A Survey on Ethical Principles of AI and Implementations

Jianlong Zhou, Fang Chen, Adam Berry
Data Science Institute
University of Technology Sydney
Sydney, Australia
{Jianlong.Zhou, Fang.Chen, Adam.Berry}@uts.edu.au

Mike Reed, Shujia Zhang, and Siobhan Savage
Reejig Pty Ltd
Sydney, Australia
{Mike, Shujia, Siobhan}@reejig.com

***Abstract*—AI has powerful capabilities in prediction, automation, planning, targeting, and personalisation. Generally, it is assumed that AI can enable machines to exhibit human-like intelligence, and is claimed to benefit to different areas of our lives. Since AI is fueled by data and is a distinct form of autonomous and self-learning agency, we are seeing increasing ethical concerns related to AI uses. In order to mitigate various ethical concerns, national and international organisations including governmental organisations, private sectors as well as research institutes have made extensive efforts by drafting ethical principles of AI, and having active discussions on ethics of AI within and beyond the AI community. This paper investigates these efforts with a focus on the identification of fundamental ethical principles of AI and their implementations. The review found that there is a convergence around limited principles and the most prevalent principles are transparency, justice and fairness, responsibility, non-maleficence, and privacy. The investigation suggests that ethical principles need to be combined with every stages of the AI lifecycle in the implementation to ensure that the AI system is designed, implemented and deployed in an ethical manner. Similar to ethical framework used in biomedical and clinical research, this paper suggests checklist-style questionnaires as benchmarks for the implementation of ethical principles of AI.**

Keywords*—*AI, ethical principles, implementation

I. Introduction

A. *Artificial Intelligence*

Artificial Intelligence (AI) is typically defined as an interactive, autonomous, self-learning agency with the ability to perform cognitive functions in contrast to the natural intelligence displayed by humans, such as sensing and moving, reasoning, learning, communicating, problem solving (see Figure 1) [1]–[3]. It has powerful capabilities in prediction, automation, planning, targeting, and personalisation, and is claimed to be the driving force of the next industrial revolution (Industry 4.0) [4]. It is transforming our world, our life, and our society and affects virtually every aspect of our modern lives. Generally, it is assumed that AI can enable machines to exhibit human-like cognition, and it is more efficient (e.g. higher accuracy, faster, working 24 hours) than humans in various tasks. Claims about the promise of AI are abundant and growing related to different areas of our lives. Some examples are: in human's everyday life, AI can recognise objects in images, it can transcribe speech to text, it can translate between languages, it can recognise emotions in images of faces or speech; in traveling, AI makes self-driving cars possible, AI enables drones to fly autonomously, AI can predict parking difficulty by area in crowded cities; in medicine, AI can discover new uses for existing drugs, it can detect a range of conditions from images, it enables the personalised medicine; in agriculture, AI can detect crop disease, and spray pesticide to crops with pinpoint accuracy; in finance, AI can make stock trades without human intervention, and handle insurance claims automatically; AI can identify potentially threatening weather in meteorology; AI can even conduct various creative work, such as paint a van Gogh painting, write poems and music, write film scripts, design logos, recommend songs/films/books you like.

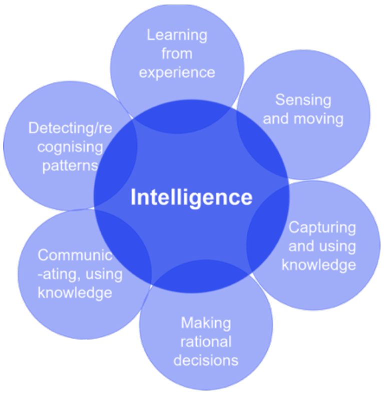
Learning from experience
Sensing and moving
Detecting/re cognising patterns
Intelligence
Capturing and using knowledge
Communic -ating, using knowledge
Making rational decisions

Figure 1. AI can make things as smart as humans or even smarter.

The diverse and ambitious claims about the promise of AI motivate wide adoptions of AI in various sectors including public services, retail, education, healthcare and others. For example, AI enables the monitoring of climate change and natural disasters, enhances the management of public health and safety, automates administration of government services, and promotes productivity for economic wellbeing of the country. AI also helps to prevent human bias in criminal justice, enables the efficient fraud detection (e.g. in welfare, tax, trading), enhances the protection of national security (e.g. with face recognition), and others.

However, AI may cause adverse effects to humans. For example, AI usually requires huge volumes of data especially personal data in order to learn and make decisions, the concern of privacy becomes one of important issues in AI [5]. Because AI can do many repetitive work and other work more efficiently than humans, people also worry about that they will lose their jobs because of AI. Furthermore, the highly developed Generative Adversarial Networks (GANs) can generate natural quality faces, voices, and others [6], which may be used to do harmful things in the society.

