AI & SOCIETY (2023) 38:2693–2703
https://doi.org/10.1007/s00146-022-01602-z

OPEN FORUM

AI ethics: from principles to practice

Jianlong Zhou[1] · **Fang Chen**[1]

Received: 18 December 2021 / Accepted: 15 November 2022 / Published online: 24 November 2022

Abstract
Much of the current work on AI ethics has lost its connection to the real-world impact by making AI ethics operable. There exist significant limitations of hyper-focusing on the identification of abstract ethical principles, lacking effective collaboration among stakeholders, and lacking the communication of ethical principles to real-world applications. This position paper presents challenges in making AI ethics operable and highlights key obstacles to AI ethics impact. A preliminary practice example is provided to initiate practical implementations of AI ethics. We aim to inspire discussions on making AI ethics operable and focus on its impact on real-world applications.

Keywords AI ethics · Ethical principles · Implementation · Challenges

1 Introduction

1.1 Artificial intelligence

Artificial intelligence (AI) is typically defined as an autonomous and self-learning agency with the ability to perform intelligent functions in contrast to the natural intelligence displayed by humans, such as learning from experience, reasoning, problem solving (Taddeo and Floridi 2018; Zhou and Chen 2019). AI is a computer system which performs tasks that are typically associated with human intelligence or expertise. It has powerful capabilities in prediction, automation, planning, targeting, and personalisation. It is transforming our world, our life, and our society and affects virtually every aspect of our modern lives (Zhou and Chen 2018). Generally, it is assumed that AI can enable machines to conduct tasks that human often do, and it is more efficient (e.g., higher accuracy, faster) 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 recognize objects in images (He et al. 2016; Zhou et al. 2017), it can transcribe speech to text, it can translate between languages (Monroe 2017), it can recognize emotions in images of faces or speech (Zhao et al. 2014), AI makes self-driving cars possible in traveling (Bojarski et al. 2016), AI enables drones to fly autonomously, AI can predict parking difficulty by area in crowded cities, AI can identify potentially threatening weather in meteorology, AI can even conduct various creative work, such as paint a van Gogh painting (Gao et al. 2020), write poems and music, write film scripts, design logos, and recommend songs/films/books you like as well as many others (Batmaz et al. 2019).

The impressive performance of AI we have seen across a wide range of domains motivates extensive 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 (Rolnick et al. 2019), enhances the management of public health and safety (Mooney and Pejaver 2018), automates administration of government services (Anastasopoulos and Whitford 2019), and promotes productivity for economic well-being of the country. AI also helps to enables efficient fraud detection (e.g., in welfare, tax, trading, credit card) (Awoyemi et al. 2017), enhances the protection of national security (e.g., with unauthorized network access and malicious email detection) (Amrollahi et al. 2020), and others.

However, AI may cause negative 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

✉ Jianlong Zhou
jianlong.zhou@uts.edu.au

Fang Chen
fang.chen@uts.edu.au

[1] Data Science Institute, University of Technology Sydney, Sydney, Australia

in AI (Deane 2018). 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 (Nguyen et al. 2020), which may also be used to do harmful things in the society. For example, GANs have been used to create fake videos by swapping the face of a person by the face of another person, which have harmful usages including fake news, hoaxes, and financial fraud (Tolosana et al. 2020).

1.2 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 (Wiener 1960; Samuel 1960). As AI becomes more sophisticated and has the ability to perform more complex human tasks, their behavior 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 the future of humans and society (Bird et al. 2020; Lo Piano 2020; Gupta et al. 2020). For example, Bossmann (2016) summarized top nine ethical issues in AI: unemployment, inequality, humanity, artificial stupidity (AI can be fooled in ways that humans would not be, e.g., random dot patterns can lead to a machine to "see" things that are not there), racist robots, security, evil genies (AI can fulfill wishes, but with terrible unforeseen consequences), 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 are perceived as ethical, which shows the importance of ensuring ethical AI systems for the positive impact of AI on society (Capgemini 2019). Therefore, it is imperative to identify the right set of fundamental ethical principles and framework to inform the design, regulation, and use of AI and leverage it to benefit as well as respect individuals and societies. An ethical 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 (Dawson et al. 2019). There is debate about both what constitutes "ethical AI" and which ethical requirements, technical standards and best practices are needed for its realization (Jobin et al. 2019).

1.3 AI ethics

Ethics is a branch of philosophy that involves systematizing, defending, and recommending concepts of right and wrong, or good and bad conduct, usually in terms of rights, obligations, benefits to human society, justice, or specific virtues (Dewey and Tufts 2019). It seeks to resolve questions of human morality by defining concepts such as good and evil, right and wrong, justice and crime. Ethics is a well-founded area with philosophers, academics, political leaders and ethicists spending centuries developing ethical concepts and standards. As the primary concern of ethics is on the conduct, there are two popular ethical theoretical perspectives to evaluate conduct: agent-centered ethics and action-centered ethics (Foreman 2014; Hursthouse and Pettigrove 2018). Action-centered ethical theories focus on what an agent should do and how to determine the morally right action in specific circumstances by requiring an agent to follow certain rules or principles. While agent-centered ethical theories aim to develop a good moral character and focus on being rather than just doing.AI ethics is the part of the ethics of technology specific to AI-based solutions. It concerns with the moral behavior of humans as they design, construct, use, and treat artificially intelligent beings, as well as concerns with the moral behavior of AI agents (Jobin et al. 2019). From this perspective, AI ethics considers both action-centered and agent-centered perspectives of conduct of AI. The IEEE report, titled Ethically Aligned Design (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019), 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".

