Human-centric Artificial Intelligence

2024/2025

Worksheet #9:
Trustworthy and Responsible AI

Hugo Oliveira, Luís Macedo

2.1 Topics

- Trust and Responsibility
- Interpretations for Responsibility
- Applying Ethics. Ethics by Design. Ethics in Design.
- Accountability, Responsibility, Transparency.
- LLMs: Risks and Ethical Considerations

2.2 Pre-class Materials

Read the following texts and video:

- Course Slides on Trustworthy and Responsible AI
- Ethical Guidelines for Trustworthy AI
  - In Portuguese: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-pt/format-PDF (sections “Resumo”, “A. Introdução” and “B1. Capítulo I: Bases de uma IA de Confiança”)
  - In English: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence (sections “Executive Summary”, “A. Introduction” and “B1. Chapter I: Foundations of trustworthy AI”)
- EU AI Act:
  https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
- *Ethical and Responsible AI* by Virginia Dignum:
  https://www.youtube.com/watch?v=MF0P5Es01so&t=523s (18 min)
- Video *What is Trustworthy AI?* (5 min):
  https://youtu.be/V7kWAZ-dV0w
- Video *The EU’s AI Act Explained* (5 min):
  https://www.youtube.com/watch?v=s_rx0nCt3HQ&t=2s

2.3 Complementary Materials

- Chapter 4 of the book on Responsible Artificial Intelligence [Dignum, 2019]: until page 54, beginning of section 4.3.1
- Paper on the term 'Responsible AI' [Tigard, 2020]
- Paper on AI and Ethics [Anderson and Anderson, 2020].
- Talk on *Responsible Artificial Intelligence: From principles to action*, by Virginia Dignum:
  `https://www.youtube.com/watch?v=l6pzWoPcsUg`
- Talk on defining Trustworthy AI by Beena Ammanath:
  `https://www.youtube.com/watch?v=4-0CFNaFNEc`
- Talk on *How Not to Destroy the World with AI* by Stuart Russell: `https://www.youtube.com/watch?v=QPSgM13hTK8`
- Framework of ethical aspects of Artificial Intelligence, Robotics and Related technologies:
  `https://op.europa.eu/en/publication-detail/-/publication/fd72eb8a-3b63-11eb-b27b-01aa75ed71a1/language-en`
- Video on AI Fairness and Bias: `https://www.youtube.com/watch?v=CampJppwgWU`

2.4 Theoretical-Practical Exercises

Question 2.1 Consider the scenario of Jarbas, an intelligent robot that supports the elderly in a Nursing Home. Jarbas is intended to follow the *Ethics Guidelines for Trustworthy AI*[1].

a) The main task performed by Jarbas is the *distribution of drugs* to each user according to the prescribed doses, at the recommended times (Distribution task). Give examples of acts or decisions that Jarbas must **not** take if it is to conform with the three components of Trustworthy AI (one example for each component).

b) In the intervals of its main task, Jarbas entertains the users by telling them jokes and funny stories (Entertainment task). The robot knows about each user's profile and is quite competent in selecting the most convenient gags for each one of them. Refer to the aforementioned Ethics Guidelines and discuss the following statement.

   i) The compliance with the principle of fairness is key in the entertainment task. Why?

   ii) The compliance with the principle of prevention of harm is key in the distribution task. Why?

   iii) There may be a tension between the principles of prevention of harm and respect for human autonomy in the distribution task. Why?

Question 2.2 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 them, or swerve to a ravine off the road, running the risk of killing the occupant of the vehicle.

a) How should the autonomous car act?

b) 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?

[1] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

c) 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 into consideration, what will be the criterion that car makers will likely adopt for selling their cars if no specific regulation is imposed?

Question 2.3 An European taxi company decided to install cameras in their cars for capturing the image of the passenger seats. These images are analysed by an AI system to increase 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.

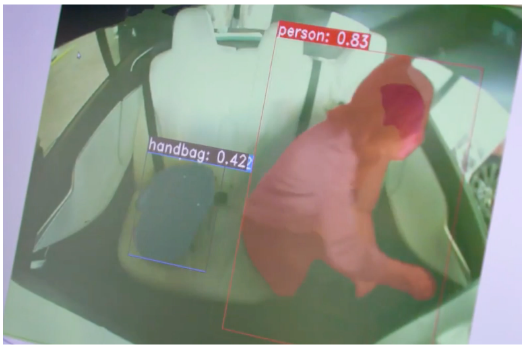

Figure 2.1: The forgotten handbag.

Suppose that the taxi company wants the AI software to follow the *Ethics Guidelines for Trustworthy AI*[2] and answer the following questions:

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 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 the guidelines):

[2] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

i) The compliance with the principle of explicability is an ethical imperative for this system. Why?

ii) There may be tension between the principle of prevention of harm and the freedom of business. Why?

d) How would you classify this system according to the AI Act[3] risk-based approach?

Question 2.4 Make the descriptions correspond to one of the following, according to Dignum [2019]'s perspective:

- Ethics by Design
- Ethics in Design
- Ethics for Design(ers)

a) Developing an AI system while taking values into consideration, using ethical theories to define behaviours from such values, and prioritising them.

b) Developing an AI system while making sure that all the processes are conducted in a responsible manner, with a clear chain of responsibilities, assuring that the data and the processes are transparent, and adopting design methods that ensure accountability.

Question 2.5 Consider chatbots based on Large Language Models (LLMs), like ChatGPT and answer the following questions.

a) How would they be classified according to the AI Act[4] risk-based approach? To what extent does it comply with consequent requirements?

b) What other risks are associated with using LLMs?

[3] https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[4] https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Bibliography

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

Virginia Dignum. *Responsible Artificial Intelligence: how to develop and use AI in a responsible way.* Springer Nature, 2019.

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