According to surveys by Mckinsey [7], the leading sectors in AI adoption today are mainly high tech and telecommunications, automotive and assembly, financial services, resources and utilities, media and entertainment, consumer packaged goods followed by transportation and logistics as well as others. All these adoptions will ultimately

978-1-7281-2547-3/20/$31.00 ©2020 IEEE

help to deliver a better quality of human life with manageable cost of living, better environment, and easy access of transport for time saving, etc.

B. Ethical Concerns on AI

Since diverse and ambitious claims of AI as well as its possible adverse effects to humans and society as mentioned above, it faces ethical challenges ranging from data governance, including consent, ownership, and privacy, to fairness and accountability and others. The debate about the ethical concerns on AI dates from the 1960s [3], [8]. As AI becomes more sophisticated and has the ability to perform more complex human tasks, their behaviour can be difficult to monitor, validate, predict and explain. As a result, we are seeing increasing ethical concerns and debate about the principles and values that should guide AI's development and deployment, not just for individuals, but for humanity as a whole and for future of humans and society [9]–[11]. Therefore, it is imperative to identify 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. Bossmann [12] summarised top nine ethical issues in AI as the following: unemployment, inequality, humanity, artificial stupidity, racist robots, security, evil genies, singularity, and robot rights.

Research found that ethics drive consumer trust and satisfaction, and consumers would place higher trust in a company whose AI interactions they perceived as ethical, which shows the importance of ensuring that AI systems are ethical for the positive impact of AI on society [13]. Therefore, an ethics framework for AI needs to set up to guide the development and deployment of AI. An ethics framework for AI is about updating existing laws or ethical standards to ensure that they can be applied in the context of new AI technologies [14]. There is debate about both what constitutes "ethical AI" and which ethical requirements, technical standards and best practices are needed for its realization [15].

This paper reviews the current efforts to ethical framework of AI and shows the most prevalent ethical principles of AI that the current work focuses on. The implementations of ethics of AI from principles to practices are then investigated to find effective ways to make ethical principles of AI actionable. The institutions specifically set up for ethics of AI are also exemplified and standards on ethical AI are presented. Finally, some intuitive suggestions on the implementation of ethical principles of AI are discussed. In summary, the primary contributions of this paper include:

- Identifying fundamental ethical principles of AI that the current work focuses on;
- Highlighting various approaches for the implement-tation of ethics of AI for actionable ethics;
- Proposing checklist-style questionnaires as benchm-arks for the implementation of ethics of AI.

II. Preliminary Knowledge

A. Ethics

Ethics is a branch of philosophy that involves systematising, defending, and recommending concepts of right and wrong conduct, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues [16]. It seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, justice and crime. There are three major areas of study within ethics recognised today [16]: meta-ethics, normative ethics, and applied ethics. Of these three major areas, normative ethics is the study of ethical action, investigating the set of questions that arise when considering how one ought to act, morally speaking. Normative ethics examines standards for the rightness and wrongness of actions [16]. The main streams within normative ethics include [16]: deontological ethics, dirtue ethics, and consequentialist ethics.

Ethical AI is mainly related to normative ethics especially the deontological ethics which emphasises principles of obligation/duty (e.g. Immanuel Kant was one of philosophers in this stream). The example questions in this stream are: what is my duty? What are the right rules to follow?

B. AI Ethics and Ethical AI

Ethics is a well-founded area with philosophers, academics, political leaders and ethicists spending centuries developing ethical concepts and standards. Various countries also set up different laws based on ethical standards. However, there is no commonly agreed ethics standards for AI because of its complexities and relatively new area. *AI ethics* is the part of the ethics of technology specific to AI based solutions. AI ethics concerns with the moral behaviour of humans as they design, construct, use and treat artificially intelligent beings, as well as concerns with the moral behaviour of AI agents [17]. The IEEE report, titled Ethically Aligned Design [18], argues that the three highest level ethical concerns that should drive AI design are to,

- "Embody the highest ideals of human rights",
- "Prioritize the maximum benefit to humanity and the natural environment", and
- "Mitigate risks and negative impacts as A/IS (Autonomous and Intelligent Systems) evolve as socio-technical systems".

It is imperative to build ethics into algorithms, otherwise AI will make unethical choices by design [19].

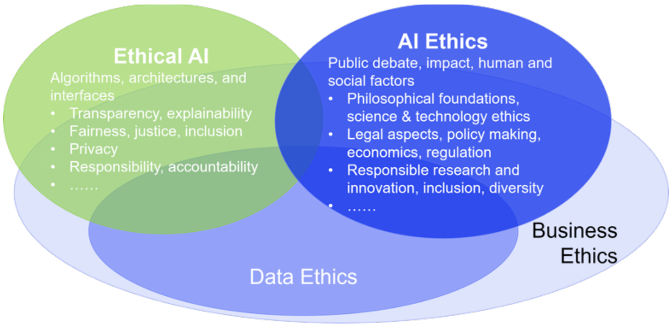
Ethical AI
Algorithms, architectures, and interfaces
• Transparency, explainability
• Fairness, justice, inclusion
• Privacy
• Responsibility, accountability
• ……
AI Ethics
Public debate, impact, human and social factors
• Philosophical foundations, science & technology ethics
• Legal aspects, policy making, economics, regulation
• Responsible research and innovation, inclusion, diversity
• ……
Business Ethics
Data Ethics

Figure 2. AI Ethics disciplinary landscape (adapted from [20]).

