An Overview of Artificial Intelligence Ethics

Article *in* IEEE Transactions on Artificial Intelligence · July 2022

DOI: 10.1109/TAI.2022.3194503

4 authors, including:

Changwu Huang
Southern University of Science and Technology
32 PUBLICATIONS **604** CITATIONS

SEE PROFILE

An Overview of Artificial Intelligence Ethics

Changwu Huang, *Member, IEEE*, Zeqi Zhang, Bifei Mao, and Xin Yao, *Fellow, IEEE*

***Abstract*—Artificial intelligence (AI) has profoundly changed and will continue to change our lives. AI is being applied in more and more fields and scenarios such as autonomous driving, medical care, media, finance, industrial robots, and internet services. The widespread application of AI and its deep integration with the economy and society have improved efficiency and produced benefits. At the same time, it will inevitably impact the existing social order and raise ethical concerns. Ethical issues, such as privacy leakage, discrimination, unemployment, and security risks, brought about by AI systems have caused great trouble to people. Therefore, AI ethics, which is a field related to the study of ethical issues in AI, has become not only an important research topic in academia, but also an important topic of common concern for individuals, organizations, countries, and society. This paper will give a comprehensive overview of this field by summarizing and analyzing the ethical risks and issues raised by AI, ethical guidelines and principles issued by different organizations, approaches for addressing ethical issues in AI, methods for evaluating the ethics of AI. Additionally, challenges in implementing ethics in AI and some future perspectives are pointed out. We hope our work will provide a systematic and comprehensive overview of AI ethics for researchers and practitioners in this field, especially the beginners of this research discipline.**

***Impact Statement*—AI ethics is an important emerging topic among academia, industry, government, society, and individuals. In the past decades, many efforts have been made to study the ethical issues in AI. This article offers a comprehensive overview of the AI ethics field, including the summary and analysis of AI ethical issues, ethical guidelines and principles, approaches to address AI ethical issues, and methods to evaluate the ethics of AI. Additionally, some challenges and future perspectives are discussed. This article will help researchers to obtain sufficient background and a bird's eye view of AI ethics, and thus facilitate their further investigation and research.**

***Index Terms*—Artificial Intelligence, AI Ethics, Ethical Issue, Ethical Theory, Ethical Principle.**

I. Introduction

Artificial intelligence (AI) [1] has achieved rapid and remarkable development during the last decade. AI technologies such as machine learning (ML), natural language processing, and computer vision are increasingly permeating and spreading to various disciplines and aspects of our society. AI is increasingly taking over human tasks and replacing human decision-making. It has been widely used in a variety of sectors, such as business, logistics, manufacturing, transportation, health care, education, state governance and etc.

The application of AI has brought about efficiency improvement and cost reduction, which are beneficial for economic growth, social development, and human well-being [2]. For instance, the AI chatbot can respond to clients' inquiries at any time, which will improve the customers' satisfaction and the company's sales [3]. AI allows doctors to serve patients in remote locations through telemedicine services [4]. It is no doubt that the rapid development and wide application of AI are already affecting our daily life, humanity, and society.

However, at the same time, AI also poses many significant ethical risks or issues for users, developers, humans, and society. Over the past few years, many cases in which AI produced poor outcomes have been observed. For instance, in 2016, the driver of an electric Tesla car was killed in a road accident after its Autopilot mode failed to recognize an oncoming lorry [5]. Microsoft's AI chatting bot, Tay.ai ,was taken down because it became racist and sexist only less than a day after she joined Twitter [6]. There are many other examples concerned with the failure, fairness, bias, privacy, and other ethical issues of AI systems [7]. More seriously, AI technology has begun to be used by criminals to harm others or the society. For example, criminals used AI-based software to impersonate a chief executive's voice and demand a fraudulent transfer of $ 243,000 [8]. Therefore, it is urgent and critical to address the ethical issues or risks of AI so that AI can be built, applied, and developed ethically.

AI ethics or machine ethics [9] is an emerging and interdisciplinary field concerned with addressing ethical issues of AI [10]. AI ethics involves the ethics of AI, which studies the ethical theories, guidelines, policies, principles, rules, and regulations related to AI, and the ethical AI, that is, the AI that can uphold ethical norms and behaves ethically [11]. The ethics of AI is a prerequisite to building ethical AI or to making AI behave in an ethical manner. It involves the ethical or moral values and principles that determine what is morally right and wrong. With appropriate ethics of AI, ethical AI can be built or implemented through some methodologies and technologies.

