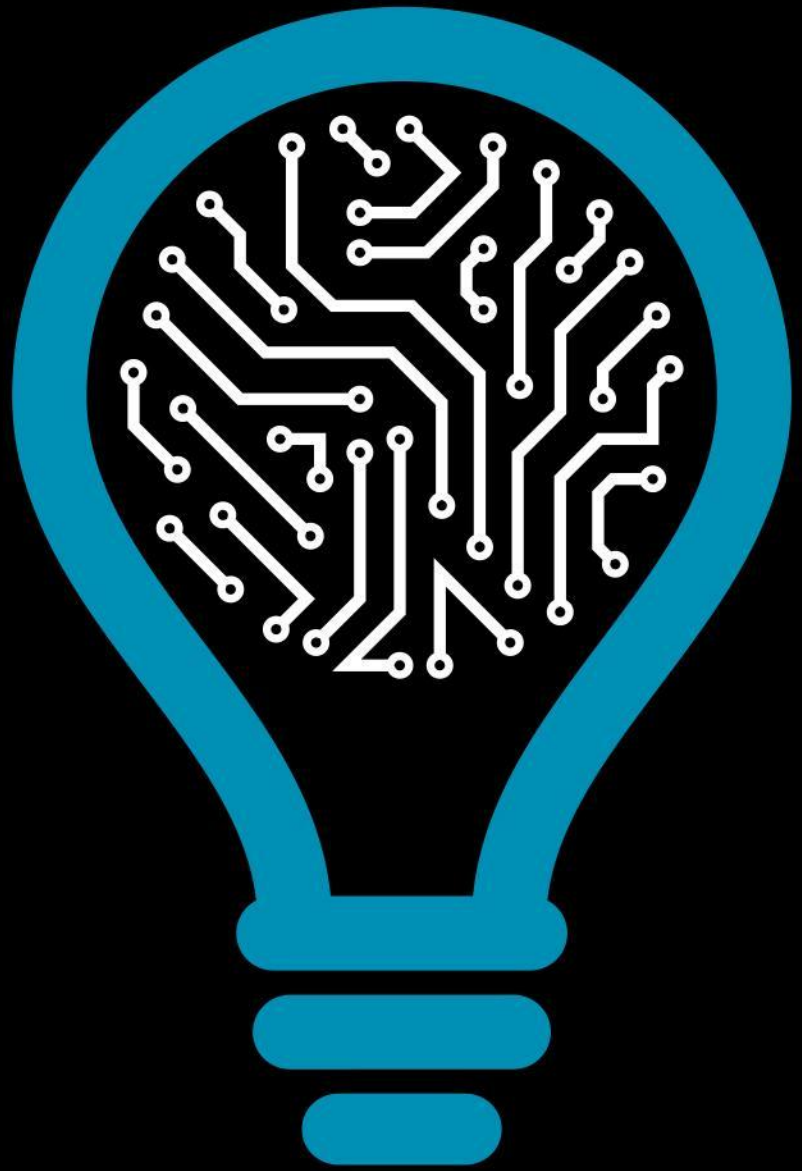
Responsible AI - Lecture 2



TAIA - Advanced Topics on Artificial Intelligence

Tiago Filipe Sousa Gonçalves

tiago.f.goncalves@inesctec.pt | tiagofs@fe.up.pt

INESCTEC

INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE

INESCTEC

Outline

1. Principles of Responsible AI (beyond Interpretability)

2. Law meets AI: Friend or Foe?

3. Take-home messages and further readings

INESCTEC

1. Principles of Responsible AI (beyond Interpretability)

Responsible AI relies on fundamental principles

- **Responsible AI** is a framework that guides how we should address the challenges around artificial intelligence from both an **ethical, technical and legal** point of view[1]
  - We must resolve ambiguity for where responsibility lies if something goes wrong!

- This framework relies on fundamental principles[2]:
  - **Accountability**
  - Interpretability
  - **Fairness**
  - **Safety**
  - **Privacy**