Generally, AI solutions are trained with a large amount of data for different business purposes. Data is at the core of AI, while business requirements and end users of AI determine functions of AI and how it will be used. Therefore, both data ethics and business ethics contributes to AI ethics. As shown in Figure 2, AI ethics needs active public debate by considering AI impact, as well as human and social factors. It is built based on different aspects such as philosophical foundations, science and technology ethics, legal aspects, responsible research and innovation for AI as well as others. Ethical principles describe what is expected in terms of right and wrong and other ethical standards. Ethical principles of AI refer to ethical principles that AI should follow on the "do's" and "don'ts" of algorithmic use in society. Ethical AI refers to AI algorithms, architectures and interfaces that follow ethical principles of AI, such as transparency, fairness, responsibility and privacy. Figure 2 summarises an overview of AI ethics disciplinary landscape.

III. Current Efforts to Ethical Framework of AI

In order to mitigate various ethical concerns as reviewed previously, national and international organisations including governmental organisations, private sectors as well as research institutes 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. For example, the European Commission has published "Ethics Guidelines for Trustworthy AI" [21], emphasising that AI should be "human centric" and "trustworthy". The United Kingdom's national plan for AI explores AI ethics from different angles including inequality, social cohesion, prejudice, data monopolies, criminal misuse of data, and suggestions for the development of an AI Code [22]. Australia also has published its AI ethics framework [14], which uses a case study based approach to investigate core ethical principles for AI and proposes a toolkit for implementing ethical AI.

Besides governmental organisations, big leading companies such as Google [23] and SAP [24] publicly released their AI principles and guidelines. Furthermore, professional associations and non-profit organisations such as Association of Computing Machinery (ACM) also issued their recommendations for ethical AI. The Institute of Electrical and Electronics Engineers (IEEE) has launched the "IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems" "to ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity" [25]. IEEE also set up the P7000 standards projects specifically for the standards for the future of ethical intelligent and autonomous technologies.

This section investigates questions of what constitutes ethical principles of AI and which ethical requirements are needed for its realisation.

A. Monitoring of Ethical Compass

Different parties have active discussions on unlocking AI values in business. For example, KPMG [26], [27] argued that ethical frameworks of AI are essential to maximise the benefits to society. KPMG has active discussions in various aspects of ethics (e.g. bias, privacy) [27], and raises key questions on designing the principles (e.g. who should have a say in the design of these principles and standards, monitor their adoption and police their impact as the development and use of AI increases?) [26]. Therefore, KPMG proposes nine ways to monitor ethical compass in AI [28]:

- Start the ethical discussion within the company;
- Define or update the company's core values and corporate social responsibility focus;
- Evaluate how the company historically makes tough decisions such as deciding when and where to close down operations and open new locations;
- Consider how core values extend into technologies;
- Knowing the complexity of automation decisions, many companies establish an ethics committee or board that typically includes outside experts;
- Follow through by establishing metrics to track the residual effects of automation;
- Consider the work of AI-oriented non-profits in strategy;
- Get involved in education programs – preschool through college – to help make sure there's a workforce skilled for the future;
- Cyber and physical security should be integral to protect intellectual property.

KPMG specifically launches a framework of "AI in Control" to help organizations realize value from AI technologies while achieving imperative objectives like algorithm integrity, explainability, fairness and agility, aiming to address the trust gap in AI [29].

B. Ethical Principles of AI

There is a rapid increase in the number and variety of ethical guidelines for AI. Jobin et al. [15] made an in-depth investigation on ethical principles of AI and identified 84 documents related to ethical principles or guideline for AI (Jobin et al. failed to identify China AI ethical risk analysis [30] in their survey and therefore there are 85 documents in total). Algorithm Watch [31] also maintains an AI ethics guidelines global inventory, which provides a global landscape of AI ethics and is a work in progress.

Various parties identified slightly different ethical principles of AI because of their background or other reasons. For example, ethical principles identified by Data61 CSIRO in Australia include: human, social & environmental wellbeing, human-centred values, fairness, privacy protection and security, reliability & safety, transparency & explainability, contestability, and accountability [14]. While ethical principles identified by IEEE include: human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse, and competence [18]. The comparison of these shows that there are some common ethical principles of interest of AI from various parties.

C. Analysis of Ethical Principles of AI

Jobin et al.'s [15] investigation found that no single ethical principle is explicitly endorsed by all existing ethical guidelines reviewed, but there is an emerging convergence around the following principles: transparency, justice and fairness, responsibility, non-maleficence, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity as shown in Figure 3 (China AI ethical risk analysis is included in this figure), which also shows a developing convergence in the global policy landscape.