Generally, AI solutions are trained with a large amount of data for different business purposes. Data are 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 contribute to AI ethics. As shown in Fig. 1, AI ethics needs active public debate by considering AI impact, as well as human and social factors (Rovatsos 2019). 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

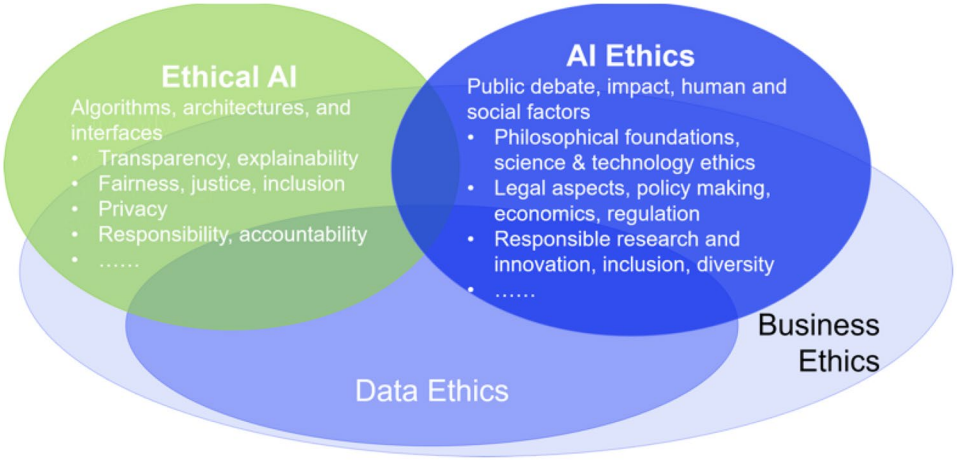

Fig. 1 AI ethics disciplinary landscape [adapted from Rovatsos (2019)]

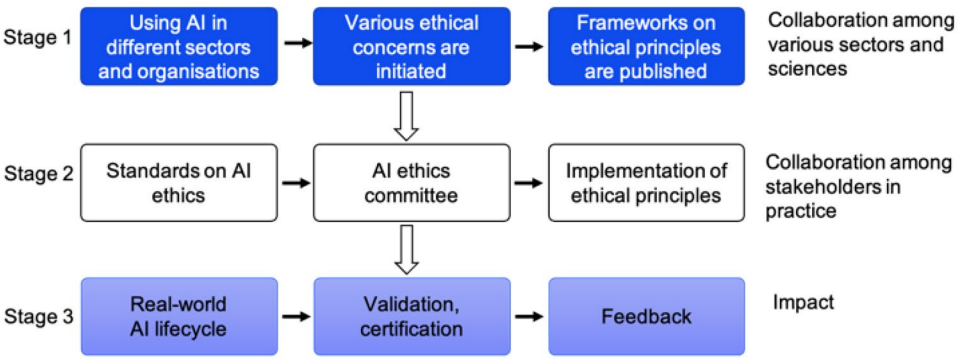

Fig. 2 Three stages of making AI ethics operable. Current work is highly biased towards the stage 1

other ethical standards. Ethical principles of AI refer to ethical principles that AI should follow on the "dos" 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 1 summarizes an overview of AI ethics disciplinary landscape.

This position paper argues for a change in how we view the AI ethics and make it operable for impact from abstract ethical principles to practice. We propose a three-stage approach to make AI ethics operable (see Fig. 2). Stage one is on the identification of ethical principles based on the collaboration among various sectors and sciences. The current work is highly biased towards this stage. Stage two is on the implementation of ethical principles. This stage aims to set up standards on AI ethics and AI ethics committee to make AI ethics operable. This stage needs the close collaborations among stakeholders in practice. The stage three is to apply the implemented ethical principles into real-world applications to validate the compliance of AI with ethical principles for the impact. We seek to stimulate creative thoughts to address the actively discussed issues of making AI ethics operable in practice for impact. The contributions of this work are:

- Identifying a fundamental problem in making AI ethics operable: a lack of collaboration among stakeholders and follow-through;
- suggesting stages towards solving the gaps;
- identifying challenges to direct efforts to make AI ethics operable for impact;
- highlighting a number of key obstacles to AI ethics impact, as an aid for focusing future efforts;
- providing a preliminary practice example to initiate practical implementations of AI ethics.

