Klein, Felix

How companies succeed in developing ethical artificial intelligence (AI)

2023, 4 S.

How companies succeed in developing ethical artificial intelligence (AI)

felix.kleinge@gmail.com
FELIX KLEIN

The rapid advancement of artificial intelligence (AI) [2] has the potential to bring great benefits to society, but also raises important ethical and moral questions. To ensure that AI systems are developed and deployed in a responsible and ethical manner, companies must consider a number of factors, including fairness [11], accountability, transparency, privacy, and consistency with human values [6]. This essay provides an overview of the *key considerations* for building an ethical AI system and briefly discusses the *challenges* [1], including the importance of developing AI systems with a clear understanding of their potential impact on society and taking steps to mitigate any potential negative consequences. This essay also highlights the need for *continuous monitoring and evaluation* of AI systems and outlines a strategy, namely an enterprise-wide *"Ethics Sheet for AI tasks"* [3], to ensure that AI systems are used in an ethical and responsible manner within the company. Ultimately, building an ethical AI system requires a commitment to transparency, accountability, and a clear understanding of the ethical and moral implications of AI technology, and the company must be aware of the long-term consequences [8] of using a non-ethical and morally questionable AI system.

1 WHAT IS AI?

One of the problems with the term AI today is that there appears to be no clear definition that describes AI technologically, ethically, and conceptually [2][5]. There is a problem with the exact identity associated with the term AI. The basic problem is that there is also no precise definition of intelligence [2], as different views lead to different definitions. The two views listed in the paper [5] are, *a)* that intelligence is defined by human mind and its adaptability, and *b)* that AI systems are distinct from the human mind even when machines mimic the human mind. This results in five different definitions of AI systems in use today.

(1) **AI by structure**
An AI system can be defined by its brain-like architecture and is implemented by a *neuron-like process* [4], for example, Neural Networks (NN).
(2) **AI by behavior**
An AI system is defined by its *imitation of human behavior* [2]. This can be proven by the Turing test and is realized in many companies by chatbots for customer service.
(3) **AI by capacity**
An AI system is defined by the *system's ability to solve hard problems* in optimal time. It is said to have a human-like problem-solving capacity, such as a Convolutional Neural Network (CNN), used for image recognition.
(4) **AI by function**
An AI system *maps input data* to an output. An example of functional AI would be a recommendation system used by online retailers, such as Amazon, to suggest products to customers based on their browsing and buying behavior.
(5) **AI by principle**
An AI system is defined by the fact that *human intelligence can be described and reproduced* [4]. These are AI systems, like those that use rule-based algorithms to make decisions, such as a fraud detection system, that are used by every bank.

These five definitions lead to five different research and business objectives between which each company must decide in order to achieve its business goals. To talk further about AI systems in this essay, we use *our own definition* [2][5][13], to which the term AI refers in this essay, and which offers the most superficial definition possible:

AI refers to the ability of computers and machines to perform tasks that would normally require human intelligence, such as recognizing speech and images, solving problems, and making decisions. *AI* is achieved through a combination of advanced algorithms and machine learning techniques that allow computers to learn from data and make predictions.

2 AN ETHICAL AI

The second hurdle is to define what ethical AI is for the company. We try to provide a framework for this term. Furthermore, building an AI system that is considered morally and socially responsible, fair and transparent, and consistent with human values involves several important steps, which we explain below.

The framework
Ethical AI refers to the development and use of AI systems in ways that are considered morally and socially responsible [14], fair, transparent, and consistent with human values [1]. It is about creating AI systems that respect privacy, protect human rights, minimize harm, and operate fairly, impartially, and trustworthy [3]. The goal of ethical AI is to ensure that AI systems are used in ways that promote the well-being of society [16] and do not perpetuate or exacerbate existing biases or inequalities.

Below, we list the steps any company must take to ensure that AI is considered ethical and moral. These steps can be introduced as a form of an *"Ethics Sheets for AI"* [3] in the company and can serve as a standard for the development, implementation and use of ethical AI in the company:

(1) **Define the problem:**
Clearly define the problem that the AI system is being developed to solve and ensure that it is consistent with human values and ethics.

Author's address: Felix Klein felix.kleinge@gmail.com.

(2) **Identify biases:**
AI systems can perpetuate existing biases and inequalities. Therefore, it is important to identify and address potential sources of bias in the data and algorithms used to develop the system. Last but not least, it is beneficial if all developers involved are aware that they too have unconscious biases and incorporate them into the data or algorithms.
(3) **Diverse training data:**
To ensure that the AI system is fair and unbiased, use a variety of training data that represents a broad cross-section of the population.
(4) **Introduce transparency**
Ensure that the AI system is transparent in its decision-making process and provides an explanation for the decisions it makes. Transparency can be implemented, for example, through the use of decision trees and natural Language Generation Algorithms (NLG).
(5) **Ensure privacy:**
Protect user privacy by taking appropriate security measures and complying with privacy regulations.
(6) **Promote accountability:**
Assign clear responsibilities and accountability for the development and deployment of the AI system.
(7) **Monitor and evaluate:**
Monitor and evaluate the AI system regularly to identify and address any ethical concerns or unintended consequences.
(8) **Engage stakeholders:**
Engage with relevant stakeholders, including professionals, ethicists, and the public, to gain a better understanding of the ethical implications of the AI system and consider their feedback.

