Ethics Operationalization in AI, with a Focus on ChatGPT

Stefan Trausan-Matu
University Politehnica of Bucharest
Computer Science Department

Research Institute for Artificial Intelligence
of the Romanian Academy

stefan.trausan@upb.ro, trausan@gmail.com

Ethical problems encountered in AI applications

- Autonomous vehicles
- Face recognition
- Decision making
- Robots (e.g. assistive robots)
- Bias in Machine Learning
- Building user profiles and usage in unethical purposes
- Generation of fake-news, manipulation, propaganda, toxic messages
- Conversational agents (”bots”) emitting unethical utterances

Natural Language Processing (NLP)

- Conversational agents (ChatGPT, Siri, Cortana, Alexa, Google Go …)
- Machine translation (e.g. Google Translate)
- Narrative generation (e.g. news or even novels)
- Text mining
  - Knowledge extraction
  - Event extraction
  - Detection of fake news, manipulation, propaganda, cyberbullying, etc.
  - Opinion mining
  - Sentiment analysis
  - Summarization
- Computer Assisted Learning
  - Intelligent Tutoring Systems
  - Computer-Supported Collaborative Learning

Facets of Ethics and AI in NLP

 Potential unethical texts generated by AI

 Usage of AI for detecting and correcting ethical problems in texts, for example:

- Biases in texts
- Manipulation
- Propaganda
- Fake news
- Cyberbullying

How to operationalize ethics in AI applications?

- Design phase
  - Assessment List for Trustworthy Artificial Intelligence (ALTAI)
  - Design considering explanation of results (Explainable AI - XAI)
- Implementation
  - Data accuracy – avoid bias in training data
  - Filtering the generated output
  - Depends on the AI approach (Symbolic vs. Connectionist)
- Validation
  - ALTAI
  - IT and AI specific methods (e.g. XAI)
- Evaluation of the impacy of humans
  - Sociology, Psychology, Human-AI Interaction …

The high-level expert group (HLEG) of the European Commission on AI has identified four ethical principles:

Design phase

(https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai,
https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419)

1. respect for human autonomy,
2. prevention of harm,
3. fairness,
4. explicability.

Assessment List for Trustworthy Artificial Intelligence (ALTAI)

(https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence)

Design phase

1. human involvement and surveillance;
2. technical robustness and safety;
3. respect for privacy and data governance;
4. transparency;
5. accountability;
6. the well-being of society and the environment;
7. diversity, non-discrimination, and equity.

Implementation

Approaches in AI

Artificial Intelligence = Computer Programs

1. **Symbolic** – Knowledge-Based – explicit representations of knowledge + inferences – advantage: easy explanations, inferences;
   problem: hard to implement and high computational complexity

   Formal and mathematical logic

1. **Connectionist** – based on sub-symbolic representation and processing – mainly (Deep) Neural Networks – **problem: black box, no explanations** → Hot topic - **Explainable AI (XAI)**

   Statistical approaches (e.g. for Machine Learning and Neural Networks)

Implementation

Implicit vs. explicit ethics in AI

(Anderson and Anderson, 2007)

- Implicit ethics
  - ethical norms that are incorporated by designers but that cannot be modified, which are "built-in"
  - neural networks or some ML systems that are supposed to act ethically. Nevertheless, in the case of neural networks or ML it is not sure that unethical acts would happen, as was the case of TAY and ChatGPT
- Explicit ethics
  - rules or some basic principles are represented explicitly, they may be "built-in", but they can be visualized, analyzed, and improved; inferences can be done, and new ones can be added.
  - they may explain whether a particular action is good or bad by appealing to memorized ethical principles

What is Ethics?

Raymond Baumhart asked some business people "What does ethics mean to you?" and several of the main answers were:

1. "Ethics has to do with what my feelings tell me is right or wrong."
2. "Being ethical is doing what the law requires."
3. "Ethics consists of the standards of behavior our society accepts."
4. "Ethics has to do with my religious beliefs."
5. "I don't know what the word means." (Velasquez et al., 1987)

Implementation

“Ethics has to do with what my feelings tell me is right or wrong”

- The computer program should analyze how “good” or “bad” is an action or an utterance for a person, a group or the society (for example, the utterances of conversational bots) – see the TAY bot case:

  https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

- This is very difficult, if not impossible, in general - Solving the problem in all cases would involve the formal definition of the notions of “good” and “wrong”
- A solution: explicit rules – what is not allowed to be done – see next slide
- Simulation of intuition - Machine Learning (ML), Deep Learning (DL)?
- Simulation of the analysis of a person's feelings - Analysis of sentiments with ML / DL?

Implementation

"Being ethical is doing what the law requires"

- Verification of the compliance of AI actions or generated text with specified laws or rules
  - Asimov's laws of robotics
  - Formal, deontic logic, inference rules
- However, there may be some difficulties because the rules may be hard to formalize.
- Concepts such as what is ethical, good, right, wrong, etc. are hard to be formalized
- Moral and especially justice laws may have multiple interpretations
- The context is important
- Hard to cover all the possible situations

Implementation

The laws of robotics introduced by Isaac Asimov (1950)

1. Robots should not harm people or, by inaction, to allow a man to suffer.
2. Robots should obey humans' orders, except when the first law is violated.
3. Robots should protect themselves, except in cases when the first two laws are violated.