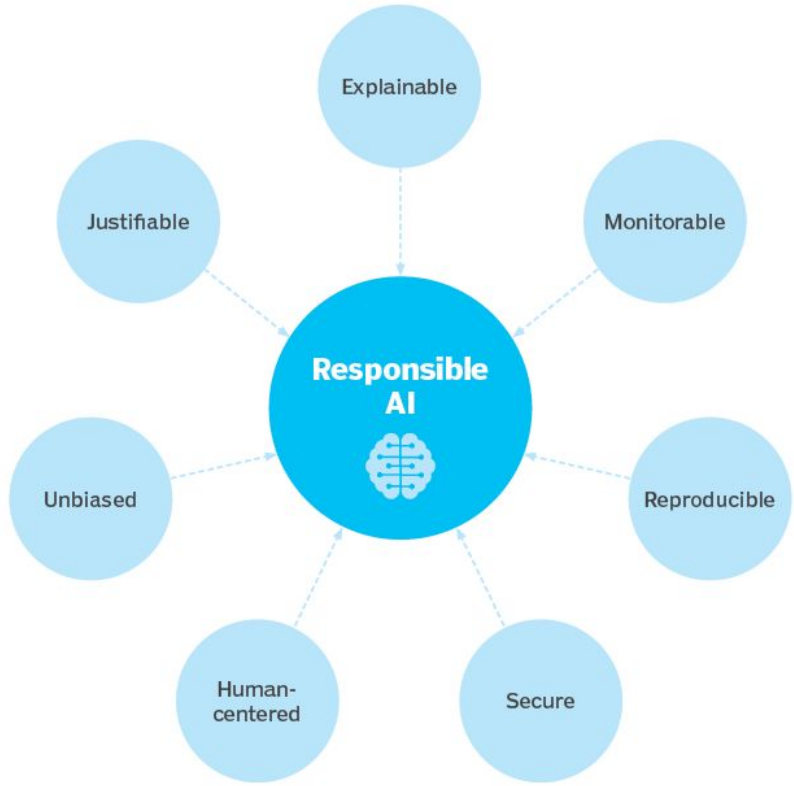


Figure - Responsible AI (Image from [1])

Sources: [1] https://searchenterpriseai.techtarget.com/definition/responsible-AI, [2] https://towardsdatascience.com/what-is-responsible-ai-548743369729

Accountability: who will take the responsibility?[1]

- **People** should be **accountable** for AI systems[1]
  - This principle is the **baseline**, hence, all the other principles can be seen as **branches**

- Some important ideas are:
  - Implement and use a **human-centered design approach**[2]
  - Identify **multiple metrics to assess training and monitoring**, and ensure that these metrics are **appropriate for the context and goals** of our system[2]
  - Examine our **raw data** (e.g, missing values, incorrect labels, biases, feature redundancy)[2]
  - Understand the **limitations** of our databases and models[2]
  - Learn **best practices from software quality engineering** to make sure the AI system is working as intended and can be trusted[2]
  - Continue to **test**, **monitor** and **update** the system after deployment[2]

Sources: [1] https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6, [2] https://ai.google/responsibilities/responsible-ai-practices/?category=general,

Fairness: how to deal with bias?[1]

- AI models learn from existing **data collected from the real world**, and so an accurate model **may learn** or **even amplify problematic pre-existing biases** in the data based on **sensitive characteristics**[2]

- Regarding this issue, we should:
  - Interact with **social scientists**, **humanists**, and **other relevant experts** for our product to understand and account for various perspectives[2]
  - Consider how the technology and **its development over time will impact different use cases** (e.g., what outcomes does this technology enable)[2]
  - **Assess fairness in our datasets** (e.g., identifying representation and corresponding limitations)[2]
  - Check the system for **unfair biases**[2]
  - **Analyse the performance** of the system, taking into account **the different metrics** we've defined[2]

Sources: [1] https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6, **[2]** https://ai.google/responsibilities/responsible-ai-practices/?category=fairness,

Safety: achieving reliable and safe AI systems[1]

- Safety and security intend to ensure that AI systems **behave as intended**, **regardless** of how **attackers try to interfere**[2]

- Regarding this issue, we should:
  - Consider if **there are incentives to make the system misbehave**[2]
  - Identify what **unintended consequences** would result from the **system making a mistake**, and assess the **likelihood and severity of these consequences**[2]
  - Build a **rigorous threat model** to understand **all possible attack vectors**[2]
  - Research into **adversarial machine learning**, as it continues to offer **improved performance for defenses**[3] and **provable guarantees**[2, 4]
  - Check if **there are other vulnerabilities** in the AI supply chain[2, 5]

Sources: [1] https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6, [2] https://ai.google/responsibilities/responsible-ai-practices/?category=security, [3] Kannan et al. "Adversarial Logit Pairing", [4] Wong and Kolter "Provable defenses against adversarial examples via the convex outer adversarial polytope", [5] Gu et al. "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain"

Privacy: can AI reveal aspects of its training data?[1]

- AI systems must **prioritise and safeguard consumers' privacy** and **data rights** and provide **explicit assurances** to users about how their personal data will be used and protected[1]

