Ethics, trust, and explainability in artificial intelligence (AI).

Breaking the Black Box of AI and Understanding AI Ethics

As artificial intelligence (AI) continues to evolve at a rapid pace, becoming part of more and more business processes, and as regulatory and customer pressures continue to mount, we need to address some very deep ethical issues. AI designers and developers in particular can help reduce bias and discrimination by addressing these areas of ethical consideration and using supporting toolkits.

If you were told that your loan application had been rejected without any discernible reason, would you accept that decision? And if you knew that an autonomous vehicle could be manipulated to misinterpret speed signs, would you drive it? Certainly not.

We humans live by communally established ethical norms that are enforced by laws, rules, social pressure, and public discourse. Although ethics and values may vary over time and between cultures, it has played a critical role in decision-making since early human civilization.

The issue of ethics is not new in business either. But in an age when artificial intelligence (AI) is rapidly evolving, moving into more and more business processes and supporting decision-making, we need to address very deep, ethical questions without delay.

Ethical stumbling blocks of AI

In 2019, customer complaints surfaced accusing Apple's credit scoring algorithm for Apple Card applicants of gender discrimination. And security researchers at McAfee used a simple trick to trick Tesla's intelligent cruise control. To do so, the researchers stuck a 2-inch strip of tape on a 35-mph speed sign (making the middle part of the 3 a little longer), and the car's system misinterpreted it as 85 mph and adjusted its speed accordingly.

Responsible use of data has thus become central to competitive advantage.

While consumers are concerned about societal issues such as shared prosperity, inclusion, and the impact of AI on employment, companies are focused on organizational impacts, such as:

- Regulatory authorities as well as the European Union are working on corresponding **legal frameworks** that are increasingly obligating companies. On April 21, 2021, the European Commission presented the first ever legal framework for AI, which attempts to divide AI into risk classes.
- If a model unfairly discriminates against a certain group of customers, it could cause serious **reputational damage**.
- Transparent decision making builds **trust with customers** and increases their willingness to share data. For example, 81% of consumers say they have become more concerned about how companies use their data over the past year.

Example: IBM Ethics Guidelines

Accordingly, many companies have already imposed their own guidelines for ethical AI. IBM, which has been striving for responsible innovation for more than 100 years, circumscribes the following five focus areas when developing responsible AI systems:

- **Accountability**: AI designers and developers are responsible for ethical AI systems and their outcomes.
- **Value Alignment**: AI should be designed to align with the norms and values of your user group.
- **Explainability**: AI should be designed so that humans can understand its decision-making process.
- **Fairness**: AI must be designed to minimize bias and promote inclusion.
- **User data rights**: AI must be designed to protect user data and retain user control over data access and use.

Future of ethical AI systems

The criteria and metrics for ethical AI systems will ultimately depend on the industry and use case in which they are deployed. But AI designers and developers can help reduce bias and discrimination by addressing these five areas of ethical consideration.

AI systems must remain flexible enough to be continually maintained and improved as ethical issues are discovered and addressed. Various dashboards or toolkits available for development can assist in the process. Such as “AI Fairness 360” an open-source toolkit that can help investigate, report and mitigate discrimination and bias in AI models. Or the open source toolkit “AI Explainability 360” that can help with explainability of AI algorithms.

Summary

Ethical decision making is not just another form of technical problem solving, but must be embedded in the AI design and development process from the beginning. Ethical, human-centered AI must be designed and developed to be consistent with the values and ethics of the society or community it affects.

For companies using AI, this must be a top priority. It must be ensured that every employee understands the risks and feels responsible for the success of AI in the company.