Human-Centered Artificial Intelligence

2024/2025

Worksheet #1:
Human-Centered Artificial Intelligence
Introduction

Luís Macedo
Amílcar Cardoso, Hugo Oliveira

2.1 Topics

- AI definition
- History of AI
- AI Paradigms
- Artificial, Intelligent, Autonomous Agents, and Multiagent Systems
- AI Risks
- Requirements for Human-centered AI

2.2 Pre-class Readings

Analyse the following texts (available in the course's repository in UCStudent):

- Luís Macedo. AI Paradigms and Agent-based Technologies. In: Human-centred AI: A Visual Scientific Quest. Springer, 2024.
- Russell and Norvig [2010, chap 26]
- HLEG-AI [2020 (accessed September 15, 2022]

View the following talk:

- Russell [2020 (accessed September 15, 2022]

2.3 Theoretical Exercises

Question 2.1 AI definition

How is Artificial Intelligence (AI) defined?

What are the two main aspects of AI?

What is the relationship between AI and human intelligence?

What are some core components of intelligence?

How do the scientific and engineering aspects of AI relate to each other?

Why is AI considered important in contemporary society?

Question 2.2 History of AI

Who are some historical figures that contributed to the early development of AI?

How has AI evolved historically?

Question 2.3 Artificial, Intelligent, Autonomous Agents, and Multiagent Systems

What are the main models/architectures of artificial intelligent agents?

Question 2.4 AI Paradigms

What are AI paradigms?

What are the four types of AI according to Russell and Norvig?

What is the difference between strong AI and weak AI?

What is symbolic AI, and how does it work?

What is connectionist AI, and how does it differ from symbolic AI?

What is statistical AI, and what role does machine learning play in it?

How do symbolic, connectionist, and statistical AI paradigms relate to each other?

What are Pedro Domingos' five tribes of AI?

How do hybrid AI systems combine multiple paradigms?

What are the key differences between logical and non-logical symbolic AI?

Question 2.5 Risks of AI

What concerns exist regarding the impact of AI?

What is suggested about the future impact of AI on our lives?

Why do some experts advocate for more stringent regulatory measures for AI?

What are catastrophic AI risks?
What are the main categories of AI risks?
What are the risks associated with the malicious use of AI?
What are the risks related to the AI arms race?
How do organizational risks affect AI safety?
What are rogue AI systems, and why are they dangerous?
What is the difference between strong and weak AI in terms of risks?
How can technical safety mechanisms help mitigate AI risks?
What ethical guidelines and governance frameworks are needed for AI?
How does AI pose risks to societal well-being and human rights?
What is the role of global cooperation in mitigating AI risks?
What future risks does AI present, and how should they be addressed?

2.4 Theoretical-Practical Exercises

Question 2.6 Consider the following dilemma: an autonomous car sees a child crossing the road in front of it and realizes it is impossible to stop the car in due time; there are two alternatives: hit the child, running the risk of killing him/her, or swerve to a ravine off the road, running the risk of killing the occupant of the vehicle. How should the autonomous car act?

Question 2.7 Consider the same dilemma. If the car kills someone in such circumstances, who should be responsible for the tragic event? The owner of the car? The maker of the car? The maker of the AI software? The programmer of the AI software? Any other person or entity intervening in the process?

Question 2.8 If you were to buy an autonomous car, would you prefer to buy one programmed to give priority, in a dilemma situation as the one described above, to save occupants' lives, or to save pedestrians' lives? Taking your answer in consideration, what will be the criterion that car makers will likely adopt for selling their cars?

Question 2.9 Consider section 1.3 of Russell and Norvig [2010] concerning the history of AI. See also the following links with a few current AI developments and applications:

- Voice:
  https://www.descript.com/overdub

- Video:
  https://www.youtube.com/watch?v=ohmajJTcpNk
  and
  https://www.technologyreview.com/2017/05/01/152061/real-or-fake-ai-is-making-it-very-hard-to-know/

- Robotics:
  https://youtu.be/fn3KWM1kuAw
  and
  https://youtu.be/6Zbhvaac68Y

- Painting:
  https://genekogan.com/works/style-transfer/
  and
  https://medium.com/x8-the-ai-community/neural-style-transfer-using-deep-learning-to-gene

Do you think these may be threats to human race? What are the advantages and disadvantages of AI for these kinds of applications?