- Regarding this issue, we should:
  - Identify whether our AI model can be trained **without the use of sensitive data**[2]
  - **Anonymise** and **aggregate** incoming data using best practice data-scrubbing pipelines (e.g., removing **personally identifiable information (PII)** and outlier or metadata values that might **allow de-anonymisation**)[2]
  - Train our models using **federated learning**[3], where a fleet of devices coordinates to train a shared global model from locally-stored training data
  - Perform tests based on **"exposure" measurements**[4] or **membership inference assessment**[5] to estimate whether our model is **unintentionally memorising** or **exposing sensitive data**

Sources: [1] https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6, [2] https://www.ibm.com/artificial-intelligence/ai-ethics-focus-areas#2652220, [3] https://ai.google/responsibilities/responsible-ai-practices/?category=privacy, [4] https://research.googleblog.com/2017/04/federated-learning-collaborative.html, [5] Carlini et al. "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks", [6] Shokri et al. "Membership Inference Attacks against Machine Learning Models"

2. Law meets AI: Friend or Foe?

Why the EU got it right: regulating data before AI!

- The **General Data Protection Regulation (GDPR)**[1] is a privacy and security law drafted and passed by the European Union (EU), that imposes obligations onto organizations anywhere
  - This goes back to the **right to privacy**, part of the **1950 European Convention on Human Rights**: “Everyone has the right to respect for his private and family life, his home and his correspondence”[2]

- Moreover, the EU-GDPR is all about **people’s privacy rights**[1]:
  - The right to be informed
  - The right of access
  - The right to rectification
  - The right to erasure
  - The right to restrict processing
  - The right to data portability
  - The right to object
  - Rights in relation to automated decision making and profiling

Sources: [1] https://gdpr.eu/, [2] https://www.echr.coe.int/Documents/Convention_ENG.pdf

Can we trust AI? Towards “Trustworthy AI” in the EU

- The European Commission appointed a group of experts to provide advice on its artificial intelligence strategy: **High-Level Expert Group on AI**[1]

- **According to the Guidelines, trustworthy AI should be:**
  - **Lawful**: respecting all applicable laws and regulations
  - **Ethical**: respecting ethical principles and values
  - **Robust**: both from a technical perspective while taking into account its social environment

- Several important **guidelines were proposed**[2]:
  - **Human agency and oversight**: AI systems should empower human beings
  - **Technical Robustness and safety**: AI systems need to be resilient and secure
  - **Privacy and data governance**: data governance mechanisms must be ensured
  - **Transparency**: the data, system and AI business models should be transparent
  - **Diversity, non-discrimination and fairness**: AI systems should be accessible to all
  - **Societal and environmental well-being**: AI systems should benefit all human beings
  - **Accountability**: ensure responsibility and accountability for AI systems and their outcomes

Sources: [1] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai, [2] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

The AI Act and how it will impact our lives

- The AI Act[1] is a **document proposed by the European Commission** that contains several **harmonised rules**[2] regarding **AI applications**, emphasising that its approach is shaped by EU values and **risk-based**, ensuring both **safety** and **fundamental rights protection**

- What does the AI Act propose?[2]
  - **Prohibition of unacceptable AI practices** (e.g., social scoring)
  - **Regulation of high-risk AI systems** (e.g., AI used in the context of recruitment)
  - **Conformity assessmen**t (i.e., under the EU product safety framework)
  - **Transparency obligations for potentially deceptive AI systems**
  - **Ex post market surveillance** (i.e., post-market monitoring system)
  - **Governance** (i.e., authorities must be appointed for the application and implementation)
  - **Pre-emption of national AI regulatory frameworks** (i.e., regulated by the EU)
  - **Monitoring and enforcement** (i.e., done by the Member States)
  - **Compliance with the prohibitions and regulatory requirements**

Sources: [1] https://www.eipa.eu/publications/briefing/the-artificial-intelligence-act-proposal-and-its-implications-for-member-states/, [2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

3. Take-home messages and further readings

A (tentative) fair and accurate summary of this lecture

- **The development of data-driven artificial intelligence applications is pushing the limits of the applications of these algorithms in our lives:** this rapid evolution motivated the need to ethical, legal and technical regulatory frameworks based on specific principles: **accountability**, **interpretability**, **fairness**, **safety**, **privacy**

- At the European level, there have been several proposals to regulate data and AI-based applications: **EU-GDPR**, **Trustworthy AI Initiative** and **AI Act**

- Open regulatory challenges will focus on the impacts of AI in **ethics**, **transparency**, **fairness**, **safety**, **sociology** and **sustainability**[1, 2, 3, 4]