If these steps are followed, it is possible to develop an AI system that is consistent with human values and ethics and contributes to a more positive and equitable future [3][8][14].
There are several AI systems that have been developed and deployed with the goal of adhering to ethical principles and values and have been created using the guidelines just mentioned. To illustrate the versatility of ethical AI systems and their use, here are a few examples:

- **Fairness tools**
Tools designed to help identify and address biases in data and algorithms, ensuring that AI systems are fair and impartial. Fair representation learning [17] is a popular technique in this area, here AI systems learn fair and non-discriminatory representations of data.
- **Privacy-preserving AI**
AI systems that protect user privacy by implementing appropriate security measures and respecting data protection regulations. For example, homomorphic encryption a technique that allows computations to be performed on encrypted data without decrypting it, thereby protecting the privacy of the data and maintaining the data security of customers.
- **Explainable AI (XAI)**
AI systems that provide transparent and understandable explanations for the decisions they make (*Decision Trees* [18]).
- **Human-centered AI**
AI systems that are designed to prioritize the well-being of humans and align with human values and ethical principles, such as Siri from Apple or Alexa from Amazon. These AI systems are designed to perform tasks that are useful for humans in everyday life.
- **AI for social good**
AI systems that are deployed to solve societal challenges, such as improving healthcare, reducing poverty, and promoting sustainability. In this field, AI-assisted medical diagnoses in particular are well known and already standard in the treatment of patients in many countries.
- **AI in education**
AI systems that are designed to enhance the learning experience and support educational outcomes. One example is the Intelligent Tutoring System, which provide students with individualized feedback and guidance to reach their maximum potential.

These are just a few examples of AI systems that have been developed with the goal of aligning with ethical principles and values.

It is important to note that while these systems are designed with ethical considerations in mind, the actual impact of AI systems depends on how they are used and implemented in practice.

3 CHALLENGES

There are several challenges associated with the development and deployment of AI systems. These challenges highlight the importance of developing AI systems in a responsible and ethical manner, taking into account the potential impact on society. It is therefore necessary for any company to address these challenges proactively and effectively:

I. Significant computational effort and cost
With increasing complexity of calculations and larger models specialized for Big Data projects, the computational effort increases. Increased computational effort is always associated with acquisition costs and operating costs of new hardware. In particular, the increased power consumption that increases with the complexity of AI systems is a worldwide economic problem. Therefore, the development and implementation of such AI systems is costly and demands an ever increasing amount of our natural resources. The scarcity of rare natural resources requires a discussion on whether an AI system should be created or not [8].

II. Bias and fairness [1]
Bias and fairness in AI systems refers to the potential for AI systems to make unfair or discriminatory decisions [8] due to biases in the data [11] used to train the AI system or in the algorithms used to build the AI system. Bias can occur in a number of ways, including the following:

- **Training data**
AI systems are trained using large amounts of data, and if that data is biased, the AI system will learn and retain that

bias. The company must ensure that AI systems are trained on diverse, balanced and representative data.
- **Algorithm design**
  Biases can also occur in the development of AI algorithms, especially in the criteria used to make decisions. Designing algorithms that are transparent and interpretable to allow bias detection and correction is an essential requirement for a good AI system.
- **Humans as a flaw in the system**
  AI systems often depend on human decision makers to interpret the results produced by the AI system and act accordingly. When human decision makers have biases, those biases can affect how the AI system is used. However, developers also have unconscious biases that may be reflected in the selection, processing, and use of data and algorithms.

The impact of bias in AI systems can be significant, especially in areas such as employment, lending, and criminal justice, where unfair and discriminatory outcomes can have a significant impact on individuals and society.

III. Explainability and transparency - The black box problem
Ensuring explainability and transparency in AI systems [7][12] is a critical aspect of responsible AI development, as it enables companies, especially decision makers, stakeholders, and users, to understand how AI systems make decisions. This creates trust in the system, which is important not at least because companies are held accountable for AI systems and their actions. Some approaches to strengthen explainability and transparency are:

- **Designing algorithms**
  Using explainable AI algorithms and models, such as decision trees and linear regression, can make it easier to understand how AI systems make decisions.
- **Interpretability tools**
  Tools and techniques such as feature weighting [19], partial dependency diagrams, and counterfactual analysis can be used to make AI systems more interpretable and transparent.
- **Explanation generation**
  Explanation generation techniques, such as rule-based models and NLG, can be used to generate explanations for decisions made by AI systems, making it easier for decision makers to understand how AI systems make decisions.
- **Data provenance and transparency**
  Clear and comprehensive documentation of the data used to train AI systems, as well as the algorithms and models used, can help ensure transparency and accountability of AI systems.
- **Regular monitoring and evaluation**
  Regular monitoring and evaluation of AI systems can help identify and correct biases and ensure that AI systems make their decisions in a fair and transparent manner.

