Trustworthy AI: Pillars and Requirements

Luís Macedo

University of Coimbra

September 23, 2024

Table of Contents

Introduction to Trustworthy AI

- The European Commission's High-Level Expert Group (HLEG) on AI defines **trustworthy AI** based on three fundamental pillars:
  1. **Lawfulness**: AI systems must comply with applicable laws and regulations.
  2. **Ethics**: AI must ensure ethical principles such as fairness, non-maleficence, and respect for human rights.
  3. **Robustness**: AI systems should be technically robust and secure to avoid unintentional harm.
- Trustworthy AI ensures that systems are safe, reliable, and respect fundamental rights.

Pillar 1: Lawfulness

- **Lawfulness** ensures that AI systems comply with all relevant laws, regulations, and legal frameworks.
- Legal compliance includes respecting:
  - **Privacy and Data Protection** laws (e.g., GDPR in Europe).
  - **Non-discrimination** and **equality**.
  - **Consumer protection**, intellectual property, and other regulatory requirements.
- AI should follow existing laws and contribute to establishing new legal standards where necessary.

Pillar 2: Ethics

- AI must uphold **ethical principles**:
  - **Fairness**: AI should avoid biases and treat individuals equally.
  - **Autonomy**: Respect for human decision-making and control over AI systems.
  - **Transparency**: AI decisions should be explainable and understandable.
  - **Non-maleficence**: AI must not cause harm to individuals or society.
- Ethical AI promotes **human dignity** and **societal well-being**.

Pillar 3: Robustness

- **Robustness** ensures that AI systems function reliably in a range of environments.
- AI systems should:
  - Be **technically robust** and capable of handling errors or unexpected situations.
  - Ensure **security** by protecting against adversarial attacks and vulnerabilities.
  - Incorporate mechanisms for **fallback plans** or human intervention in critical situations.
- Robust AI systems minimize risks of failures, malfunctions, and exploitation.

Seven Key Requirements for Trustworthy AI

- The **European Commission's High-Level Expert Group (HLEG)** outlines seven core requirements for achieving trustworthy AI:
  1. **Human Agency and Oversight**
  2. **Technical Robustness and Safety**
  3. **Privacy and Data Governance**
  4. **Transparency**
  5. **Diversity, Non-discrimination, and Fairness**
  6. **Societal and Environmental Well-being**
  7. **Accountability**
- These requirements address legal, ethical, and technical dimensions.

Requirement 1: Human Agency and Oversight

- AI should empower individuals, giving them the ability to make informed decisions.
- AI systems must:
  - Respect human **autonomy**.
  - Incorporate mechanisms for **human oversight** (e.g., human-in-the-loop, human-on-the-loop).
  - Support decision-making, without undermining human control.

Requirement 2: Technical Robustness and Safety

- AI systems must be resilient and secure.
- Key aspects:
  - Ensure **resilience** to attacks, errors, and unexpected conditions.
  - Incorporate **fallback plans** or safe modes.
  - Provide **accuracy** and **reliability** across various tasks.

Requirement 3: Privacy and Data Governance

- Ensure AI systems respect **privacy** and provide secure handling of data.
- Key elements:
  - Uphold **data minimization** principles.
  - Provide **user control** over their data.
  - Comply with data protection laws (e.g., **GDPR**).

Requirement 4: Transparency

- AI systems must be **transparent** about their capabilities and limitations.
- Requirements:
  - Ensure **explainability** of AI decisions and processes.
  - Disclose when AI is being used (e.g., in decision-making systems).
  - Provide documentation and **traceability** of the system's actions.

Requirement 5: Diversity, Non-discrimination, and Fairness

- AI must ensure **fair treatment** of all individuals and avoid discrimination.
- Key aspects:
  - Avoid **bias** in AI algorithms and training data.
  - Promote **inclusive design** that reflects diverse users.
  - Ensure that AI systems do not reinforce existing inequalities.

Requirement 6: Societal and Environmental Well-being

- AI should contribute to the well-being of society and the environment.
- Key considerations:
  - AI systems should align with the **Sustainable Development Goals (SDGs)**.
  - Minimize **environmental impact** of AI, including energy use.
  - Promote **ethical use** of AI for social good (e.g., in healthcare, education).

Requirement 7: Accountability

- AI systems must be designed with clear mechanisms for **accountability**.
- Key aspects:
  - Ensure **responsibility** for AI decisions (e.g., audits and assessments).
  - Provide **traceability** of AI processes and outcomes.
  - Guarantee legal recourse and clear responsibility in case of harm or misuse.

Conclusion

- Trustworthy AI is based on the pillars of **lawfulness**, **ethics**, and **robustness**.
- The seven key requirements proposed by the European Commission's High-Level Expert Group (HLEG) provide a comprehensive framework for building AI systems that are safe, fair, and beneficial for society.
- Ensuring that AI systems are lawful, ethical, and robust will foster trust and innovation in the AI ecosystem.

Literature References

- European Commission, "Proposal for a Regulation laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act)," COM(2021) 206 final, 2021.
- European Commission High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," 2019.
- UNESCO, "Recommendation on the Ethics of Artificial Intelligence," 2021.
- United Nations, "Transforming our world: the 2030 Agenda for Sustainable Development," 2015.
- European Union, "General Data Protection Regulation (GDPR)," Regulation (EU) 2016/679, 2016.
- Luciano Floridi, "Ethics of Artificial Intelligence: The Role of Principles," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2019.