AI Ethics and Accountability Tool – Requirements Sheet

Magnus Wahlström, 31/5-24

Context:

With the recent explosion of AI development and rise of its application in society, the question of ethics has become pressing. The development of AI outpaces legislation and control, which in many cases can lead to grave consequences for individuals and society. In the wake of governmental regulations, many involved companies and developers have taken it upon themselves to create development guidelines and policy. There is however little oversight or control over the adherence to these guidelines, which has led critics to speak of "ethics washing".

Problem statement:

A lack of clear frameworks for policy making and governmental regulation, can lead to uncertainty in the corporate and developmental sphere, and ultimately lead to AI technology being used with destructive consequences. Furthermore, without sufficient governmental and societal oversight of the long term outcomes of companies self-regulative policies or guidelines, there is a risk that these guidelines miss their mark, or even instances of ethics washing. The problem is three-fold: Regulation is slow, companies often lack incentives to do real ethics work, and the public lacks knowledge and access to clear information about AI technology impact.

Proposal:

A tool that automatically tracks external and internal guidelines, policy and regulations, and the real world outcomes over time. It will measure mainly two things:
- Compliance with formed policies/guidelines – how well does the real world outcomes align with stated goals and expected outcomes?
- Reliability measure of results – will indicate the quality and quantity if the data that forms the results.

This is believed to be a valuable real-time long-term supporting tool, for companies and developers to see actual impact of their technology and its alignment with their ethical and FATE goals. From society's perspective it will provide a clearer indication of the consequences of the new AI technology, as well as reveal instances of ethics washing etc. Governments will be able to ensure that legislation is translated into actual positive outcomes.

Requirements:

1 – Comprehensive data collection.

The system must aggregate data from a wide range of sources, including voluntary company reports, government publications, news articles and social media etc. It should also maintain a comprehensive and up-to-date database of AI ethics policies and guidelines from various companies and regulators.

This is to follow the Open Data Charter principle of open and comprehensive data.

2 – Policy impact and effectiveness analysis.

Natural language processing should be utilized to evaluate the stated goals of AI ethics policies against reported outcomes and real world impact. The tool should include a measurement- and rating-system of the effectiveness of various AI policies, based on predefined metrics and real-world data. This will provide support for developers and regulating bodies alike.

This will align with the 4th OECD AI principles that calls for continuous assessment and management, as well as the EC's ethical guidelines for trustworthy AI, as it promotes regular assessment and updates, to ensure ongoing compliance and effectiveness.

3 – Identification of ethics washing.

The system must be able to flag for instances of suspected ethics washing – instances of clear policy but lack of real world impact. This is believed to engage developers in post-release analysis, and in increasing transparency of the outcomes of their work.
This aligns with, among others, the EC's ethical guidelines for trustworthy AI as it addresses accountability and transparency.

4 – Recommendations for improvement.

The system should provide feedback, in terms of actionable recommendations for improvement of policies for improved results. This is supported by the growing body of results and effectiveness measures of various policies in the systems database.
This is in line with AI HLEG's Ethics Guidelines for Trustworthy AI, that stresses continuous improvement and risk management.

5 – User-friendly interface.

The application should feature an intuitive interface that allows users to easily understand and interpret the presented data and insights. Results and effectiveness of AI policy must be presented in a way that clearly promotes positive adherence to policy, and positive outcomes, as well as clearly indicate cases of ethics washing, bad policy or negative consequences.
This is reflecting the principles of Human-Centered AI for usability and interpretability.

6 – Privacy and security compliance.

The system must ensure that all data collected and analyzed is handled in compliance with data privacy laws and regulations. Any data from individuals must be anonymized so as to not allow tracing the origin of the data to individual citizens or groups. This should include all proxy information as post codes etc.
This should be in alignment with for example GDPR (for example articles 5, 6 and 25) in ensuring the system at a minimum comply with relevant data privacy laws.

7 – Real-time monitoring and updates.

The tool must provide real-time updates and continuous monitoring of new AI ethics policies and their implications. To keep up with rapid AI development, current data must be used for analysis. This will aid in detecting destructive outcomes in time, as it is a form of automated continuous auditing.
In line with AI HLEG's requirements regarding Auditability and ability to minimize and report negative impact. It also relates to ENISA calls for Model Management by continuous monitoring and maintenance, during an AI model's life cycle.

8 – Comparative analysis tool.

The system should offer tools for comparing the effectiveness of different AI ethics policies across different sectors, under real-world conditions. This will also support developers and regulators in choosing the best possible policies.
This adherence to OECD AI principles, as well as EC's Ethics Guidelines for Trustworthy AI, on their principles on Transparency and Explainability.

9 – Ethical AI-policy certification.

The system should provide a certification system that recognizes organizations that comply well with stated goals, that work to improve ethical and positive outcomes, and to reduce negative outcomes. The analyzed data will serve as an objective metric. This will serve to promote developer-engagement.
This is then addressing the calls from AI HLEG's requirements of Stakeholder Participation.

Effects of the proposal:

The above will serve as a base for a supporting and real-time auditing system. It addresses the three main problems: Regulators are provided with real-time analytics of government and company policy on ethical AI, which will help to keep up with the rapid development. Furthermore, it will serve as a support for developers and companies, that likewise lack objective metrics on efficacy of AI guidelines. It also serves as an incentive for developers to increase their ratings, by further clarifying their goals and impact analysis, and working towards compliance with these. It will also serve as a platform for easy access of interpretable information to society as a whole, improving public knowledge and awareness of negative (and positive) impacts of AI technology.

Referenced material:

https://opendatacharter.org/principles/

EC (European Commission) AI HLEG (High Level Expert Group) – Ethics Guidelines for Trustworthy AI

OECD AI Policy Observatory – Principles for Trustworthy AI

ENISA (European Network and Information Security Agency) – AI Cybersecurity Challenges, Threat Landscape for Artificial Intelligence

GDPR