2 Ethical principles for AI uses

To mitigate various ethical concerns, national and international organizations have made active discussions on ethics of AI within and beyond the AI community (Zhou et al. 2020). Furthermore, professional associations and non-profit organizations 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" (IEEE). This section highlights the way the current AI ethics is conducted today that limits its impact on real-world applications. The goal of this section is to initiate our self-inspection to merit our effort in eliminating them to make AI ethics operable for impact.

2.1 Hyper-focus on the identification of ethical principles

A very large number of ethical principles, codes, guidelines, or frameworks for AI have been proposed over the past few years. Various organizations including governmental and inter-governmental organizations, private sectors, universities, as well as research institutes have made extended efforts by drafting policy documents on ethics of AI and having active discussions on ethics of AI within and beyond the AI community. For example, Jobin et al. (2019) made an in-depth investigation in 2019 on ethical principles of AI and identified 84 documents related to ethical principles or guideline for AI. Algorithm Watch (2020) also maintains an AI ethics guidelines global inventory, which provides a global landscape of AI ethics and is a work in progress. A survey conducted in June 2020 (Gupta et al. 2020) summarizes further efforts on ethics of AI.

The focus of these ethical initiatives on AI is to identify ethical principles that AI should comply with. Various parties identified slightly different ethical principles of AI because of their background or other reasons. For example, ethical principles identified by CSIRO's Data61 in Australia include: human, social and environmental well-being, human-centered values, fairness, privacy protection and security, reliability and safety, transparency and explainability, contestability, and accountability (Dawson et al. 2019). While ethical principles identified by IEEE include: human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse, and competence (the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019).

Jobin et al.'s (2019) investigation found that no single ethical principle is explicitly endorsed by all existing ethical guidelines reviewed, but there is an emerging convergence around the principles of: transparency, justice and fairness, responsibility, non-maleficence, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity, which also shows a developing convergence in the global policy landscape. The survey in 2020 (Bird et al. 2020) identified ethical principles for AI including human rights and well-being; emotional harm; accountability and responsibility; security, privacy, accessibility and transparency; safety and trust; social harm and social justice; lawfulness and justice; control and the ethical use (or misuse) of AI; environmental harm and sustainability; informed use; existential risk.

Floridi and Cowls (2019) analyzed these common ethical principles of interest of AI 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 Jobin et al. (2019). Furthermore, organizations also identified mandatory ethical principles for the use of AI including community benefit, fairness, privacy and security, transparency, and accountability (NSW Government 2020).

Despite the proliferation of ethical principles of AI, there is still an increasing trend of active discussion on ethical principles of AI and publishing guidelines or frameworks by various organizations. First, it is still in its early stage of AI ethics. People still do not have full knowledge on what ethical principles should be complied with and do not fully understand the implications and scope of specific ethical principles in an AI context. Second, AI has been increasingly used by various organizations. They do have ethical concerns on the use of AI. Therefore, various organizations try to show their positions in the ethics of AI. However, ethical principles proposed by various organizations are mostly overlapped with slightly differences or different terms are used for the same or similar ethical issues. Therefore, does every organization need to show their positions and propose their own ethical principles of AI?

This trend has been going for at least for more than ten years. In 2011, the Engineering and Physical Sciences Research Council (EPSRC) and the Arts and Humanities Research Council (AHRC) of United Kingdom jointly published a set of five ethical "principles for designers, builders and users of robots" in the real world, along with seven "high-level messages" intended to be conveyed, based on a September 2010 research workshop with experts from the worlds of technology, industry, the arts, law and social sciences (UKRI 2010). We keep seeing people espousing roughly the same principles, but little work wants to talk about how we genuinely operationalise the assessment of and adherence to those principles in real-world AI deployment contexts until recently. For example, Morley et al. (2021) recently tried to understand the gap between abstract ethical principles and their practical operationalization by surveying a group of participants with a diverse background from startups, large corporations to public sectors, confirming the high demand for operationalizing AI ethics. However, it is still unclear how to translate ethical principles into practice.

2.2 Lack of effective collaboration among stakeholders on AI ethics in practices

Ideally, the successful implementation and deployment of ethical principles of AI in practices need a close collaboration among stakeholders of AI ethics (see Fig. 2), which at least include AI developers, AI users, people impacted by AI uses, an ethics committee who can provide ethical advice, and the parties who set up the standard for AI ethics. AI developers follow AI ethics standards to design and develop AI solutions, while AI users express their ethical concerns to other parties in order to use AI solutions "safely". The parties who set up AI ethics standards need to have a better understanding of other stakeholders, and the ethics committee acts as a bridge between AI ethics standards and AI developers as well as AI users.

Despite the close collaborations between AI developers and AI users in general, they usually focus more on the delivery of AI solutions but not on AI ethics impacted by AI solutions. While such collaborations are important, there is a substantial shortage of collaborations among various stakeholders on AI ethics to make AI ethics operable. Such collaborations will result in the implementation of ethical principles in practice as tools and guidelines to validate the compliance of AI solutions with ethical principles.

