6 Managerial, Ethical and Privacy Considerations

6.1 Ethics: Explain the difference between consequentialism and categorical moral reasoning at the example of the trolley problem

- Consequentialism: “The greatest good for the greatest number” (utilitarianism, Bentham 1748–1832)
- Categorical Moral Reasoning: “Act only according to that maxim whereby you can, at the same time, will that it should become a universal law.” (Kant, 1724–1804)

6.2 Ethics: Fairness and Transparency. What was the problem in the dog-vs-wolf and doctor-vs-nurse examples?

- A complex and intransparent classifier (a convolutional neural network) was trained on image data
- Aim was to classify images as either showing a dog or a wolf, or showing a doctor or a nurse
- The classifier seemingly performed well by looking at its error on a test set, but
- The classifier used characteristics of the image that should not have been relevant:
  - dog-vs-wolf: snow/outside area vs. indoor area in the background
  - doctor-vs-nurse: long hair was used as discriminator, because training (and test) set did not include an equal number of doctors with long hair
- Thus, data set needs to be carefully designed not to include biases or problematic attributes

6.3 Trustworthy AI: What are the three components of a trustworthy AI, according to the European Ethics guidelines?

(X) lawful - respecting all applicable laws and regulations

(X) ethical - respecting ethical principles and values

(X) robust - both from a technical perspective while taking into account its social environment

6.4 Trustworthy AI: Explain the 7 key requirements that AI systems should meet in order to be deemed trustworthy?

(X) Human agency and oversight: AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches

(X) Technical Robustness and safety: AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.

(X) Privacy and data governance: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.

(X) Transparency: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

(X) Diversity, non-discrimination and fairness: Unfair bias must be avoided, as it could could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

(X) Societal and environmental well-being: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.

(X) Accountability: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured.

6.5 Privacy: Technically, is it sufficient to anonymise data by deleting obvious identifiers such as name, address, and social security number? Give arguments / provide examples, why this is / isn't the case!

- Not sufficient, as pairing of data records with external data, such as data from population registers, might allow re-identification of individuals.
- Examples: Massachusetts Group Insurance Case

6.6 Privacy: What are shadow profiles? Why do they pose a problem regarding consent?

- Profiles about **non-**users, obtained by combining data that has been disclosed by users
- Non-users did not give consent to processing of their data, might not even know their data has been disclosed
- Examples: Friends in the address books of Social Network users, such as Facebook/Whatsapp/Friendster/...

6.7 Privacy: Explain the differences between the following privacy models

(X) k-Anonymity: grouping into k-undistinguishable records; no consideration that all members of a group might share the same sensitive attribute value, e.g., a positive test result

(X) l-Diversity: expands k-Anonymity by requiring that in each group there are at least l different values for sensitive attributes present (thus, knowing someone belongs to a group doesn't disclose a sensitive attribute value)

(X) t-Closeness: expands l-Diversity by requiring distribution of sensitive attributes to be similar between each group and the whole population

(X) ϵ-Differential Privacy: requires that a single record can be removed without (considerably) changing the outcome an analysis

6.8 Privacy: Explain the following GDPR-related concepts:

- ( ) Concent: must be given explicitly, withdrawal of consent must be as easy as giving
- ( ) Breach notification: within 72 hours of discovery authorities must be notified
- ( ) Right to access one own's personal data: upon request, a person must obtain copies of their data free of charge
- ( ) Right to be forgotten: upon request, a person's data must be deleted
- ( ) Data portability: upon request, a person should obtain their data in a commonly used and machine-readable format (e.g., for re-using it at another institution)
- ( ) Privacy by design: collect only the data necessary to complete one's business, and restrict access to those who need it
- ( ) Record keeping and data protection officers: documentation of collection/processing/storage of personal data, and point-of-contact for internal/external GDPR-related requests

Appendix

References

[Ala-Pietila, 2019] Ala-Pietila, P. e. a. (2019). Ethics guidelines for trustworthy AI.

[European Parliament and Council of the European Union, 2016] European Parliament and Council of the European Union (2016). European general data protection regulation (GDPR).