Figure 3 shows that the most prevalent principle is transparency followed by justice and fairness, which indicates a moral priority to require transparent processes throughout the entire AI lifecycle (from transparency in the design and development of algorithms to transparent uses of AI). It also indicates that justice and fairness need to be adequately addressed to avoid inequality because of AI uses.

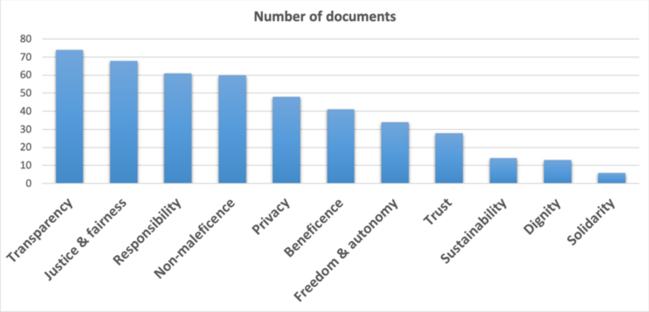

Figure 3. Ethical principles identified in existing AI guidelines.

As investigated above, a very large number of ethical principles, codes, guidelines, or frameworks have been proposed over the past few years. However, the "principle proliferation" of AI may result in overwhelm and confuse, and cause questions such as whether these principles are overlapping and converge upon a set of agreed-upon principles, or diverge, with significant disagreement over what constitutes "ethical AI" [32]. Floridi and Cowls [32] analysed these principles and identified an overarching framework consisting of five core principles for ethical AI: beneficence, non-maleficence, autonomy, justice, and explicability. Different terms express justice, e.g. "fairness". Different terms also express explicability, e.g. "transparency", "understandable and interpretable". Therefore, these results align with the investigated results presented in [15].

IV. From Principles to Practice for Ethical AI

As reviewed above, various sets of ethical principles and frameworks for AI were published typically from industry (e.g. Google, IBM, Microsoft, Intel), government (e.g. UK Lords Select Committee, European Commission's High-Level Expert Group), and academia (e.g. Future of Life Institute, IEEE, AI4People). These principles act as normative constraints on the "do's" and "don'ts" of AI in society.

Once identified, ethical principles should be translated into viable toolkits and guidelines to shape AI-based innovation and support the practical application of ethical principles of AI. Toolkits and guidelines on how to apply ethical principles into the design, implementation, and deployment are highly necessary. Despite various efforts in ethical principles of AI as reviewed, uncertainty remains regarding how ethical principles should be implemented in AI [15]. The main challenges of implementing ethical principles and guidelines of AI include: complexity, variability, subjectivity, and lack of standardisation, including variable interpretation of each of the ethical principles [33].

IEEE proposes two practices of ethics to follow: ethical design and sustainable development, and focuses on "societal implications of conventional and emerging technologies, including intelligent systems" [34]. Ethical design as a concept has been popularized [18]. Sustainable development has had great attention due to the United Nations' Sustainable Development Goals for 2030. Furthermore, the concept of extended intelligence is proposed to integrate humans and machines instead of thinking about machine intelligence in terms of humans versus machines [35]. This is based on the understanding that "instead of trying to control or design or even understand systems, it is more important to design systems that participate as responsible, aware and robust elements of even more complex systems" [35]. The following subsections investigate various approaches on the implementation of ethical principles of AI.

A. *AI Lifecycle and Typology of Applied AI Ethics*

A typical AI application lifecycle usually includes different stages 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 AI application lifecycle delineates the role of every stage in data science initiatives ranging from business to engineering. It provides a high-level perspective of how an AI project should be organized for real and practical business value with the completion of every stages.

Morley et al. [33] constructed a typology by combining the ethical principles with the stages of the AI lifecycle to ensure that the AI system is designed, implemented and deployed in an ethical manner. The typology indicates that each ethical principle should be considered at every stage of the AI lifecycle. The full typology from Morley et al. [33] can be found from https://tinyurl.com/AppliedAIEthics. The typology provides a brief snapshot of what tools are currently available to AI developers to encourage the progression of ethical AI from principles to practice. The typology found that the current interest in the practice of "ethical AI", and thus the availability of tools and methods, is not evenly distributed across the AI lifecycle. Most attention for all the ethical principles is focused on interventions at the early input stages or the model testing stages. No tools or methods were found for ensuring value-alignment at the deployment stage and very few tools or methods were found for promoting autonomy during the middle building and testing stages from [33].

B. *Operationalising AI Ethics*

A workshop from Safeforce Research [36] identified three groups of approaches in operationalising AI ethics: 1) Socialisation/education; 2) Processes to kickoff every project; 3) Tools or processes throughout the product development. It is regarded that educating everyone in the organization is the first task for people to get familiar with AI ethics.