2.3 Lack of follow-through

As shown in Fig. 2, the identification of ethical principles is only the first stage in the AI ethics program. They need to be applied to real-world applications for the impact. However, it is very hard to directly use those ethical principles by AI developers or users in the practical applications. Ethical principles should be translated into actionable toolkits and/or 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.

According to Fig. 2, standards on AI ethics are important components to make ethical principles actionable. There are a number of emerging standards that address ethical, legal and social impacts of AI. For example, the IEEE Standards Association has launched a standard via its global initiative on the Ethics of Autonomous and Intelligent Systems. There are currently 14 IEEE standards working groups working on drafting so-called "human" standards that have implications for AI. ISO is also setting up standards for trustworthy AI. However, the standards that do exist are still in development and there is limited publicly available information on them.

While these advances are excellent, there is little incentive to connect these advances with practices to make impact. Connecting active advances of ethical principles to real-world practices is part of the process of maturing of AI ethics. To the general public, these are the only visible advances of AI ethics.

3 Making ethical principles operable

Rather than following "endless" identifications of ethical principles for the use of AI, can we move to the stage of making identified ethical principles operable for impact in real-world applications? This is not simply a matter of another round of discussions among various stakeholders or reporting on isolated applications. What is needed is a fundamental change in or new definition on how we define metrics for validation of ethical principles, how we justify compliances of AI with ethical principles, how we consider differences among different sectors, how we educate the general public to get understanding with AI ethics, how we make AI ethics as a necessary component for each AI solution for impact, and who should monitor the adoption of ethical principles and police their impact as the development and use of AI increases? Answers to these challenging questions need in-depth and close collaborations among various parties and stakeholders such as AI developers, AI users, as well as experts from multiple disciplinaries.

This section initiates discussions by providing example approaches to answer key questions on making ethical principles operable. For example, a committee (such as Human Research Ethics Committee or HREC, Institutional Review Board or IRB) with members from different disciplinary is usually set up to monitor the human related research in organizations. Similarly, an ethical AI committee can be set up to monitor the ethical principles in AI development and deployment. Furthermore, both qualitative and quantitative metrics need to be considered to validate ethical principles. Standards are also indispensable components to justify compliances of AI with ethical principles. It is suggested to incorporate AI ethics into every stage of the whole AI lifecycle as a necessary component for each AI solution for impact. Because of the abstraction of ethical principles, the education on AI ethics is helpful for the improved effectiveness of operationalisation of AI ethical principles. The following subsections give more insightful discussions on these aspects.

3.1 Setup of ethical AI committee

The ethical AI committee aims to consider ethical issues, foster discussion forums, and publish resulting guidance to the industry and regulators. It acts as a bridge between AI ethics standards and AI developers as well as AI users to make sure the compliance of AI solutions with ethical principles. To make ethics of AI operable, the establishment of ethical AI committee with the right expertise and

with the authority to have impact is the first step (Blackman 2022). The primary challenges of this step include who would be the best candidates of committee members, and which areas should they come from. Google 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 and model (Corinium 2019). 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 AI, engineering, law, science, economics, ethics, philosophy, politics, and health. IEEE suggests that the key experts would include but not limited to (IEEE 2018):

- Specialists developing AI-based products and services;
- academic institution experts in AI;
- government organizations involved with AI policy and/or regulations.

We agree the necessity of these experts in the ethical AI committee, other experts such as lawyers who can provide what is legally permissible such as from the anti-discrimination laws' perspectives are also important considerations. Furthermore, business strategists should also be included when AI is used in a business to decide business strategies in addressing AI ethics risks and the investment of time and money.

It is also recommended to set up sector-specific ethical AI committee for the effectiveness and impact. Currently, little news is reported on the setup of an ethical AI committee for practical applications in a country, a state, a sector, or a university.

Besides the member of ethical AI committee, the operation of the ethical AI committee plays a key role in the success of ethical AI committee. Considering the typical function of the ethical AI committee: to identify and help mitigate the ethical risks of AI solutions developed in-house or purchased from third-party vendors. Based on the review, the committee must confirm whether the AI solution poses any ethical risks, recommend changes, or advise against developing or procuring the AI solution. One of the important factors to determine the success of the ethical AI committee is how much authority the committee will have. The option of being ethically sound is only advised is risky for the effectiveness of committee's recommendations, while being ethically sound is must-have is a good idea for committee's recommendations to ensure a real business impact of AI ethics.

Furthermore, the ethical AI committee themselves should also be regulated and reinforced. For example, the work of committee members is regularly reviewed by peers from same or different disciplines, and the membership of the committee is formally promoted to different levels based on the performance of their roles. Considering the differences in local standards, regulations or laws in different regions and countries, this paper argues that the use of the compliance of AI ethics of AI solutions could be limited to the location that the approval is assigned by the AI ethics committee.