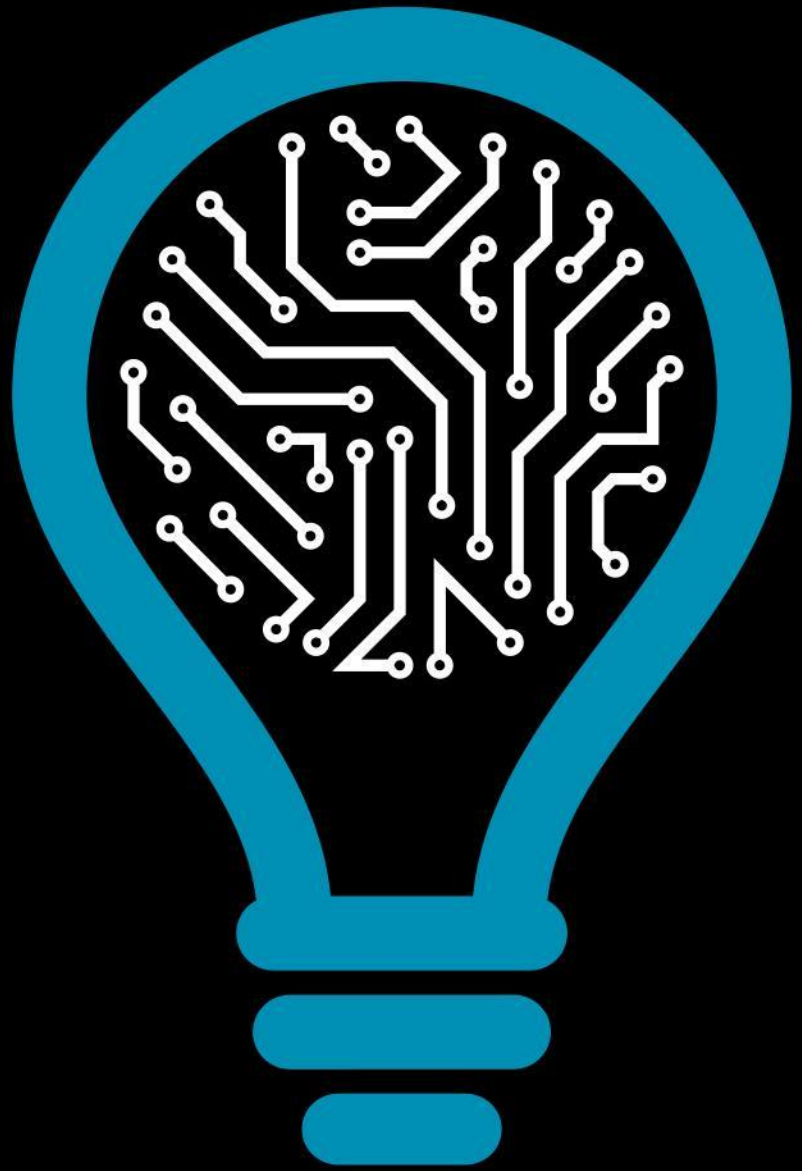
- **Multidisciplinary work** is, more than ever, of utmost importance and useful!

Sources: [1] https://www.forbes.com/sites/cognitiveworld/2020/02/20/ai-laws-are-coming/?sh=655a2497a2b4, [2] https://www.wsj.com/articles/google-pushes-sensible-ideas-for-how-to-regulate-ai-11579521003?mod=article_inline, [3] https://www.nytimes.com/2021/04/16/business/artificial-intelligence-regulation.html?smtyp=cur&smid=tw-nytimes, [4] https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics

Further readings...

- Carlini et al. “The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks”
- Shokri et al. “Membership Inference Attacks against Machine Learning Models”
- Papernot et al. “Scalable Private Learning with PATE”
- Kannan et al. “Adversarial Logit Pairing”
- Wong and Kolter “Provable defenses against adversarial examples via the convex outer adversarial polytope”
- Gu et al. “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain”
- Montenegro et al. “Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis”
- Pessach and Shmueli “Algorithmic Fairness”
- https://www.ibm.com/blogs/research/2019/09/adversarial-robustness-360-toolbox-v1-0/
- https://www.forbes.com/sites/anniebrown/2021/07/02/ais-role-in-the-future-of-data-privacy/?sh=b23e83918c0d

Responsible AI - Lecture 2



TAIA - Advanced Topics on Artificial Intelligence

Tiago Filipe Sousa Gonçalves

tiago.f.goncalves@inesctec.pt | tiagofs@fe.up.pt

INESCTEC

INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE