UTS

Ethics of AI: From Principles to Practice

Data Science Institute, UTS

November 2020

Fang Chen and Jianlong Zhou

Email: fang.chen@uts.edu.au

Executive Summary

Artificial Intelligence (AI) is changing the world around us dramatically. It has powerful capabilities in prediction, automation, planning, targeting, and personalisation. With these capabilities, AI technologies are claimed to enable machines to exhibit human-like cognition, automate car-driving, protect our privacy, promote productivity, relieve workers of repetitive or dangerous tasks, and other human tasks. Ultimately, AI will change the way people are living, working, and entertaining. AI is, therefore, regarded as the driving force of the fourth industrial revolution.

As AI becomes more sophisticated and has the ability to perform more complex human tasks, we are seeing increasing concerns of what AI will be in decades and what it means, not just for business, but for humanity as a whole and for future of humans and society. Furthermore, since AI is fuelled by data, it faces ethical challenges related to data governance, including consent, ownership, and privacy. AI is a distinct form of autonomous and self-learning agency and thus raises unique ethical challenges. Ethics is specifically becoming one of major concerns related to AI uses. As a result, there are increasing active debates about the ethical principles and values that should guide AI's development and deployment in recent years.

The ethics of AI is the part of the ethics of technology specific to AI systems. It is sometimes divided into a concern with the moral behaviour of humans as they design, make, use and treat AI systems, and a concern with the behaviour of machines, in machine ethics. Ethical principles describe what is expected in terms of right and wrong and other ethical standards. Besides identifying the right set of fundamental ethical principles to inform the design, regulation, and use of AI and leverage it to benefit as well as respect individuals and societies, it is imperative to implement identified ethical principles in practical AI applications.

Ethical Principles of AI

Ethical principles of AI refer to ethical principles that AI should follow on the "do's" and "don'ts" of algorithmic uses in society. There is a rapid increase in the number and variety of ethical principles and guidelines for AI from various parties including governmental and inter-governmental organisations, private sectors, universities, as well as research institutes. They have made extended efforts by developing expert committees on AI, drafting policy documents on ethics of AI, and having active discussions on ethics of AI within and beyond the AI community. In particular both the public sector (governmental and inter-governmental organisations) and the private sector (companies and private sector alliances) have made significant efforts to ethical guidelines, indicating that the ethical challenges of AI concern both public entities and private enterprises.

Various parties identified slightly different ethical principles of AI because of their background or other reasons. For example, the ethical principles identified by IEEE include: human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse, and competence. While the ethical principles identified by CSIRO Australia are human, social and environmental wellbeing, human-centred values, fairness, privacy protection and security, reliability and safety, transparency and explainability, contestability, and accountability. However, there is a convergence around the principles of *transparency, justice and fairness, responsibility, non-maleficence, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity.*

Besides different governmental agencies, companies, and laboratories publishing their principles, guideline or codes on ethics of AI, some new institutes also have been established in recent years to specifically focus on investigations related to ethics of AI. Furthermore, in recognition of the increasingly pervasive role of AI based decision making systems and growing public concerns regarding the "black box" nature of such systems, the IEEE Standards Association launched the IEEE P7000 series of standards projects which address specific issues at the intersection of technological and ethical/societal considerations. The IEEE P7000 series empowers innovation across borders and enable societal benefit.

Mandatory Ethical Principles of AI

Despite the proliferation of ethical principles, it is argued that the some ethical principles are especially significant for AI solutions and must be implemented, which are also called mandatory ethical principles. These ethical principles are: community benefit, transparency, fairness, accountability, and privacy.

- **Community benefit:** AI should deliver a clear community or government benefit or insight. This principle should be a default principel for all AI solutions.
- **Transparency:** Transparency refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learns to adapt to its environment, and to the governance of the data used created. Three principle segments of transparency are validated: traceability, communication, and explainability.
- **Fairness:** Fairness concerns with AI's decision making in the equal treatment or equitability of decisions based on people's performance or needs. It is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics. Three principle segments of are validated: bias avoidance, accessibility and universal design, and stakeholder participation.
- **Accountability:** Accountability is about a clear acknowledgement and assumption of responsibility and "answerability" for actions, decisions, products and policies. Four principle segments are validated: auditability,

minimising and reporting negative impact, documenting trade-offs, and ability to redress.

- **Privacy:** From the digital perspective, privacy implies on the ability to control how data especially personal data is being collected, stored, modified, used, and exchanged between different parties. Three principle segments are validated: respect for privacy and data protection, quality and integrity of data, and access to data.

From Principles to Practices

The identified ethical principles should be translated into viable toolkits to shape AI-based innovation and support the practical application of ethical principles of AI. Despite the proliferation of ethical principles of AI in recent years, uncertainty remains regarding how ethical principles should be implemented in practice. The main challenges of implementing ethical principles and guidelines of AI lie in: complexity, variability, subjectivity, and lack of standardisation, including variable interpretation of the “components” of each of the ethical principles.

The stages of the AI lifecycle range from business and use-case development, design phase, training and test data procurement, building AI application, testing the system, deployment of the system to monitoring performance of the system. The ethical principles are suggested to be combined with every stage of the AI lifecycle to ensure that the AI system is designed, implemented and deployed in an ethical manner. However, the current practice of ethical principles of AI is mostly focused on interventions at the early input stages or the model testing stages in the AI lifecycle.

The Implementation of Ethical Principles of AI

It is a challenge to implement identified ethical principles of AI to make them actionable in order to validate the compliance of AI solutions with these ethical principles in practical applications.

In our implementation, we propose to implement ethical principles of AI both qualitatively and quantitatively. A series of checklist-style questionnaires are used to seek validations for the ethics around AI solutions. Therefore, two categories of questionnaires are provided for the validation: qualitative questionnaires and quantitative questionnaires. Qualitative questionnaires aim to evaluate the compliance of AI with ethical principles by developing qualitative questions on ethical principles and collecting responses from AI developers. For example, qualitative questions are asked to validate any measures used for the protection of data privacy. Quantitative questionnaires aim to evaluate the compliance of AI with ethical priciples by developing quantitative approaches to measure the compliance of AI with ethical principles. For example, quantitative measures are developed to validate the explanability of AI solutions and check the fairness of both data and models.

Our framework for the validation of ethical principles of AI is implemented as a web-based online platform to allow the effective validation of ethical principles for AI solutions. The main components included in the platform include: users of the platform (AI suppliers, validators, and administrators), projects to be validated, questionnaires, and validation outputs. Questionnaires can be customised for different projects to meet specific requirements.

By considering the dynamics of investigations on ethical principles of AI, our platform is designed as an open platform so that new ethical principles and corresponding qualitative and quantitative questions can be added easily. After the validation is finished, a summary of the validation based on the checked questions is provided to AI solution providers on the ethical aspects they have done and items that can be improved from the ethical perspective for AI solutions.

This framework has been used to validate a talent shortlisting platform developed by an AI company with talent data from big consulting companies.

The offered framework with a series of checklist-style questionnaires can be used by organisations who seek validations for the ethics around their AI solutions. The responses from questionnaires have to meet the necessary standards so that a certificate can be issued to AI solutions on their compliance with ethical principles. This ethics certification would apply both to the AI solution itself and to the organisation's processes in producing AI.