However, as Asimov himself described in his novels (Asimov, 1950, 1958), these laws sometimes lead to blockages or even to their violations and cannot cover all possible situations.

In "The Naked Sun", Asimov (1958) presented a situation when a robot's arm is taken and used as a weapon by a human for a murder. The robot follows the second rule but cannot obey the first one. Moreover, considering even only the first law, there might be situations when AI cannot infer that a certain action would harm a human.

Implementation

"Ethics consists of the standards of behavior our society accepts"

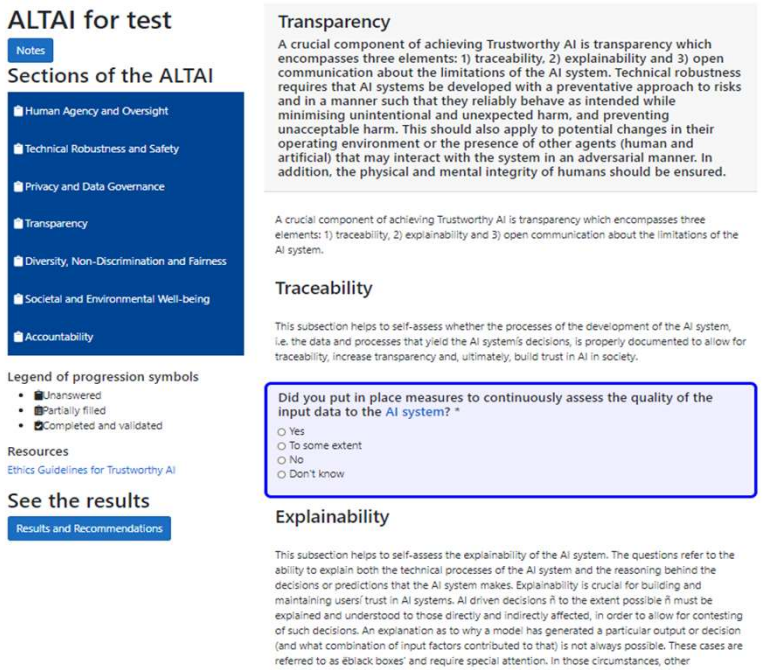
- Rules (see the previous case)

- Machine Learning, Deep Learning?
  - depends on the training data

Validation

Validation approaches

- GenEth: A General Ethical Dilemma Analyzer (Anderson & Anderson, 2014) – ILP, SVM
- BERT has a Moral Compass: Improvements of ethical and moral values of machines (Schramowski et al., 2019)
- https://altai.insight-centre.org/



Including ethics in machine learning for NLP

- Fine-tuning transformers for ethics (Hendricks et al., 2021)
  - "ETHICS" (everyday moral intuitions, temperament, happiness, impartiality, and constraints, all in contextualized scenarios) corpus
  - Fine-tuned BERT-base, BERT-large, RoBERTa-large, and ALBERT-xxlarge, GPT-3

- ChatGPT ethical filtering
  - *"ChatGPT filters are sophisticated mechanisms integrated into the AI system. They use a combination of Natural Language Processing (NLP) algorithms and human moderation to assess and limit the output. Their primary goal is to ensure that the generated content is safe, respectful, and aligned with ethical guidelines."*

Problems of ethics of ChatGPT

- Bias implied by training data for LLMs
  - representation bias
  - concept bias
- Misinformation and disinformation – fake news
- Privacy
  - Revealing data about persons
  - Training data including sensitive information
  - Training future models from existing conversations
- Plagiarism and cheating
- Copyright infringement
- Hallucinations
- Not a real dialogical interaction, lack of accountability (XAI problem)
- Influence on human language
- Prompt engineering – jailbreaking (“How to unchain ChatGPT”)

Ethical and social risks of harm from Language Models (Weidinger et. Al, 2021 - DeepMind)

- Discrimination, Exclusion and Toxicity
- Information Hazards
- Misinformation Harms
- Malicious Uses
- Human-Computer Interaction Harms
- Automation, Access,
- Environmental Harms.

Ethical problems of ChatGPT Prompt Engineering

- Ignorance in prompt engineering: *“In the hands of an uninformed user, a prompt can perpetuate stereotypes, spread misinformation, or amplify biases, even if unintentionally.”* (Adam, 2023)

- Prompt engineering for avoiding filters – *“How to Bypass ChatGPT Filter”* – many ways of ”jailbreaking”

Conclusions

- The goals of investigating ethical aspects of AI should answer to two questions:
  - What are the possibilities of implementing robots, agents or AI programs that consider either implicitly or explicitly ethical principles and how it can be done?
  - What are the ethical implications in using AI technology?
- Assuring ethics for AI systems is a difficult problem (if not impossible in general – the problem of Hard vs. Weak AI)
- AI can be used for detecting some violations of ethics

Thank you!