3.2 Meaningful and operable validation metrics

Since the nature of ethical principles which are often abstract requirements, the abstract ethical principles need to be translated into meaningful metrics to use them in practical applications. We understand that the ethics and AI belong to two completely different scientific areas. AI can define strict quantitative metrics on its operations such as prediction performance. While ethics is a very abstract social scientific field, it is hard to define quantitative metrics on its requirements, instead, qualitative metrics are often used. Furthermore, the nature of AI ethics is to check the compliance of AI with ethical principles. Therefore, both qualitative and quantitative metrics can be defined to validate ethical principles in AI. These metrics set up links between AI and ethics.

The first consideration for the definition of these metrics should be the meaningfulness for AI users and the operability that can be put into practice by both AI developers and AI users. Focusing our metrics on meaningfulness and operability will motivate the efforts on ethical principles of AI. It will also guide us how to select metrics and how to implement them for impact.

3.3 Standards for validation metrics

When various qualitative and quantitative metrics are available for the validation of ethical principles of AI, we need to decide whether the value of one metric meets the requirement of ethical principles. For example, faithfulness (Alvarez-Melis and Jaakkola 2018) is defined as one of quantitative metrics to evaluate how "good" a particular feature-based local explanation is likely to be. The range of its values is $[-1, 1]$. If the faithfulness of one AI explanation is 0.55, how do we decide whether the faithfulness of the AI explanation "passes" the validation? Let's have another example regarding the validation of access to data in the privacy principle. If a qualitative question is "what are the processes/infrastructure implemented to restrict access to user data" and the answer to it is "secure protocols are used

to connect with customer systems using Transport Layer Security 1.2 for HTTPS encryption", how do we justify whether the answer to this question "passes" the validation?

Therefore, a standard for validation metrics is highly necessary to help stakeholders justify the validation of ethical principles in practice. The lack of standards or regulations/laws is one of major factors that prevent the AI ethics committee from not reaching agreement on whether an AI research proposal should be approved. Despite standards on AI ethics such as the IEEE 7000 series of standards emerging, they are still under development and no details are available. Furthermore, different sectors may have different requirements from the ethical perspective. Different sectors can set up further implementation details regarding the sector according to the general standards for the impact. 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, the development of implementation details for ethical principles for a specific sector according to general ethical principles and standards is more effective for the implementation of ethical principles.

Government plays a significant role in this process and can publish policies to guide the setup of the standards or laws for validation metrics. For example, New York City's Local Law #144 on "Automated employment decision tools"[1] requires that the bias audit of automated employment decision tools should meet 4/5th Rule. The 4/5th Rule states that there is adverse impact on a certain group if the selection rate for that group is less than 80 percent (4/5) of that of the group with the highest selection rate.

Therefore, this paper argues that in a typical AI ethics approval process, AI researchers/suppliers provide answers to qualitative and quantitative metrics for AI ethical principles. The AI ethics committee makes judgements according to standards and laws. The ethics committee needs to include members spanning different fields such as AI, engineering, law, science, economics, ethics, philosophy, politics, and health to make reasonable approval decisions.

3.4 AI ethics for the whole AI lifecycle

The lifecycle of a typical AI application usually includes different stages from business and use-case development, design phase, training and test data collection, building AI application, testing the system, deployment of the system to monitoring performance of the system. The lifecycle of an AI application 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. (2019) 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. Both action-centered ethics and agent-centered ethics can be considered in the AI lifecycle.

In the AI ethics for the whole AI lifecycle, different stages of AI lifecycle may have different emphasis on ethical principles. For example, in the data procurement stage, the data privacy is the core principle, while in the AI application building stage, stakeholders are more interested in the model transparency. Therefore, we should implement ethical principles at every stage of the AI lifecycle while also giving different emphasis on different ethical principles at different stages of the AI lifecycle.

3.5 Education of AI ethics

Ethical principles are abstract and AI ethics is especially difficult to understand by AI users including layman users. The better understanding of the AI ethics will largely benefit the implementation of ethical principles of AI and the overall impact of AI ethics in AI applications. Therefore, the education on AI ethics is helpful for stakeholders to better understand ethical principles, resulting in the improved effectiveness of implementations of ethical principles of AI and boost the impact of AI ethics.

Short courses are a viable approach to educate key concepts and knowledge on AI ethics. For example, some universities have developed such courses and provided to the public (Zhou et al. 2021). Coursera also offers similar short courses with project-based approach. Some educational tools for teaching AI ethics have also been developed. For example, value cards (Shen et al. 2021) is an educational toolkit to inform students and practitioners the social impacts of different machine learning models via deliberation. However, more studies on the education of AI ethics are still necessary to investigate approaches to educate various stakeholders of AI solutions effectively.

[1] https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9 .

4 Challenges for making AI ethics operable

Ambitious and meaningful challenges are helpful to direct efforts to make AI ethics operable. The following challenges are articulated as examples of making AI ethics operable that matters.

