attention" to give importance to specific parts of data sequences, leading to better handling of longer correlations between data (more effective when dealing with complex data)

Applications: Natural language processing

Limitations of ANNs

CNNs and LSTMs

- CNNs keep the bias in the training data, leading to biased unfair image recognition (What else?)
- Both CNNs and LSTMs rely on an enormous amount of data to be trained (privacy issues)
- Both have explainability issues due to being black box models (lack of transparency)

Transformers

- Transformers need significant computational power and resources (justice issues and environmental impact)
  - Not everyone can afford training transformers
  - Substantial energy consumption (What else?)
  - Carbon footprint
- Transformers can be dangerous as they amplify the biases in the data.

A case study- Facial recognition systems

Systems that identify and recognize individuals based on their facial features (used by social media, phone production companies, and law enforcement). They match the features of the input images and videos by comparing them with the data in the databases.

Technology: CNNs

Privacy: Very large number of images they need, The privacy of users' data must be guaranteed by the platforms

Justice and fairness (Discussion) data biases in relation to race and gender may lead to unequal decisions

Transparency: no one can scrutinize why they are flawed

Responsibility: When mistakes happen, it is not clear who is responsible

Statistics-based AI methods and systems

- Naive Bayes classifier: based on Bayes theorem, it relies on the prior probabilities of each class before getting data estimates for posterior probability and predicting the class with the highest probability given a set of features.

- Linear regression: uses statistical techniques to derive a linear equation model to predict the relationship between an output with a set of input variables.

Systems

- Term frequency-inverse document (TF-ID): a statistical measure for assessing the importance of a word, using
  - how often a term appears in a document and how common it is across all documents gives a score suggesting the importance of the term.

  Application: text mining
- Content-based filtering: a statistical recommendation technique that suggests items based on features and the user's preference.
  - Components: items, item features, user profile, matching process, and recommendation generation.

  Application: Image and video recommendations

A case study – Social media algorithms

Social media algorithms (used by platforms such as Facebook, YouTube, and Twitter) show content relevant to the user's preferences to increase engagement.

Technologies: the combination of the following

- Inputs: likes, shares, comments, and watch time as input
- recommendation systems to recommend content relevant to users' past behavior
- computer vision analyzes the interacted images and videos and makes recommendations accordingly.

- *Transparency*: (Discussion) Opaque. Most of the time it is not clear why certain content is shown to users
- *Justice and fairness*: (Discussion) Reinforcing personal biases creates an echo chamber (risk of being radicalized)
- *Security*: These are not designed to cause harm, but have the potential for negatively affecting people with harmful content: misinformation, hate speech, and extreme views.
- *Responsibility*: (Discussion) When the harmful content is amplified, who is responsible?
- *Privacy*: Collect an enormous amount of data from users. It is unclear what type of data is collected.