Ensuring explainability and transparency in AI systems requires a combination of technical and governance approaches, as well as ongoing attention and efforts to ensure that AI systems are used in a responsible and ethical manner.

IV. Privacy
Ensuring privacy in AI systems is also critical, as AI systems often process large amounts of sensitive personal data, including personal data, medical records, and financial data. Several approaches can be taken to ensure privacy in AI systems, including:

- **Data minimization**
  Minimizing the amount of personal data collected and processed by AI systems can help reduce privacy risks, through anonymization, for example.
- **Data protection**
  Protecting personal data through encryption and secure storage can help prevent unauthorized access to sensitive information. Like the homomorphic encryption that has already been mentioned.
- **Data governance**
  Implementing effective data governance policies and procedures, including data sharing agreements and data retention policies, can help ensure that personal data is used in a responsible and ethical manner.
- **Transparency**
  Ensuring that individuals are informed about the data collected and processed by AI systems, and that they have given consent for that data to be used, can help increase transparency and trust in AI systems.
- **Regular privacy and security audits**
  Regular privacy and security audits of AI systems can help detect and prevent privacy violations and data breaches.

It is important to balance the benefits of AI with the need to protect privacy and handle personal data responsibly.

V. Responsibility and liability
The responsibility and liability of AI systems [8][14] is a complex issue that raises many questions about who is responsible for the actions of AI systems and the results they produce. To ensure that AI systems are used responsibly and ethically, some key principles can be followed, including:

- **Human control**
  Ensuring that AI systems are designed to operate under human control and that human decision makers have the ability to intervene and make decisions as needed.
- **Algorithm accountability**
  Ensuring that AI algorithms are developed in a transparent and accountable manner, with clear and documented decision-making processes, can help build trust in AI systems and increase accountability.
- **Assigning responsibility and liability**
  Determining who is responsible and liable for the actions of AI systems and the results they produce, including manufacturers, developers, and users, and establishing clear lines of responsibility and liability.

- **Ethical and legal compliance**
  Ensure that AI systems are developed and deployed in accordance with ethical and legal principles, including privacy, data protection, and anti-discrimination laws.
- **Regular monitoring and evaluation**
  Regular monitoring and evaluation of AI systems can help detect and correct biases and ensure that AI systems make decisions in a fair and transparent manner.

Establishing responsibility and liability for AI systems requires a comprehensive approach involving a range of stakeholders, including government regulators, industry, and civil society.

VI. Regulation
There is a lack of clear and consistent regulations governing the development and use of AI systems [8], which can make it difficult to ensure that AI systems are used in an ethical and responsible manner, but companies can take various regulatory actions on their own, including:

- **Internal policies and procedures**
  Develop internal policies and procedures to regulate the development, deployment and use of AI systems, including data protection, privacy and ethical principles. The recommendation here is again to introduce a company-wide *"Ethics Sheets for AI Tasks"* [3].
- **Employee training**
  Train employees on responsible AI practices, including data protection, privacy, and ethical considerations.
- **Technical safeguards**
  Implement technical safeguards, such as encryption and secure storage, to protect sensitive data and prevent unauthorized access.
- **Third-party review and certification**
  Third-party review and certification of AI systems to ensure they are developed and deployed in a responsible and ethical manner.
- **External audits and assessments**
  Conduct regular external audits and assessments of AI systems to detect and correct biases and ensure they are making decisions in a fair and transparent manner.
- **Stakeholder engagement**
  Engaging stakeholders, including customers, employees, and civil society, to understand their concerns and ensure that your AI systems is developed and deployed in a responsible and ethical manner.

4 WHY DOES MY COMPANY NEED TO DEVELOP ETHICAL AI

The widespread use of AI systems in everyday life is increasing every year. The risk of these systems acting unethically or being abused is high. To illustrate, here are some examples of AI systems that have been criticized for their potential negative impact:

- **Facial recognition technology**
  Some facial recognition technologies have been criticized for their potential to create bias and racial profiling, as well as their potential to violate privacy rights.
- **AI-assisted predictive policing**
  Predictive policing systems that use algorithms to predict crime hot spots have been criticized for promoting bias and discrimination in the criminal justice system and raising serious ethical concerns.
- **AI-assisted hiring algorithms**
  Some AI-assisted hiring algorithms have been criticized for increasing bias and discrimination in the hiring process because they can reinforce existing biases in the data on which they are trained. Furthermore, applicants have an advantage tricking companies into writing keywords in their application in white font, or altering the metadata of the file so that they are most likely to be invited for an interview.
- **AI-assisted weapons**
  Some countries are developing autonomous weapons systems that use AI to make lethal decisions, raising ethical and moral concerns about the use of such technologies in warfare.