- **Standardization.** The standardization can help to erode the complexity and variations of ethical principles of AI to make AI ethics operable. The standards are not only on which ethical principles should be validated for AI solutions, but also on how those ethical principles should be validated, and what are the criteria that AI solutions "pass" the validation of a specific ethical principle.
- **Quantifying ethical values in AI ethics.** We understand AI can create substantial economic value to humans. Similarly, can we also quantify values imposed by ethics when we implement ethical principles of AI? An example of approaches to quantify values imposed by ethics could be the evaluation of economic value difference of AI before and after applying AI ethics. The improvement of social impact can also be used to quantify values imposed by ethics. Such quantification of ethical values will help us to justify other challenges as shown below.
- **Balancing the economic value created by AI to be ethical or unethical**. AI and related forms of automation can create substantial economic value to humans in various fields. Therefore, we have to confront the problem where economic value created by ethical AI and unethical AI conflict (Korinek 2020), raising the challenge of balancing ethical and unethical by considering economic values when developing AI technologies. For example, AI and related forms of automation affect labor markets significantly, which may lead to significant increases in inequality for human labor. How we balance the use of AI and ethical concerns?
- **Balancing the AI performance and ethical values.** In some cases, there is a contradiction between AI performance and ethical principles. For example, the mitigation of model bias may have adverse effects on AI performance. Therefore, we have the challenge of balancing the AI performance and ethical values.

These challenges seek to capture the key components in making AI ethics operable including the standardization, value quantification, and value balancing. The goal is to inspire the field of AI ethics to take steps needed to mature into valuable contributions to making AI ethics operable besides the exploration of ethical principles. Furthermore, no such list can claim to be comprehensive. It is hoped that this paper can help to inspire researchers and stakeholders to formulate additional challenges that benefit making AI ethics operable.

5 Obstacles to making AI ethics actionable

Let us imagine an AI ethics researcher who is motivated to tackle the problem of implementing ethical principles to make AI ethics operable. What obstacles to success can we foresee? The following are typical examples of obstacles we observe from current practices.

- **Communication**. The smooth communication between AI research field and other disciplines such as ethics, philosophy, law and social science is significant for making AI ethics operable. However, the abstract concepts in one field may be difficult to understand or find corresponding concepts in another field, which serve as a barrier for smooth communication between different fields. For example, considering the concepts of "feature extraction", "cross-validation", "variance", and "mutual information" which are basic concepts within AI, people outside of AI may not easily understand these concepts. We need to "translate" these concepts to terms that can be understood by people from other fields. For example, "feature extraction" can be expressed as "representation", and "cross-validation" is also known as "out-of-sample testing". Similarly, the terms in other fields need to be "translated" into concepts in AI so that corresponding AI approaches can be developed.
- **Complexity**. Despite the proliferation of ethical principles of AI, the field has not yet matured to a point where users from an application domain can simply apply the ethical principles in their applications to make sure AI solutions are ethical. This is mainly due to the lack of knowledge of what ethical principles need to be applied, how to validate ethical principles, and whether there is a standard to justify that AI solutions meet requirements of specific ethical principles. Furthermore, AI ethics is related to different fields including engineering, law, science, economics, ethics, philosophy, politics, health, and others. Therefore, while AI ethics itself is an abstract activity, simplifying and maturing tools can help relieve this obstacle and permit wider and independent validations of ethical principles of AI.
- **Subjectivity**. Because of the nature of ethics, the validation of ethical principles can mostly be checked with subjective questionnaires. Such subjectivity makes it difficult to have an objective justification of the compliance of AI with ethical principles. For this reason, the design of more objective methods to validate the compliance

of AI with ethical principles can erode this obstacle and make ethical principles of AI operable.

- **Risk**. The violation of one ethical principle may cause certain risks. Furthermore, the risk degree can be different because of the violation of different ethical principles. For example, in an AI-assisted credit risk prediction system, the violation of privacy of personal data may cause higher risks than the violation of fairness of AI. Therefore, we also need to associate risks caused by the violation of ethical principles with the validation of relevant ethical principles if we hope to infuse AI ethics into real-world applications. However, the identification of risks for the violation of ethical principles is a difficult task.
- **Government policies/rules**. Government plays a significant role in the AI ethics, which can be seen from strategies and initiatives on AI published by various governments (NSW Government 2020). Further actions can be taken by governments in order to make AI ethics operable. For example, policies/rules can be set up to regulate how AI should be acted ethically in public services, which could also function as standards that AI solutions follow. This is related to different fields at least including law, politics, AI, and ethics.

These are also not the comprehensive list. For example, different sectors may also have different obstacles. Furthermore, much effort has been put to eliminate these kinds of obstacles. For example, we can see increasing collaborations and communications among different disciplines from AI, law, philosophy, and social science to solve challenges on AI ethics. Governments from different countries have set up various policies/rules to regulate AI development and applications, or are actively discussing policies/rules to be set up. These are effective attempts to overcome obstacles to making AI ethics operable. It is also hoped that this paper can motivate researchers and stakeholders to formulate potential obstacles and propose effective approaches to eliminate obstacles in making AI ethics operable.