It is found that: 1) review boards and discussion forums with experts are valuable resources for AI ethics operations; 2) Checklists are useful for AI ethics implementations to check each aspects of principles; and 3) The importance of documentation throughout is crucial and provides transparency, accountability, and consistency. Examples of these approaches include model cards [37], datasheets [38], feedback from reviews and decision outcomes.

C. *Toolkits and Methods for Ethical AI*

1) *Duties of Different Parties*

The research suggests that organizations trying to focus on ethics in AI must take a targeted approach to make systems fit for purpose. Capgemini [13] recommends a three-pronged approach to build a strategy for ethics in AI that embraces all key stakeholders :

- For business leaders and those with a remit for trust and ethics: strategy and code of conduct for ethical AI and policies for practices play important roles.
- For the customer and employee-facing teams, ethical usage of AI and education of customers are important.

- For AI, data and IT leaders and their teams, AI systems transparent and understandable are key concerns.

2) Datasheets for Datasets

Data plays a critical role in AI and the characteristics of datasets fundamentally influences a model's behavior [38]. By analogy to the electronics industry where every component is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information, Gebru et al. [38] proposed that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, preprocessing/cleaning/ labeling, recommended uses, distribution, and maintenance. A set of questions and workflow are provided to cover the information that a datasheet for a dataset might contain. For example, for the information on motivation of dataset to be covered, the questions to be documented include [38]:

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
- Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
- Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

3) Model Cards

In order to clarify the intended use cases of AI models and minimize their usage in contexts for which they are not well suited, Mitchell et al. [37] recommended that released AI models be accompanied by documentation detailing their performance characteristics, called model cards, to encourage transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups and intersectional groups that are relevant to the intended application domains [37].

4) Social contract

Caron and Gupta [39] proposed that the adoption of AI can be thought of as a form of social contract. In order to enable a social contract to arise for the adoption and implementation of AI, the development of 1) a socially accepted purpose, through 2) a safe and responsible method, with 3) a socially aware level of risk involved, for 4) a socially beneficial outcome, is the key.

D. Checklist-Style Assessment

The Institute for Ethical AI & Machine Learning [40] proposed eight principles for responsible ML development. The eight principles are a practical framework put together by domain experts to provide guidance for technologists to develop machine learning systems responsibly. These eight principles of responsible machine learning include [40]: human augmentation, bias evaluation, explainability by justification, reproducible operations, displacement strategy, practical accuracy, trust by privacy, and data risk awareness.

The assessment criteria for responsible ML principles are also proposed [40]. The assessment criteria are then used to build a framework, which converts the principles for responsible machine learning into an actionable checklist [41].

E. A four-step implementation approach

Accenture proposed a four-step implementation approach for ethical AI [42]. The four steps are:

- Set up an ethical AI committee: The ethical AI committee aims to consider ethical issues, foster discussion forums and publish resulting guidance to the industry and regulators.
- Gather the development of ethical codes and participate actively in the development internationally: These activities help to pick up on the latest internationally acceptable thinking of ethics of AI.
- Develop core ethical principles: Engage with various stakeholders to develop core fundamental ethical principles.
- Encourage the development of sector specific codes: Instead of general ethical principles, it is more effective to develop sector specific ethical principles.

These steps provide important points on how to start the implementation of ethics of AI in practice.

V. Institutions on Ethical AI

Different governmental agencies, companies, and laboratories published their principles, guideline or codes on ethics of AI. Some new institutes have been established in recent years to specifically focus on investigations related to ethics of AI. Table I lists examples of this kind of institutes. Besides, many universities set up research labs focusing on human-centred AI especially ethical aspects in AI.

Table I. Institutes on ethics of AI

Institute	Country	Note
AI Now Institute	USA	A research institute examining the social implications of AI
Future of Life Institute	USA	Work to mitigate existential risks facing humanity, particularly existential risk from advanced AI
The Institute for Ethical AI and Machine Learning	UK	Carry out research into processes and frameworks that support the responsible development, deployment and operation of machine learning systems
Humanising Machine Intelligence	Australia	Unite world-leading researchers in the social sciences, philosophy, and computer science to work closely together to make substantial progress towards moral machine intelligence
Gradient Institute	Australia	Progress the research, design, development and adoption of ethical AI systems

AI Ethics Lab	USA, Turkey	Provide ethics guidance to researchers, developers, and legislators
Data Ethics Site	USA	Investigate philosophical and ethical dilemmas at the intersection of big data and human experience
Montreal AI Ethics Institute	Canada	Create tangible and applied technical and policy research in the ethical, safe, and inclusive development of AI

VI. IEEE Standards on Ethical AI

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 [43]. The IEEE P7000 series empowers innovation across borders and enable societal benefit. IEEE also has launched the Ethics Certification Program for Autonomous and intelligent Systems (ECPAIS), which aims to create specifications for certification and marking processes advancing transparency, accountability and reduction of algorithmic bias in Autonomous and Intelligent Systems [44]. However, these works are still in progress and not fully ready for industry applications.

VII. Discussions

Ethics of AI is becoming one of the mostly discussed topics in recent years as AI is widely used in different domains for prediction, automation, planning, targeting, and personalisation as well as others. This leads to the "principle proliferation" of AI with a very large number of ethical principles, codes, guidelines, or frameworks have been proposed over the past few years. However, it is still a challenge task to implement ethics in AI in practical applications. Following Accenture's four-step implementation approach for ethical AI [42] as reviewed above, this section discusses implementation paths to ethical AI and proposes potential approaches that can be used to implement ethical principles of AI for a specific sector in order to make ethics of AI operable.

A. *Ethical AI Committee*

In order to make ethics of AI actionable, the establishment of ethical AI committee is the first step. The primary challenges of this step include who would be the best candidates of committee members, and which areas should they come from. Google ever shut down its External Advisory Board for AI just a week after forming it, which shows how challenging it is to choose candidates for an ethical AI committee. The committee members need to at least understand how AI works and how to pull the ethics out of the data [45]. However, legal or social experts are good at ethical issues related to data governance, but they may not be familiar with how an AI model such as deep learning model is built with a large number of parameters as AI experts be. The conversation about AI ethics is a philosophical discussion and needs to be elevated to a sufficiently high level from different fields. Therefore, committee members can be experts that span the fields of engineering, law, science, economics, ethics, philosophy, politics, and health. IEEE suggests that the key experts would include but not limited to [44]:

- Specialists developing AI based products and services;
- Academic institution experts in AI;
- Government organisations involved with AI policy and/or regulations.

B. *Ethical Principles of AI for a Specific Sector*

As reviewed in previous sections, the most endorsed ethical principle of AI is transparency, followed by fairness, responsibility/accountability, non-maleficence, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity [15]. Accenture also suggests to develop sector specific codes [42]. This is because that different sectors have different emphasis on ethical principles. For example, in high stake applications such as AI-supported diagnostics, the transparency of the system is one of key principles for consideration. While in an AI-assisted recruiting system, unfair discrimination against individuals, communities or groups would be the main issue to avoid. Therefore, despite the common ethical principles as reviewed above, the development of ethical principles for a specific sector instead of general ethical principles is more effective for the implementation of ethical principles.

C. *Checklist-Style Assessment*

Different approaches can be used to explain an AI model/algorithm for its transparency [46], assess bias of an algorithm for its fairness in prediction [47], or enhance privacy with algorithms [48]. All these algorithmic approaches for the enhancement of ethics of AI can be effective for AI experts/developers to detect/assess ethical issues of AI models/algorithms, but can cause further confusions or concerns for domain users of AI because of complexities of those algorithmic approaches.

The widely used ethical framework for biomedical and clinical research suggests various checklist-style questionnaires as benchmarks [49]. Despite theoretical weaknesses in its framework and some practical problems in the implementation, this still became the default ethics governance model in biomedical research [50]. Besides, like life itself, all research entails some risks. Clinical research usually offers individual participants a favorable net risk-benefit ratio [49]. The principle of a favorable net risk-benefit ration in clinical research requires fulfilling three benchmarks [49]:

- The risks the research should be delineated and minimised. Researchers should identify the type, probability and magnitude of the risks of the research;
- The type, probability, and magnitude of the benefits of the research should be identified. The benefits to individual participants, such as health improvements, are relevant;
- The risks and potential benefits of the clinical research interventions to individual participants should be compared.

Furthermore, informed consent is widely used in many human related research fields such as biomedical and clinical research. Valid informed consent requires that the consenting person has the capacity to understand risks and benefits, make decisions, receive relevant information about the study, understand that information, and consent voluntarily without coercion [49].

Similarly, checklist-style assessment can be used to implement ethical principles of AI. As reviewed in the previous section, different checklist-style questionnaires have been proposed for the ethics of AI. Besides, all AIs also entail some risks. The risks of AI should be identified, including the type, probability and magnitude of the risks. The benefits of AI should also be identified. The comparison of the risks and benefits of AI should be revealed to users. The approach similar to the informed consent can be used to inform AI users about risks and potential benefits of AI for the AI-informed decision making.

The advantages of the checklist-style assessment for the implementation of ethics of AI at least include:

- It is easy to understand by AI domain users;
- It has the high operability for both AI developers and domain users of AI, including implementation, updating and extension;
- It has the high potential for setting up implementation standards of ethical AI, especially for a specific sector.

The algorithmic approaches for the detection and assessment of ethical issues of AI can be a supplement for the checklist-style assessment of ethics of AI.

In summary, despite the proliferation of ethical principles of AI, there is no a widely accepted list of ethical principles of AI with which that both AI developers and AI users comply. Furthermore, it is a challenge to implement ethical principles of AI in practice because of their complexities. There is also a lack of ethical standards that are used to certify AI solutions.

