Abdelhamid Mehri – Constantine 2 University
Faculty of New Information and Communication Technologies
Department of Fundamental Computer Science and Its Application

ETHICS AND GOVERNANCE OF ARTIFICIAL INTELLIGENCE

MASTER 1 SCIENCE DE DONNÉES ET INTELLIGENCE ARTIFICIELLE (SDIA)

DR ILHAM KITOUNI

2023-2024

LESSON 3 : FAIRNESS AND BIAS IN AI

ETHICAL PRINCIPLES FOR AI

An overview of ethical principles that should guide the development and use of AI systems. These principles include:

- **Beneficence:** AI systems should be developed and used in a way that benefits humanity.
- **Non-maleficence:** AI systems should not be developed or used in a way that harms humanity.
- **Autonomy:** AI systems should be designed to respect human autonomy.
- **Justice:** AI systems should be developed and used in a fair and just manner.

INTRODUCTION TO FAIRNESS IN AI

- Justice and fairness are fundamental concepts in society.
- They are also important in the development and use of artificial intelligence (AI).
- **Question:** What is fairness in AI?

INTRODUCTION TO FAIRNESS IN AI

- Fairness in AI means that AI systems should not discriminate against individuals or groups of people. However, AI systems can be biased in a number of ways.
- **Example:** An AI algorithm used to predict recidivism risk can be biased against people of color.

ALGORITHMIC BIAS

Different types of algorithmic bias, including:

- **Representation bias:** This occurs when the data used to train an AI system is not representative of the population that the system will be used on. For example, a facial recognition system that is trained on a dataset of mostly white faces may have difficulty recognizing black faces.
- **Selection bias:** This occurs when the AI system selects data in a way that introduces bias. For example, an AI system used to predict recidivism risk may be more likely to select data on people who have already been convicted of crimes, which could lead to the system being biased against certain groups of people.

ALGORITHMIC BIAS

- **Measurement bias:** This occurs when the AI system measures data in a way that introduces bias. For example, an AI system used to measure employee performance may be biased against women if it measures performance based on criteria that are typically seen as being more important for men.

TECHNIQUES FOR REDUCING BIAS IN AI

- There are a number of techniques that can be used to reduce bias in AI systems.
- One technique is to use a more diverse dataset to train the system.
- **Example:** A facial recognition dataset that includes a greater diversity of people will help to reduce bias in the system.

USING A MORE DIVERSE DATASET TO TRAIN AI SYSTEMS

- A diverse dataset is a dataset that includes a variety of people, including people of different races, ethnicities, and sexes.
- **Example:** A facial recognition dataset that includes images of people of all races, ethnicities, and sexes will help to reduce bias in the system.

TECHNIQUES TO DETECT AND REMOVE BIAS FROM AI SYSTEMS

- There are a number of techniques that can be used to detect and remove bias from AI systems.
- One technique is to use adversarial training techniques.
- **Example:** Adversarial training involves creating synthetic data that is designed to expose the biases of an AI system. The AI system is then trained on this synthetic data, which helps to reduce its biases.

CASE STUDY 1: THE COMPAS ALGORITHM AND RACIAL BIAS

- COMPAS is an algorithm used to predict recidivism risk.
- A study found that COMPAS was biased against black men.

CASE STUDY 2: AMAZON REKOGNITION AND BIAS AGAINST PEOPLE OF COLOR

- Amazon Rekognition is a facial recognition system used by police.
- A study found that Amazon Rekognition was biased against people of color.

CASE STUDY 3: BIAS IN SOCIAL MEDIA

- How AI systems can introduce bias in social media.
- For example, AI systems used to recommend content to users may be biased towards certain types of content or users.
- Additionally, AI systems used to filter content may be biased against certain groups of people or types of content.
- Analyse the potential consequences of biased AI systems in social media.
- For example, biased AI systems could lead to the spread of misinformation or could promote discrimination.

CASE STUDY 4: BIAS IN RECRUITMENT AI SYSTEMS

- Explain the potential for bias in AI systems used in hiring and recruitment.
- For example, AI systems used to screen resumes or interview candidates may be biased against certain groups of people.
- Discuss the ethical implications of biased AI systems in recruitment.
- For example, biased AI systems could lead to qualified candidates being overlooked or could lead to discrimination in the workplace.

CASE STUDY 5: BIAS IN HEALTHCARE AI SYSTEMS

- Analyse the potential for bias in AI systems used in healthcare.
- For example, AI systems used to diagnose diseases or recommend treatments may be biased against certain groups of people.
- Discuss the potential impact of biased AI systems in healthcare.
- For example, biased AI systems could lead to misdiagnosis or could lead to patients receiving suboptimal treatment.

CASE STUDY 6: BIAS IN FINANCIAL AI SYSTEMS

- How AI systems can be biased in the financial sector.
- For example, AI systems used to approve loans or set credit scores may be biased against certain groups of people.
- The potential ethical implications of biased AI systems in finance.
- For example, biased AI systems could lead to people being denied access to financial services or could lead to people being charged higher interest rates.

Q&A AND DISCUSSION