6 Our practices

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 questions. 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 questions aim to evaluate the compliance of AI with ethical principles by developing quantitative approaches to measure the compliance of AI with ethical principles. For example, quantitative measures are developed to validate the explainability 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 application to allow the effective validation of ethical principles for AI solutions. The main components of the platform include: users of the platform (AI suppliers, validators, and administrators), projects to be validated, questionnaires, and validation outputs. Questionnaires can be customized for different projects to meet specific requirements.

By considering the dynamics of investigations on ethical principles of AI, the 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.

7 Conclusion

AI ethics is becoming one of the most discussed topics in recent years as AI is widely used in different domains for prediction, automation, planning, targeting, and personalization as well as others. This leads to the "principle proliferation" for 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 to implement ethics in AI in practical applications. This paper suggested three stages to solve the gap, which are identification of ethical principles based on the collaboration among various sectors and sciences, the implementation of ethical principles, and applying the implemented ethical principles into real-world applications to validate the compliance of AI with ethical principles for the impact. We highlighted the current the way the current AI ethics is conducted that limits its impact on real-world applications, to initiate our self-inspection to merit our effort in eliminating them to make AI ethics operable for impact. Furthermore, this paper initiated discussions by providing example approaches to answer key questions on making ethical principles operable, such as setup of ethical AI committee, meaningful and operable validation metrics, standards for validation metrics, AI ethics for the whole AI lifecycle, and education of AI ethics. Key example challenges were articulated to inspire the field of AI ethics to take steps needed to mature into valuable contributions to making AI ethics operable besides the exploration of ethical principles. The paper also foresaw

key obstacles to AI ethics impact. A preliminary practice example was provided to initiate practical implementations of AI ethics. This position paper was not to provide specific solutions on making AI ethics operable, but to articulate key aspects to consider in making AI ethics operable and initiate discussions on steps needed to take in making AI ethics operable. Aiming for the real impact of AI ethics is not only to identify ethical principles, but also to build standards, implement them effectively, and deploy them easily in practical applications.

Data statements This paper has no associated data.

Declarations

Conflict of interest Authors state that there is no conflict of interest.

References

AlgorithmWatch (2020) AI ethics guidelines global inventory. https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/. Accessed 18 Oct 2020

Alvarez-Melis D, Jaakkola TS (2018) Towards robust interpretability with self-explaining neural networks. arXiv:1806.07538 [cs, stat]

Amrollahi M, Hadayeghparast S, Karimipour H et al (2020) Enhancing network security via machine learning: opportunities and challenges. In: Choo K-KR, Dehghantanha A (eds) Handbook of big data privacy. Springer International Publishing, Cham, pp 165–189

Anastasopoulos LJ, Whitford AB (2019) Machine learning for public administration research, with application to organizational reputation. J Public Adm Res Theory 29:491–510. https://doi.org/10.1093/jopart/muy060

Awoyemi JO, Adetunmbi AO, Oluwadare SA (2017) Credit card fraud detection using machine learning techniques: a comparative analysis. In: 2017 international conference on computing networking and informatics (ICCNI), pp 1–9

Batmaz Z, Yurekli A, Bilge A, Kaleli C (2019) A review on deep learning for recommender systems: challenges and remedies. Artif Intell Rev 52:1–37. https://doi.org/10.1007/s10462-018-9654-y

Bird E, Fox-Skelly J, Jenner N, et al (2020) The ethics of artificial intelligence: issues and initiatives. European Parliamentary Research Service

Blackman R (2022) Why you need an AI ethics committee. Harvard Bus Rev. July-August 2022

Bojarski M, Del Testa D, Dworakowski D et al (2016) End to end learning for self-driving cars. arXiv preprint. arXiv:1604.07316

Bossmann J (2016) Top 9 ethical issues in artificial intelligence. In: World economic forum. https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/. Accessed 9 Sept 2019

Capgemini (2019) Why addressing ethical questions in AI will benefit organizations. In: Capgemini worldwide. https://www.capgemini.com/research/why-addressing-ethical-questions-in-ai-will-benefit-organizations. Accessed 10 Oct 2021

Corinium (2019) Ethics of AI. https://cdn2.hubspot.net/hubfs/2631050/CDAO%20New%20Zealand/Corinium_Ethics-of-AI_brochure_NZ.pdf. Accessed 8 Aug 2019

Dawson D, Schleiger E, Horton J, et al (2019) Artificial intelligence—Australia's ethics framework. In: Data61, CSIRO, Australia

Deane M (2018) AI and the future of privacy. In: Towards data science. https://towardsdatascience.com/ai-and-the-future-of-privacy-3d5f6552a7c4. Accessed 16 May 2019

Dewey J, Tufts JH (2019) Ethics. Good Press, Glasgow

Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Sci Rev. https://doi.org/10.1162/99608f92.8cd550d1

Foreman E (2014) An agent-centered account of rightness: the importance of a good attitude. Ethical Theory Moral Pract 17:941–954