VIII. Conclusion and Future Work

AI has powerful capabilities in prediction, automation, planning, targeting, and personalisation. Generally, it is assumed that AI can enable machines to exhibit human-like cognition. Claims about the promise of AI are abundant and growing related to different areas of our lives, which raises unique ethical challenges. As a result, we are seeing increasing ethical concerns related to AI uses. This paper reviewed the latest efforts to ethical framework from various parties including governmental organisations, private sectors as well as research institutes. The converged most prevalent ethical principles of AI were identified from the review. Furthermore, various discussions and tries on the implementation of ethics of AI were investigated in this paper. It was argued that the ethical principles need to be combined with every stages of the AI lifecycle to ensure the compliance of ethics of AI systems. Finally, this paper suggested checklist-style questionnaires as benchmarks for the implementation of ethical principles of AI. Our future work will focus on the implementation of ethical principles to make ethics of AI operable.

References

[1] M. Taddeo and L. Floridi, “How AI can be a force for good,” *Science*, vol. 361, no. 6404, pp. 751–752, Aug. 2018, doi: 10.1126/science.aat5991.

[2] J. Zhou and F. Chen, “AI in the public interest,” in *Closer to the Machine: Technical, Social, and Legal Aspects of AI*, C. Bertram, A. Gibson, and A. Nugent, Eds. Melbourne, Australia: Office of the Victorian Information Commissioner, 2019.

[3] A. L. Samuel, “Some Moral and Technical Consequences of Automation—A Refutation,” *Science*, vol. 132, no. 3429, pp. 741–742, Sep. 1960, doi: 10.1126/science.132.3429.741.

[4] “Industrial revolutions: the 4 main revolutions in the industrial world,” *Sentryo*, Feb. 23, 2017. https://www.sentryo.net/the-4-industrial-revolutions/.

[5] M. Deane, “AI and the Future of Privacy,” *Towards Data Science*, Sep. 05, 2018. https://towardsdatascience.com/ai-and-the-future-of-privacy-3d5f6552a7c4.

[6] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, “Deep Learning for Deepfakes Creation and Detection: A Survey,” *arXiv:1909.11573 [cs, eess]*, Jul. 2020.

[7] J. Bughin *et al.*, “Artificial Intelligence: The Next Digital Frontier?,” McKinsey Global Institute, Jun. 2017.

[8] N. Wiener, “Some Moral and Technical Consequences of Automation,” *Science*, vol. 131, no. 3410, pp. 1355–1358, May 1960, doi: 10.1126/science.131.3410.1355.

[9] E. Bird, J. Fox-Skelly, N. Jenner, R. Larbey, E. Weitkamp, and A. Winfield, “The ethics of artificial intelligence: Issues and initiatives,” European Parliamentary Research Service, Technical Report PE 634.452, Mar. 2020.

[10] S. Lo Piano, “Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward,” *Humanities and Social Sciences Communications*, vol. 7, no. 1, Art. no. 1, Jun. 2020.

[11] A. Gupta *et al.*, “The State of AI Ethics Report (June 2020),” *arXiv:2006.14662 [cs]*, Jun. 2020.

[12] J. Bossmann, “Top 9 ethical issues in artificial intelligence,” *World Economic Forum*, Oct. 21, 2016. https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/.

[13] “Why addressing ethical questions in AI will benefit organizations,” *Capgemini Worldwide*, Jul. 05, 2019. https://www.capgemini.com/research/why-addressing-ethical-questions-in-ai-will-benefit-organizations.

[14] D. Dawson *et al.*, “Artificial Intelligence - Australia’s Ethics Framework,” Data61, CSIRO, Australia, 2019.

[15] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence*, pp. 389–399, Sep. 2019.

[16] Wikipedia, “Ethics,” *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Ethics&oldid=914442232.

[17] Wikipedia, “Ethics of artificial intelligence,” *Wikipedia*. Sep. 10, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Ethics_of_artificial_intelligence&oldid=915019392.

[18] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, “Ethically Aligned Design: A vision for prioritizing human well-being with autonomous and intelligent systems,” IEEE, 2019. [Online]. Available: https://standards.ieee.org/content/

dam/ieee-standards/standards/web/documents/ other/ead1e.pdf.

[19] R. McLay, "Managing the rise of Artificial Intelligence," 2018. https://tech.humanrights.gov.au/ sites/default/files/inline-files/100%20-%20Ron% 20McLay.pdf.

[20] M. Rovatsos, "From AI Ethics to Ethical AI," Macau, China, IJCAI 2019 Tutorial, Aug. 2019.

[21] Eurpoean Commission, "Ethics guidelines for trustworthy AI," European Commission, Brussels, Dec. 2018. Accessed: Sep. 16, 2019. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai.

[22] Select Committee on Artificial Intelligence, "AI in the UK: ready, willing and able," House of Lords, UK, Apr. 2018.

[23] Google, "Artificial Intelligence at Google: Our Principles," *Google AI*. https://ai.google/principles/.

