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 I SCIENCE DE DONNÉES ET INTELLIGENCE ARTIFICIELLE (SDIA)

DR ILHAM KITOUNI

2023-2024

LESSON 6 : RESPONSIBLE AI DEVELOPMENT

KEY PRINCIPLES OF RESPONSIBLE AI

These principles are based on fundamental values such as justice, fairness, autonomy, and human dignity

- **Transparency:** AI systems should be transparent in their operation and decisions.
- **Accountability:** AI developers and users are accountable for the consequences of their actions.
- **Fairness:** AI systems should be fair and non-discriminatory.
- **Sustainability:** AI systems should be sustainable from an environmental and social perspective.

EXAMPLES OF ORGANIZATIONS ADOPTING RESPONSIBLE AI DEVELOPMENT PRINCIPLES

Many organizations, both public and private, have adopted responsible AI development principles. These organizations include:

- The World Economic Forum
- The Organization for Economic Co-operation and Development (OECD)
- The European Union
- Google
- Microsoft

INCORPORATING PRINCIPLES INTO AI DEVELOPMENT

- **Importance:** It is essential to incorporate ethical considerations into the development of AI systems to ensure that they are used **for good and not for harm.**
- **Methods:** There are many methods that can be used to incorporate ethical principles into AI development, including
  - **Risk assessment:** This involves identifying and mitigating the potential risks associated with AI systems, such as bias, discrimination, and misuse.
  - **Inclusive design:** This involves involving a diverse range of stakeholders in the design process to ensure that AI systems meet the needs of all users.
  - **User training:** This involves educating users about the functioning and limitations of AI systems to help them use them safely and responsibly.

IN AI DESIGN

Design decisions made early in the development process can have a significant impact on the ethics of AI systems.

For example, the following design decisions can have ethical implications:

- Data selection
- System objectives
- User interface design

IN AI DEVELOPMENT

Ethical considerations in AI development are specific to different phases of the development process.

- Data collection
- Model training
- Deployment

EXAMPLES OF GUIDELINES AND FRAMEWORKS

- **Available:** There are many guidelines and frameworks available to help developers incorporate ethical principles into AI development.
- **Examples:** Some examples include
  - the Alliance for Responsible AI's Code of Conduct for Artificial Intelligence,
  - the OECD's Guidelines for Artificial Intelligence,
  - the European Union's Framework for Responsible AI Development.

EXAMPLE OF OECD'S GUIDELINES

The OECD's Guidelines for Artificial Intelligence (AI) are a set of principles and recommendations that promote the responsible and trustworthy development and deployment of AI.

Developed by the OECD AI Network of Experts, which is a group of government officials, researchers, and industry representatives from over 50 countries.

The Guidelines are based on the following five principles:

1. **Inclusive growth, sustainable development and well-being**
2. **Human-centred values and fairness**
3. **Transparency and explainability**
4. **Robustness, security and safety**
5. **Accountability**

IMPACT OF THE OECD'S GUIDELINES

Some specific examples of how the Guidelines have been used:

- The European Union's draft AI Regulation is based on the OECD's Guidelines.
- The United Kingdom's Centre for Data Ethics and Innovation has used the Guidelines to develop its principles for AI.
- The World Economic Forum's Global AI Action Alliance has used the Guidelines to develop its framework for responsible AI.

TOOLS AND FRAMEWORKS FOR ETHICAL AI DESIGN

There are many tools and frameworks that can help developers incorporate ethical considerations into the design process.

- Examples of methodologies:
  - Ethical impact assessment
- Examples of tools:
  - FAIRNESS 360
  - AI Ethics Toolkit

CASE STUDY OF RESPONSIBLE AI PROJECTS

- Case Study 1: Explainable AI in Healthcare

HOW TO BUILD AN EXPLAINABLE AI SYSTEM

- **Data collection**
  - Patient clinical data
  - Blood test results
  - Medical images
- **Explainable AI model**
  - Uses techniques such as interpretable neural networks or explicit rule-based models
- **Decision explanation**
  - Indicates how the system arrived at its conclusion
  - Uses interactive visualization tools
- **Feedback and correction**
  - Doctors can request additional explanations and provide feedback

BENEFITS OF AN EXPLAINABLE AI SYSTEM

- **Transparency**
  - Doctors understand the reasons behind each recommendation
- **Continuing education**
  - Explanations serve as a continuing education tool
- **Shared decision-making**
  - Explanations facilitate communication between doctor and patient

EXPLAINABLE AI IN HEALTHCARE

Explainable AI systems exist in healthcare, but adoption varies.

- **Examples:**
  - InterpretML (Microsoft)
  - IBM Watson for Oncology
  - EXAIR (research initiative)
  - Google Health's Medical Brain
  - SHAP project

QUESTIONS AND ANSWERS

- What are the benefits of developing and deploying responsible AI?
- What are some of the challenges of developing and deploying responsible AI?

QUESTIONS AND ANSWERS

- **Answer:** Responsible AI can help to ensure that AI systems are used for good and not for harm. It can also help to build trust between AI developers and users.

- **Answer:** One challenge is that it can be difficult to identify and mitigate all potential ethical risks associated with AI systems. Another challenge is that it can be difficult to ensure that AI systems are fair and equitable to all users.

CONCLUSION

- Incorporating ethical considerations into the design process of AI is essential to ensure that AI systems are developed and deployed in a responsible manner.
- Responsible AI projects illustrate different ways to incorporate ethical considerations into AI development.

REFERENCES

- **Explainable AI: Interpreting, Explaining, and Visualizing Machine Learning by Christoph Molnar**
- **Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, *32*(11), 4793-4813.**
- **Explainable Artificial Intelligence for Healthcare: A Roadmap for Research and Development by the European Commission**