Even though AI ethics has been extensively discussed by interdisciplinary researchers for several years, it is still in its infancy [11]. AI ethics is a very broad and rapidly developing

Changwu Huang and Xin Yao are with the Research Institute of Trustworthy Autonomous Systems (RITAS), Southern University of Science and Technology, Shenzhen 518055, China and Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: huangcw3@sustech.edu.cn, xiny@sustech.edu.cn). Xin Yao is also with School of Computer Science, University of Birmingham, UK.

Zeqi Zhang and Bifei Mao are with Trustworthiness Theory Research Center, Huawei Technologies Co., Ltd., Shenzhen 518055, China (e-mail: zhangzeqi1@huawei.com, maobifei@huawei.com).

Corresponding author: Xin Yao.

research area that has received increasing attention from researchers in recent years. Although several review papers have been published during the past few years, each of them focuses on a certain aspect(s) of AI ethics, and there is still a lack of comprehensive reviews to provide a full picture of this field. For instance, a brief review of ethical issues in AI was provided in [11], AI ethics guidelines and principles were investigated in [12] and [13], [14] focused on bias and fairness in ML, [15] only reviewed the safety in reinforcement learning, [16] reviewed the security and privacy of federated learning, [17] dedicated to a survey of privacy and security issues in deep learning, [18] concentrated on explainable AI, [19] covered the key ethical and privacy issues in AI and traced how such issues have changed over the past few decades using the bibliometric approach. Thus, this paper is dedicated to presenting a systematic and comprehensive overview of AI ethics from diverse aspects (or topics), thereby providing informative guidance for the community to practice ethical AI in the future. We hope it will inform scientists, researchers, engineers, practitioners, and other relevant stakeholders, and provides sufficient background, comprehensive domain knowledge and a bird's eye view for interested people, especially for the beginners of this research discipline, so that further investigation and improvement can be pursued by them.

The main contributions of this article are as follows:

1) A comprehensive overview of AI ethics, including ethical issues and risks of AI, ethical guidelines and principles for AI, approaches for addressing ethical issues in AI, and methods for evaluating ethical AI, is provided in this review. This overview can provide a sufficient background, comprehensive domain knowledge, and a roadmap for researchers and practitioners.
2) The ethical issues and risks caused by AI are summarized, and a new categorization of AI ethical issues is proposed in Section III. The proposed new categorization is helpful for recognizing, understanding, and analyzing ethical problems in AI and then developing solutions to solve these problems. Additionally, the ethical issues associated with different stages of AI system's lifecycle are discussed.
3) An up-to-date global landscape of the AI ethics guidelines and principles is presented in Section IV, based on 146 guidelines related to AI ethics released by companies, organizations, and governments around the world. These guidelines and principles provide a high-level guidance for the planning, development, production, and usage of AI and directions for addressing AI ethical issues.
4) A review of multi-disciplinary approaches to addressing AI ethical problems, including ethical, technological, and legal approaches, is given in Section V. This not only provides an informative summary about the approaches to ethical AI but also suggests potentially different solutions to AI ethical issues from a variety of perspective rather than relying solely on technological approaches.
5) Methods for assessing or evaluating AI ethics are reviewed in Section VI. Testing or evaluating whether an AI system meets the ethical requirements or not is an essential part of AI ethics. However, this aspect is often overlooked in the existing literature. To the best of our knowledge, this paper is the first to summarize the aspect of evaluating ethical AI.
6) Lastly, some challenges in AI ethics and several future perspectives are pointed out, which provide some research questions and directions for further research in the future. This will be helpful for interested researchers and practitioners to pursue further research in AI ethics field.

The rest of the paper is organized as follows. After this introductory section, we briefly describe the review scope and methodology of this paper in Section II. A comprehensive summary of the ethical issues and risks raised from AI is given in Section III. Section IV reviews and analyzes the AI ethical guidelines and principles that have been released during the last few years. Section V describes the paradigms or approaches for addressing ethical issues in AI. Section VI discusses the approaches to evaluate the morality or ethics of AI systems or products. Section VII outlines the challenges in implementing ethics in AI and gives some future perspectives on designing ethical AI. Section VIII briefly concludes this paper.

II. Scope and Methodology

In this section, we first clarify the aspects and topics covered in this review and the links between these topics. Then we describe the methodology followed in conducting this survey, including the literature search strategy and selection criteria.