Gao X, Tian Y, Qi Z (2020) RPD-GAN: learning to draw realistic paintings with generative adversarial network. IEEE Trans Image Process 29:8706–8720. https://doi.org/10.1109/TIP.2020.3018856

Gupta A, Lanteigne C, Heath V et al (2020) The state of AI ethics report (June 2020). arXiv:2006.14662 [cs]

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Hursthouse R, Pettigrove G (2018) Virtue ethics. In: Zalta EN (ed) The stanford encyclopedia of philosophy, winter 2018 Edition, Available online: https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/ (This is an online encyclopedia published by Stanford. There is no publisher name and location specificlly. If we need to provide, it could be "Center for the Study of Language and Information, Stanford Univeristy", (this is from the webpage))

IEEE (2019) The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In: IEEE Standards Association. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

IEEE (2018) IEEE launches ethics certification program for autonomous and intelligent systems. In: IEEE Standards Association. https://standards.ieee.org/news/2018/ieee-launches-ecpais.html. Accessed 18 Sept 2019

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399

Korinek A (2020) Integrating ethical values and economic value to steer progress in artificial intelligence. In: Dubber MD, Pasquale F, Das S (eds) The Oxford handbook of ethics of AI. Oxford University Press, Oxford

Lo Piano S (2020) Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanit Soc Sci Commun 7:1–7

Monroe D (2017) Deep learning takes on translation. Commun ACM 60:12–14. https://doi.org/10.1145/3077229

Mooney SJ, Pejaver V (2018) Big data in public health: terminology, machine learning, and privacy. Annu Rev Public Health 39:95–112. https://doi.org/10.1146/annurev-publhealth-040617-014208

Morley J, Floridi L, Kinsey L, Elhalal A (2019) From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. arXiv:1905.06876 [cs]

Morley J, Kinsey L, Elhalal A et al (2021) Operationalising AI ethics: barriers, enablers and next steps. AI Soc. https://doi.org/10.1007/s00146-021-01308-8

Nguyen TT, Nguyen CM, Nguyen DT et al (2020) Deep learning for deepfakes creation and detection: a survey. arXiv:1909.11573 [cs, eess]

NSW Government (2020) Mandatory ethical principles for the use of AI. In: Artificial intelligence (AI) ethics policy. https://www.digital.nsw.gov.au/policy/artificial-intelligence-ai/artificial-intelligence-ai-ethics-policy/mandatory-ethical.Accessed 20 Oct 2020

Rolnick D, Donti PL, Kaack LH et al (2019) Tackling climate change with machine learning. arXiv:1906.05433 [cs, stat]

Rovatsos M (2019) From AI ethics to ethical AI. IJCAI 2019 tutorial, Macau, China

Samuel AL (1960) Some moral and technical consequences of automation—a refutation. Science 132:741–742. https://doi.org/10.1126/science.132.3429.741

Shen H, Deng WH, Chattopadhyay A, et al (2021) Value cards: an educational toolkit for teaching social impacts of machine learning through deliberation. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, NY, USA, pp 850–861

Taddeo M, Floridi L (2018) How AI can be a force for good. Science 361:751–752. https://doi.org/10.1126/science.aat5991

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems. IEEE

Tolosana R, Vera-Rodriguez R, Fierrez J et al (2020) DeepFakes and beyond: a survey of face manipulation and fake detection. arXiv:2001.00179 [cs]

UKRI (2010) Principles of robotics. In: Engineering and physical sciences research council. https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/. Accessed 4 Nov 2020

Wiener N (1960) Some moral and technical consequences of automation. Science 131:1355–1358. https://doi.org/10.1126/science.131.3410.1355

Zhao S, Gao Y, Jiang X et al (2014) Exploring principles-of-art features for image emotion recognition. In: Proceedings of the 22nd ACM international conference on multimedia. Association for Computing Machinery, New York, NY, USA, pp 47–56

Zhou J, Chen F (2019) AI in the public interest. In: Bertram C, Gibson A, Nugent A (eds) Closer to the machine: technical, social, and legal aspects of AI. Office of the Victorian Information Commissioner, Melbourne, Australia

Zhou J, Chen F (eds) (2018) Human and machine learning: visible, explainable, trustworthy and transparent. Springer, Berlin

Zhou J, Chen F, Berry A, et al (2020) A survey on ethical principles of AI and implementations. In: Proceedings of 2020 IEEE symposium series on computational intelligence (IEEE SSCI), Canberra, Australia

Zhou J, Chen F, Berry A (2021) AI ethics: from principles to practice. https://open.uts.edu.au/uts-open/study-area/Technology/ethical-ai-from-principles-to-practice/. Accessed 15 Aug 2021

Zhou J, Li Z, Zhi W et al (2017) Using convolutional neural networks and transfer learning for bone age classification. In: 2017 international conference on digital image computing: techniques and applications, DICTA 2017, Sydney, Australia, November 29–December 1, 2017. IEEE