[24] SAP, "SAP's Guiding Principles for Artificial Intelligence," Sep. 18, 2018. https://news.sap.com/ 2018/09/sap-guiding-principles-for-artificial-intelligence/.

[25] IEEE, "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems," *IEEE Standards Association*. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[26] K. Reid, "The opportunity of AI: Unlocking the potential," *KPMG*, Apr. 05, 2019. https://home.kpmg /au/en/home/insights/2019/03/artificial-intelligence-opportunity-unlocking-potential.html.

[27] "Artificial Intelligence and Ethics," *KPMG*, May 09, 2019.https://home.kpmg/au/en/home/insights/2019/04/artificial-intelligence-ethics.html.

[28] KPMG, "An ethical compass in the automation age." https://advisory.kpmg.us/content/dam/advisory/en/pdfs/ an-ethical-compass-in-the-automation-age.pdf.

[29] J. Samuel, "KPMG launches framework to help businesses gain greater confidence in their AI technologies - KPMG Global," *KPMG*, Feb. 13, 2019. https://home.kpmg/xx/en/home/media/press-releases/2018/12/helping-businesses-gain-trust-in-their-ai-technologies.html.

[30] China National Artificial Intelligence Standardization Group, "AI Ethical Risk Analysis Report," Apr. 2019. [Online]. Available: http://www.cesi.cn/images/ editor/20190425/20190425142632634001.pdf.

[31] "AI Ethics Guidelines Global Inventory," *AlgorithmWatch*.https://algorithmwatch.org/en/project/ ai-ethics-guidelines-global-inventory/.

[32] L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society," *Harvard Data Science Review*, vol. 1, Jun. 2019, doi: 10.1162/ 99608f92.8cd550d1.

[33] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices," *arXiv:1905.06876 [cs]*, May 2019,.

[34] G. Adamson, J. C. Havens, and R. Chatila, "Designing a Value-Driven Future for Ethical Autonomous and Intelligent Systems," *Proceedings of the IEEE*, vol. 107, no. 3, pp. 518–525, Mar. 2019.

[35] J. Ito, "Resisting Reduction: A Manifesto," Nov. 2017, doi: 10.21428/8f7503e4.

[36] K. Baxter, "Living Ethics in AI: How to Expand from Principles to Impact," *Salesforce Research*, Aug. 06, 2019. https://blog.einstein.ai/living-ethics-in-ai-how-to-expand-from-principles-to-impact/.

[37] M. Mitchell *et al.*, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2019, pp. 220–229.

[38] T. Gebru *et al.*, "Datasheets for Datasets," *arXiv: 1803.09010 [cs]*, Mar. 2018, Accessed: Oct. 17, 2019.

[39] M. S. Caron and A. Gupta, "The Social Contract for AI," Macau, China, Aug. 2019, [Online]. Available: https://aiforgood2019.github.io/papers/IJCAI19-AI4SG_paper_36.pdf.

[40] The Institute for Ethical AI and Machine Learning, "The Responsible Machine Learning Principles." https://ethical.institute.

[41] AI-RFX Committee, "AI-RFX Procurement Framework v1.0: Machine learning maturity model," The Institute for Ethical AI & Machine, UK, 2019.

[42] "An Ethical Framework for Responsible AI and Robotics." https://www.accenture.com/gb-en/company-responsible-ai-robotics.

[43] A. Koene, L. Dowthwaite, and S. Seth, "IEEE P7003$^{\text{TM}}$ standard for algorithmic bias considerations: work in progress paper," in *Proceedings of the International Workshop on Software Fairness - FairWare '18*, Gothenburg, Sweden, 2018, pp. 38–41.

[44] IEEE, "IEEE Launches Ethics Certification Program for Autonomous and Intelligent Systems," *IEEE Standards Association*, Oct. 02, 2018. https://standards.ieee.org/ news/2018/ieee-launches-ecpais.html.

[45] Corinium, "Ethics of AI," 2019. https://cdn2.hubspot.net/hubfs/2631050/CDAO%20New%20Zealand/Corinium_Ethics-of-AI_brochure_NZ.pdf.

[46] V. Arya *et al.*, "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques," *arXiv:1909.03012 [cs, stat]*, Sep. 2019.

[47] R. K. E. Bellamy *et al.*, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," *arXiv:1810.01943 [cs]*, 2018.

[48] H. B. McMahan *et al.*, "A General Approach to Adding Differential Privacy to Iterative Training Procedures," *arXiv:1812.06210 [cs, stat]*, Dec. 2018.

[49] E. J. Emanuel, D. Wendler, and C. Grady, "An ethical framework for biomedica research," in *The Oxford Textbook of Clinical Research Ethics*, Oxford University Press, 2011, pp. 123–135.

[50] C. Canca, "A New Model For AI Ethics In R&D," *Forbes*. https://www.forbes.com/sites/insights-intelai/ 2019/03/27/rethinking-ethics-in-ai-rd/.