A. Scope

The scope and topics of this paper is described as follows. Investigation of ethical issues and risks of AI is the starting point of this review, since it is because of the existence of ethical issues in AI that the research field of AI ethics exists. Thus, it is necessary and important to clarify and understand the ethical problems existed in AI. Then, the ethical guidelines and principles, which direct the development and use of AI, are reviewed. As the ethical issues of AI have attracted more and more attention from various sectors of our society, many organizations (including academia, industry, and governments) have begun to discuss and seek possible frameworks, guidelines, and principles for solving AI ethics issues. These guidelines and principles provide valuable directions for practicing ethical AI. After clarifying the existing ethical issues and guidelines, we review the approaches to solving the ethical issues in AI. We covered ethical, technological, and legal approaches, but focus more on the first two kinds of approaches (ethical and technological approaches) since the researchers in AI community may be more interested in these two categories of approaches. Last but not least, we summarize how to evaluate ethical AI, which is to assess the ethicality or morality of AI, i.e., how well the ethical problems are addressed or whether an AI system meets the ethical requirements or not. Apparently, these four aspects are essential for solving ethical issues in AI. Thus, the above four aspects constitute the main content of this paper and provide a systematic overview of AI ethics. The topics or aspects covered in this paper and the links between them are illustrated in Fig. 1.

B. Methodology

This review covers a wide variety of documents, including academic, organizational, government grey literature sources and news report. The search of relevant literature was conducted in two phases. In the first phase, the entries or keywords that reflect different terms related to AI ethics are used to search on Google Scholar, Web of Science, IEEE Xplore, ACM Digital Library, Science Direct, Springer Link, arXiv and Google. The entries or keywords used include: (ethics, ethical, responsibility, responsible, trustworthiness, trustworthy, transparent, explainable, fair, beneficial, robust, safe, private, sustainable) AND/OR (issues, risks, guideline, principle, approach, method, evaluation, assessment, challenge) AND (artificial intelligence, AI, machine learning, ML, intelligent system, intelligent agent). We mainly consider the literature published or released since 2010 and included as many related keywords as possible in titles. In the second phase, we checked the related work of literature found in the first phase, such as the cited articles and other work by the same authors of phase one.

As for the ethical AI guidelines, we only collected these documents in English (or with official English translations) and can be visited or downloaded on the internet. A full list with URL links of collected ethical AI guidelines is provided in the Supplementary Materials of this paper.

Fig. 1 The topics covered in this paper and the links between them.

III. Ethical Issues and Risks of AI

To address the ethical problems of AI, we must first recognize and understand the potential ethical issues or risks that AI may bring. Then, the necessary AI ethical guidelines, policies, principles, rules (i.e., Ethics of AI) can be formulated appropriately. With the adequate ethics of AI, we can design and build AI that behaves ethically (i.e., Ethical AI) [8]. The ethical issue of AI generally refers to the morally bad things or problematic outcomes relevant to AI (i.e., these issues and risks that are raised by the development, deployment, and use of AI) that need to be addressed. Many ethical issues, such as lack of transparency, privacy and accountability, bias and discrimination, safety and security problems, the potential for criminal and malicious use, and so on, have been identified from the applications and studies.

This section focuses on ethical issues and risks surrounding the use of AI. Firstly, four different categorizations of AI ethical issues in the literature are reviewed in Section III.A. Since these four categorizations either ignore some ethical issues or are too complicated to understand, we proposed a new categorization that classifies AI ethical issues into individual, societal, and environmental levels in Section III.B. Our proposed categorization comprehensively covers the existing ethical issues and is easy to understand, which is helpful for understanding and analyzing the ethical problems caused by AI. Besides, we attempt to map the ethical issues associated with the stages of AI system's lifecycle in Section III.C. This would be beneficial for figuring out these issues during the AI system development process.

The main goal of this section is to discuss and clarify the ethical issues of AI so that practitioners can recognize and understand these issues, and then help them to further study how to address AI ethical issues. The main contribution in this section is that we proposed a new categorization of AI ethical issues, which covers the ethical issues discussed in a clear and easy-to-understand manner. Additionally, the ethical issues associated with the stages of AI system's lifecycle is discussed.

A. Review of Categorizations of AI Ethical Issues

This subsection describes the ethical concerns or issues of AI from different perspectives by reviewing four different categorizations that were found in our collected literature. Two of them are from government reports and the other two are from academic publications. From different perspectives and categorizations, the ethical issues involved are also somewhat different. In the following, four different categorizations of AI ethical issues are reviewed subsequently. The four reviewed categorizations of AI ethical issues and our proposed categorization are listed in Table I.