These are just a few examples of AI systems that have been criticized for their potential negative impact. It is important for companies to consider the ethical implications of developing and using AI and to take steps to mitigate any potential negative impact on society and the company.
The next question that many companies ask is, why should you take the extra effort and develop ethical AI? Whether developing ethical AI is worth it depends on the specific context and goals of the company. Although developing ethical AI offers its own advantages and disadvantages:

- **Reputation**
  Companies that prioritize ethical considerations and develop AI systems that align with human values are likely to improve their reputation and brand image.
- **Long-term benefits**
  Ethical AI systems that consider the potential impact on society are likely to be more sustainable in the long run, contributing to a more positive and equitable future.
- **Regulatory Compliance**
  Ethical AI systems that comply with relevant regulations and standards, such as data protection laws and privacy regulations, are less likely to face legal or regulatory challenges.
- **Customer trust**
  Customers are more likely to trust companies that prioritize ethical considerations and develop AI systems that protect privacy and provide transparent explanations for decisions.

However, developing an AI system that is not ethically standardized can also have its benefits. For example, such systems can provide immediate financial benefits and help companies achieve short-term goals. Many companies and organizations are nevertheless investing in the development and use of ethical AI to improve their bottom line while contributing to a more positive and equitable future.
For example, companies are developing AI systems designed to improve customer experience and retention, such as chatbots that

provide personalized customer support. Companies are also investing in AI systems that help optimize business processes, such as automating supply chain management and streamlining processes. In addition, many companies are developing and implementing AI systems that contribute to the public good, such as AI-powered healthcare solutions that improve patient outcomes. Not only can these systems help make a positive impact on society, but they can also be monetized, either through direct revenue streams or through cost savings and other efficiencies.

It is important to consider both the financial potential and the ethical implications of AI development and deployment.

Ultimately, the decision between developing an ethical AI system or e.g. a money-oriented AI system that, for example, prefers to spread fake news rather than facts, just because fake news spreads 10-20 times faster [20], depends on the specific context and goals of the organization. By considering ethical issues and developing AI systems that are aligned with human values, organizations can benefit both financially and socially and contribute to a more positive and equitable future. In order to achieve the company's goals for ethical AI, we recommend creating a standard that defines processes and steps that require all the considerations, challenges, and full documentation and monitoring.

5 REFERENCES

[1] Khan, A. A., "Ethics of AI: A Systematic Literature Review of Principles and Challenges", 2021.
[2] Turing, A. M., "Computing Machinery and Intelligence", in Parsing the Turing Test, 2009
[3]Mohammad, S. M., "Ethics Sheets for AI Tasks", 2021.
[4] Sarker, I. H., "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems", 2022
[5] Wang, Pei., "What Do You Mean by 'AI'"?, 2008
[6] A. Gittens, B. Yener and M. Yung, "An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML," , 2022
[7] Birhane A., Isaac W., Prabhakaran, V., Díaz M., Elish, M. C., Gabriel, I., Mohamed, S., "Power to the People? Opportunities and Challenges for Participatory AI", 2022
[8] Kaminski, Margot E., "Regulating the Risks of AI", 2022
[9] Saygin, A. P., Cicekli, I., Akman, V., "Turing test: 50 years later", 2000 [10] Penco, C., "Updating the Turing Test Wittgenstein, Turing and Symbol Manipulation", 2012
[11] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, ,A., "A Survey on Bias and Fairness in Machine Learning", 2022
[12] Bathaee, Y., "The Artifical Intelligence Black Box And The Failure Of Intent And Causation", 2018
[13] Russel, S., Norvig, P., "Artificial Intelligence: A Modern Approach", 2010
[14] Gebru, T., "The Social Responsibility of AI Systems, 2019
[15] Gebru, T., "The Case for Ethical AI", 2020
[16] Li, F., "Artificial Intelligence and Life in 2030", 2016
[17] Edelsbrunner, H., O'Rourke, J., "Fairness Constraints: Mechanisms for Fair Classification", 2001
[18] Quinlan, R., "Decision Tree Induction: A Tool for Automated Knowledge Acquisition in Databases", 1986
[19] Wettschereck, D., Aha, D. W., Mohri, T., "A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms", 1997
[20] Vosoughi, Soroush, Roy, D., Aral, S., "The spread of true and false news online.", 2018