Question 2.10

A European taxi company decided to install cameras in their cars that capture the image of the passenger seats. These images are analysed by an AI system with the aim of increasing safety both for passengers and taxi drivers. Consider that the system has powerful image recognition capabilities that allow both the identification of common objects and common human gestures and body postures. Figure 2.1 illustrates a situation where a passenger can be immediately warned about a forgotten handbag when leaving the taxi.

Suppose that the taxi company wants the AI software to follow the Ethics Guidelines for Trustworthy AI (see HLEG-AI [2020 (accessed September 15, 2022]).

a) Give an example of a situation where a system like this **could** act in order to ensure the safety of the driver without violating the mentioned guidelines.

b) Give examples of acts or decisions that the AI system **must not** take if it is to conform with the three components of Trustworthy AI (one example for each component).

c) Classify the following sentences as True or False (refer to the relevant Ethical Principles described in HLEG-AI [2020 (accessed September 15, 2022]):

1. The compliance with the principle of explicability is an ethical imperative for this system (True/False). Why?

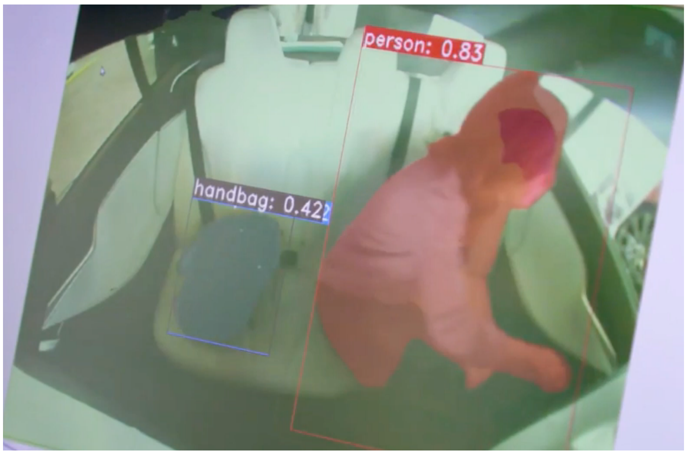

Figure 2.1: The forgotten handbag.

2. There may be tension between the principle of prevention of harm and the freedom of business (True/False). Why?

Question 2.11 What are the technical implications of the HLEG-AI ethics guidelines (see HLEG-AI [2020 (accessed September 15, 2022])?

Question 2.12 What are the main scientific or technological barriers that limit our ability to build AI systems that comply with the HLEG-AI ethics guidelines?

Question 2.13 What are the main non-technical barriers that limit our ability to build AI systems that comply with the HLEG-AI ethics guidelines?

Question 2.14 You are projected 10 years from now, and you find out that Europe is recognized as the major place-to-go for AI systems. What made this happen? What is the state of AI?

2.5 Post-class (optional) Readings

Read the following texts (available in the course's repository in UCStudent):

- Tigard [2020]
- Lauer [2020]
- Anderson and Anderson [2020]
- Turing [1950]

Bibliography

S.L. Anderson and M. Anderson. Ai and ethics. *AI and Ethics*, 2020.

HLEG-AI. *Ethics Guidelines for trustworthy AI, High-level Expert Group on Artificial Intelligence*, 2020 (accessed September 15, 2022). URL https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

Dave Lauer. You cannot have ai ethics without ethics. *AI and Ethics*, 2020.

S. Russell. *How Not to Destroy the World with AI | AAAI 2020*, 2020 (accessed September 15, 2022). URL https://www.youtube.com/watch?v=QPSgM13hTK8.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2010.

Daniel W Tigard. Responsible ai and moral responsibility: a common appreciation. *AI and Ethics*, 2020.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. ISSN 00264423.