1) Categorization Based on Features of AI, Human Factors and Social Impact

In [11], AI ethical issues are mainly discussed in three categories: ethical issues caused by the features of AI, ethical risks caused by human factors, and social impact of ethical AI issues.

a) Ethical Issues Caused by Features of AI

(1) Transparency. ML is the core technology of current AI, especially (deep) neural networks. However, it is hard to explain and understand the inference procedure of ML, which is commonly known as the "black-box". The opacity of ML makes the algorithms or models mysterious to users and even developers. This mainly leads to the transparency issue [20]. The lack of transparency not only leads to the explanatory problem, but also leads to difficulties in human monitoring and guidance of ML or AI. Thus, transparency or explainability is one of the most widely discussed downside of AI.

(2) Data Security and Privacy. The performance of current AI strongly depends on the training data. Usually, a huge amount of data, which probably includes personal data and private data, is required to train an AI model, particularly the deep learning model. The misuse and malicious use of data, such as (personal) information leakage or tampering, are serious ethical issues that are closely related to every individual,

institution, organization, and even the country. Data security and privacy are key issues encountered in the development and application of AI technology [21].

(3) Autonomy, Intentionality, and Responsibility. With the advancement of AI, current AI systems or agents, such as healthcare robots, have a certain degree of autonomy, intentionality, and responsibility [22]. Here, the autonomy of AI refers to an AI system's ability to operate without human intervention or direct control. Intentionality refers to the ability that an AI system can act in a way that is morally harmful or beneficial and the actions are deliberate and calculated [11]. Responsibility indicates that the AI system fulfill some social rule and some assumed responsibilities. However, how much autonomy, intentionality, and responsibility should an AI system be allowed is a challenging question and issue.

b) Ethical Issues Caused by Human Factors

(1) Accountability. When an AI system or agent fails in a specified task and results in bad consequences, who should be responsible. The undesirable consequence may be caused by many factors, such as the programming codes, input data, improper operation, or other factors. This brings about the so-called "the problem of many hands" [23]. Thus, accountability is an ethical issue that concerns the human factors involved in the designing, implementation, deployment, and usage of AI.

(2) Ethical Standards. As the ultimate goal of AI ethics is to create ethical AI that can follow ethical principles and behave ethically [10], it is crucial to form comprehensive and unbiased ethical standards for training or regulating AI to be ethical. To formulate ethical standards for AI, researchers and practitioners should well understand the existing ethical theories and principles [13][24].

(3) Human Rights Laws. The designer, software engineers and other participants in AI system design and application should be taught human rights laws [25]. Without training in human rights laws, they may infringe and breach essential human rights without even realizing it. The human rights laws or acts followed by different countries or regions are often different. Many different human rights laws, for instance, International Human Rights Law, International Covenant on Civil and Political Rights, International Covenant on Economic, Social and Cultural Rights, Universal Declaration of Human Rights, Charter of the United Nations, the European Convention for the Protection of Human Rights and Fundamental Freedoms and etc. [26], have been released by different governments.

c) Social Impact of Ethical AI Issues

(1) Automation and Job Replacement. As more and more factory workers are being replaced by automated systems and robots, AI will disrupt and transform the labor market. Hence, many people worry about automation and job replacement [27].

(2) Accessibility. The accessibility or availability of emerging technologies, such as AI, will have a direct impact on human well-being. However, it will be unethical and unfair if only a portion of the population benefit. Consideration must be given to developing AI products and services that are accessible to everyone, and thus the benefits of AI can be spread equally to everyone [28].

(3) Democracy and Civil Rights. Unethical AI will distort the truth and eventually lead to the loss of trust and public support for AI technology [11]. The strengths of democracies are harmed by the loss of informed and trusting communities. As democracies suffer and structural biases exacerbated, the free enjoyment of civil rights is no longer consistently available to all. Thus, democracy and civil rights must be taken into consideration in AI ethics.

2) Categorization Based on Vulnerabilities of AI and Human

In [29], the authors distinguished the ethical issues of AI into 1) ethical issues that arise because of limitations of current ML systems, which is named as "vulnerabilities in AI (especially ML)," and 2) ethical issues that arise because current ML systems may be working too well and humans can be vulnerable in the presence of or interaction with these intelligent systems, which is referred to as "human vulnerabilities".

a) Ethical Issues from the Vulnerabilities of AI

(1) ML is Data Hungry. Usually, ML requires a large amount of data to work well [30]. Therefore, this motivates companies and organizations to collect or purchase data, including sensitive personal data, even if doing so may violate the individual's right to privacy.

(2) Garbage In/Garbage Out. The performance of a ML algorithm heavily depends on the data from which it learns. If one ML algorithm is trained on insufficient or inaccurate data, it will provide undesirable results even it is well designed [31].

(3) Faulty Algorithms. Even if a ML algorithm is input with enough and accurate data, if the algorithm itself is bad, it will also make bad predictions. For example, a bad ML algorithm may not be able to recognize a pattern even if there is one or it may recognize a pattern even if there is not one, where are known as "underfitting" and "overfitting", respectively [32].

(4) Deep Learning Is a Black Box. Deep learning is a black box, which raises issues such as explainability, interpretability, and trust [33]. Even for the designers and developers of deep learning, the model is incomprehensible since it usually involves thousands or millions of connections between different neurons. Therefore, it is difficult to explain how these connections interact and why the model makes certain predictions.

b) Ethical Issues from the Vulnerabilities of Human

(1) Abuse of AI. AI technologies, such as facial recognition and image generation, can work better than humans [34]. However, ethical issues exist because people may be tempted to use them for ill. For instance, a government could use facial recognition technology to monitor its citizens, and ML can be used to fabricate photos or videos so realistic that humans cannot tell that they are fake [35]. This brings the concern about the abuse of AI technologies.

(2) Job Replacement. Since intelligent robots can perform certain tasks faster and better than humans, many people worry that robots and other AI technologies will replace a large part of current human labor in the near future [36]. Thus, people may be in fear of job replacement.

(3) Issues about Robotic Companions. As AI robots become more and more sophisticated, they have begun to be regarded as companions of humans. This raises some ethical issues about the relationship between human and robotic companions [37].

3) Categorization Based on Algorithm, Data, Application, and Long-Term & Indirect Ethical Risks

In the analysis report of AI ethical risks [38] released by the Chinese National AI Standardization General Working Group, AI ethical issues are categorized into four aspects: 1) ethical issues related to AI algorithms, 2) ethical issues related to data, 3) ethical issues related to the application of AI, and 4) long-term and indirect ethical risks.

a) Ethical Issues Related to Algorithms

(1) Algorithm Security. The AI algorithms pose several security issues. First, there is a risk of algorithm or model leakage [39][40]. Generally, the model is achieved by training it on the training data through optimizing its parameters. If the model parameters of an algorithm are leaked, a third party may be able to copy the model. This will cause economic loss to the owner of the model, since a third party obtains the same model without paying the cost of obtaining the training data. Second, the parameters of the AI algorithm model may be modified illegally by an attacker, which will cause the performance deterioration of the AI model and may lead to undesirable consequences. Additionally, in many scenarios, the output of the model is closely related to personal safety, such as in the medical and autonomous driving fields. Once there are loopholes or mistakes in the application of algorithms in these fields, it will directly harm humans and cause serious consequences [41].

(2) Algorithm Explainability. Due to the black-box characteristic of many ML algorithms [33], especially the popular deep learning or neural networks, the decision process of AI algorithms is hard to understand. The interpretability or explainability of algorithms is an essential ethical issue of AI [42], since it concerns the human right to know.

(3) Algorithmic Decision Dilemma. After obtaining the AI model, the result of the algorithm is usually unpredictable for us. In other words, even though we have designed an AI model well, we cannot foresee or predict the decisions of the algorithm and the consequence it will produce. This leads to the algorithmic decision risk or dilemma of AI. For instance, autonomous vehicles should reduce traffic accidents, but sometimes they have to choose between two evils, such as crushing pedestrians or sacrificing themselves and passengers to save pedestrians [43].

b) Ethical Issues Related to Data

(1) Privacy Protection. With the development of big data and AI, the tension between AI technology and user privacy protection has become more and more serious. Criminals have more ways to obtain personal privacy data with lower costs and greater benefits. Data security incidents have commonly occurred in recent years. Privacy protection has become a well-recognized and serious ethical issue involved by using AI [44].

(2) Recognizing and Processing Personal and Sensitive Information. Traditional laws and regulations only focus on the protection of personal and sensitive information. If the personal or sensitive information is de-identified [45] through randomization, data synthesis, and other technologies, it will no longer be regarded as personal or sensitive information and not protected by traditional laws. The subsequent usage, sharing, and transfer of such information arise some ethical issues.

c) Ethical Issues Related to Application

(1) Algorithm Discrimination. The execution results of algorithms directly affect the decision-making of AI systems. However, algorithm discrimination or bias has been seen in many applications of AI. For instance, the racial bias in criminal justice systems [46], and gender discrimination in hiring [47].

(2) Algorithm Abuse. Algorithm abuse [48] refers to the situation where people use algorithms for analysis, decision-making, coordination, and other activities, but their use purpose, use method, use range, etc., have deviations and cause adverse effects. For example, facial recognition algorithms can be used to improve the level of public security and speed up the discovery of criminal suspects, but if they are applied to detect potential criminals, or to determine whether someone has criminal potential based on their face, it is an algorithm abuse.

d) Long-Term and Indirect Ethical Risks

(1) Employment. With the fast advancement and widespread application of AI, more and more work can be completed by some AI products [27]. This will have a significant influence on the employment problem.

(2) Ownership. As AI continues to improve, the intellectual differences between AI agents and humans will gradually shrink. A series of debates on ownership will follow, such as whether the AI agent should be considered as "legal subject", whether AI products have property rights (copyrights or patent rights) [49], and so forth.

(3) Competition. Unfair competition, malicious competition, and monopolistic behaviors with technological advantages will all have an impact on social stability and market freedom, fairness, and equal value, and will seriously damage the interests of consumers and hinder the improvement of social welfare [38]. When companies, organizations or individuals use AI algorithms, they should follow competitive ethics and not go beyond legal boundaries.

(4) Responsibility. With the widespread application of AI, many cases in which AI products violate the laws or ethics, such as personal injury and algorithmic bias, have been observed. An fundamental problem that arises in these cases is who is responsible for these bad consequences [50]. For example, as autonomous driving involves multiple subjects, such as car owners, drivers, passengers, car manufacturers, autonomous driving system providers, pedestrians and etc., how should they bear responsibilities after a traffic accident.

4) Categorization Based on the Deployment of AI

In European Parliamentary Research Service's latest study on the ethical implications and moral questions brought by AI [51], the ethical issues are mapped into different categories according to the ethical impacts of AI on human society, human psychology, financial system, legal system, environment and the planet, and trust.

a) Impact on Society

(1) The Labor Market. AI has already been applied in finance, advanced manufacturing, transportation, energy development, healthcare, and many other sectors. We have already seen the impact of automation on "blue collar" jobs. As AI agents or robots become more and more sophisticated, creative, versatile, and intelligent, more jobs will be affected by

AI technologies and more positions will be obsolete. Therefore, AI technologies may put current job classes at risk, eliminate positions, cause mass unemployment in many job sectors [36]. Furthermore, discrimination in the labor market may also be an issue, for instance, people without high-skill training will be disproportionately affected by the application of AI.

(2) Inequality. AI technologies are expected to enable companies to streamline their business operations and make them more efficient and productive. However, some people argue that this will come at the expense of their human workforces. Thus, this will inevitably indicate that revenues will be split across fewer people and individuals with ownership in AI-driven companies will receive disproportionate benefits, which indeed increase social inequalities [52].

(3) Privacy, Human Rights and Dignity. AI is already affecting privacy, human rights, and dignity in many ways. For example, the Intelligent Personal Assistants (IPA), such as Apple's Siri, Amazon's Echo, and Google's Home, can learn the interests and behavior of their users, but, at the same time, the users raise concerns about the fact that they are always running and listening in the background [53]. The IPA obviously affects our privacy. AI has an important impact on democracy and people's right to private life and dignity. For instance, if AI can be used to determine people's political beliefs, then individuals may be vulnerable to manipulation. Political strategists can use this information to determine which voters are likely to be persuaded to change party affiliation and then use resources to persuade them to do so.

(4) Bias. Human bias, such as gender prejudice and racism bias, may be inherited by AI. The bias of AI may arise as a result of the training data, the value held by the developers and users, or acquired from the learning process of AI itself. Many cases of AI bias, machine bias or algorithmic bias have been reported [54]. The bias of AI will promote unexpected social bias or discrimination. Thus, bias is an ethical issue that is often talked about by the public.

(5) Democracy. The implementation and adoption of AI can threaten democracy in several ways. First, the concentration of technological, economic, and political power related to AI among a few mega corporations could allow them to pose undue influence over the government. Second, AI may damage democracy by affecting political elections [55]. With the aid of AI and big data, politicians have access to huge amounts of information that allow them to target specific voters and develop messages that will resonate with them most. Third, the increasing use of AI-based new recommenders, which present readers with news stories based on their previous reading history, reduces readers' chances of encountering different and undiscovered content, options, and viewpoints [56]. This could result in increasing societal polarization.

b) Impact on Human Psychology

(1) Relationships. AI is getting better and better at imitating human thought, experience, action, dialogue, and relationships. In the future, we will frequently interact with machines or AI products as if they are humans. This will have impacts on real human relationships and thus bring some ethical issues [57].

(2) Personhood. AI systems are increasingly taking on tasks and decisions that are traditionally performed by humans. An essential and ethical question that arise from this is that whether AI system should be endowed with "personhood" and moral or legal agency rights [58].

c) Impact on the Financial System

The application of AI in financial markets has significantly improved transaction efficiency and trading volume. Markets are very suitable for automation, because they now operate almost entirely electronically and a huge amount of data is generated at a high rate, which requires the employment of algorithms to digest and analyze it. Additionally, due to the dynamic of markets, fast reaction to information is critical [59], which provides considerable incentives to replace slow people's decision process with algorithmic decision-making. Furthermore, the rewards for effective trading decisions are considerable, which explains why companies have invested so much in AI technology.

However, the AI-based automatic trading agents may also be used maliciously to destabilize the markets or harm innocent parties in other ways. Even if they are not intended to be malicious, the autonomy and flexibility of algorithmic trading strategies, including the increasing use of ML techniques, make it difficult for people to predict how they will perform in unexpected situations.

d) Impact on the Legal System

(1) Criminal Law. According to current criminal law, a crime consists of two elements, that is, a voluntary act (or omission) and an intention to commit a crime. If AI products or robots are shown to have sufficient consciousness or awareness, then they may be the direct perpetrators of criminal offenses or responsible for negligent crimes. If we admit that AI products have their own mind, human-like free will, autonomy, or moral sense, then our criminal law and even the entire legal system will have to be revised [60].

(2) Tort Law. Tort law covers situations where one person's behavior case injury, suffering, unfair loss, or harm to another person. When an accident involving self-driving car(s) occurs, there are two legal areas that are relevant - negligence and product liability. While, today, most accidents result from driver error, which indicates that liability for accidents are governed by the negligence principle. So, in the future, the tort law, which includes many different types of personal injury claims, will be significantly affected [61] since AI products (such as self-driving cars or other intelligent robots) will involve in personal injury claims, such as the accident between self-driving cars or the injury claim where a robot harm human.

e) Impact on the Environment and the Planet

(1) Use of Natural Resources. The development and application of AI will increase the demand of many natural resources, such as rare earth metals like nickel, cobalt, graphite, and so on. As the existing supply decreases, operators may be forced to work in new and more complex environments to mine. This will increase the production and consumption rate of rare earth metals, and further damage the environment [62].

(2) Pollution and Waste. The increase in production and consumption of AI technological devices such as robots will exacerbate pollution and waste, such as the accumulation of heavy metals and toxic materials in the environment [63].

(3) Energy Concerns. Employing AI technology, particularly deep learning, generally involves training ML

models on a huge amount of data, which usually consumes large amounts of energy. According to listed data in [64], the carbon footprint of training a natural language processing model (a Transformer model) is roughly 5 times the carbon footprint of an average car across its entire lifetime.

f) Impact on Trust

AI promises numerous changes and benefits to individual's lives and the society. It is changing our daily lives in many domains, such as transportation, service industry, healthcare, education, public safety and security, and entertainment. Nevertheless, these AI systems must be introduced in ways that foster trust and understanding and respect human and civil rights [65]. The consensus among the research community is that trust in AI can only be achieved through fairness, transparency, accountability, and regulation (or control).

(1) Fairness. In order to trust AI, it must be fair and impartial. As more and more decisions are delegated to AI, we must ensure that these decisions are free from bias and discrimination [66]. Whether it is filtering through CVs for job interviews, deciding on admissions to the university, or conducting credit ratings for loan companies, it is essentially vital that decisions made by AI are fair.

(2) Transparency. Transparency is important for building trust in AI since it should be a must to know why an AI system made a particular decision, especially if that decision caused undesirable consequences or harm. In view of the fact that the autopilot of an intelligent car has led to several fatal accidents, it is clear that transparency is urgently needed to discover how and why these accidents occur, and to correct any technical or operational failures. The opacity in ML, which is well-known as black-box, is one of the main impediments to the transparency of AI [51].

(3) Accountability. Accountability [67] ensures that if an AI system makes a mistake or hurts someone, then someone can be held responsible, whether it is the designer, developer, or company selling the AI. In the event of damages, accountability is essential to establish a remedial mechanism so that victims can receive adequate compensation. Thus, accountability is crucial to ensure the trust of AI.

(4) Control. Another issue that affects the public trust in AI is the controllability of AI [68]. This is largely related to people's fear about the idea of "super-intelligence", that is, as the intelligence of AI increases to the point that it surpasses human abilities, AI may come to take control over our resources and outcompete our species, and even leading to human extinction. A related concern is that even if an AI agent is carefully designed to align its goals with human needs, it may develop unpredictable sub-goals on its own. Therefore, in order to maintain trust in AI, it is important that humans must have ultimate oversight or control on AI technology.

TABLE I List and discussion of the reviewed categorization of ethical issues of AI and our proposed categorization.

Categorization	Class	Ethical Issues	Discussion
Classification of AI ethical Issues from features of AI, human factors, and social impact [11]	Ethical Issues Caused by Features of AI	Transparency, Data Security and Privacy, Autonomy, Intentionality, and Responsibility	The impacts of AI on the environment, such as natural resource consumption and environmental pollution, are ignored.
	Ethical Issues Caused by Human Factors	Accountability, Ethical Standards, Human Rights Laws	
	Social Impact of Ethical AI Issues	Automation and Job Replacement, Accessibility, Democracy and Civil Rights	
Classification of AI ethical issues from the vulnerabilities of AI and human [29]	Ethical Issues from the Vulnerabilities of AI	ML is Data Hungry, Garbage In/Garbage Out, Faulty Algorithms, Deep Learning Is a Black Box	Several important issues, such as responsibility, safety, freedom, and environmental problems, are omitted.
	Ethical Issues from the Vulnerabilities of Human	Abuse Use of AI, Job Replacement, Issues about Robotic Companions	
Classification of AI ethical issues by ethical issues related to algorithm, data, application, and long-term & indirect ethical risks [38]	Ethical Issues Related to Algorithm	Algorithm Security, Algorithm Explainability, Algorithmic Decision Dilemma	Issues involved accountability, fairness, autonomy and freedom, human dignity, environmental problems are not included.
	Ethical Issues Related to Data	Privacy Protection, Recognizing and Proceeding Personal Sensitive Information	
	Ethical Issues Related to Application	Algorithm Discrimination, Algorithm Abuse	
	Long-Term and Indirect Ethical Risks	Employment, Ownership, Competition, Responsibility	
Classification of AI ethical Issues based on the deployment of AI [51]	Impact on Society	The Labor Market, Inequality, Privacy, Human Rights and Dignity, Bias, Democracy	Some issues, including responsibility, safety, and sustainability, are omitted and this classification is complicated and cumbersome to understand.
	Impact on Human Psychology	Relationships, Personhood	
	Impact on the Financial System		
	Impact on the Legal System	Criminal Law, Tort Law	
	Impact on the Environment and the Planet	Use of Natural Resources, Pollution and Waste, Energy Concerns	
	Impact on Trust	Fairness, Transparency, Accountability, Control	
Our categorization: Classification of AI ethical Issues at individual, societal, and environmental levels	Ethical Issues at Individual Level	Safety, Privacy & Data Protection, Freedom & Autonomy, Human Dignity	Our categorization classifies AI ethical issues from individual, societal and environmental levels. This classification is not only clear and easy-to-understand but also comprehensively covers the discussed ethical issues.
	Ethical Issues at Societal Level	Fairness & Justice, Responsibility & Accountability, Transparency, Surveillance & Datafication, Controllability of AI, Democracy and Civil Rights, Job Replacement, Human Relationship	
	Ethical Issues at Environmental Level	Natural Resources, Energy, Environmental Pollution, Sustainability	

B. Our Proposed Categorization: Ethical Issues at Individual, Societal and Environmental Levels

In the previous subsection, we have reviewed the AI ethical issues described and categorized in the literature (see Table I). However, the above presented categorizations have obvious flaws. Specifically, the categorization based on features of AI, human factors and social impact [11] obviously ignores the impact of AI on the environment, such as natural resource consumption and environmental pollution. The categorization based on vulnerabilities of AI and human [29] omits several important issues, such as responsibility, safety, and environmental problems. The categorization based on algorithm, data, application, and long-term & indirect ethical risks [38]misses the considerations of fairness, autonomy and freedom, human dignity, environmental problems and etc. Although the categorization based on the deployment of AI [51] covers ethical issues comprehensively, this classification is too cumbersome and some issues, including responsibility, safety, and sustainability, are omitted. This motivates us to further analyze and sort out AI ethical issues.

It is no doubt that AI systems mainly serve individuals or the public of society. Hence, we can analyze and clarify AI ethical issues from individual and societal perspectives. At the same time, as entities on the planet, AI products will inevitably have impacts on the environment. So, the ethical issues related to the environmental aspects also need to be considered. Therefore, in this subsection, we proposed to classify AI ethical issues at three different levels, that is, ethical issues at individual, societal, and environmental levels. Ethical issues at individual level mainly include issues that have undesirable consequence for individual human beings, their rights, and their well-being [69]. AI ethical issues at societal level consider the societal consequence that AI has brought or may bring for groups or society as a whole [69]. AI ethical issues at environmental level focus on the impacts of AI on the natural environment. Our proposed categorization is shown in Fig. 2.

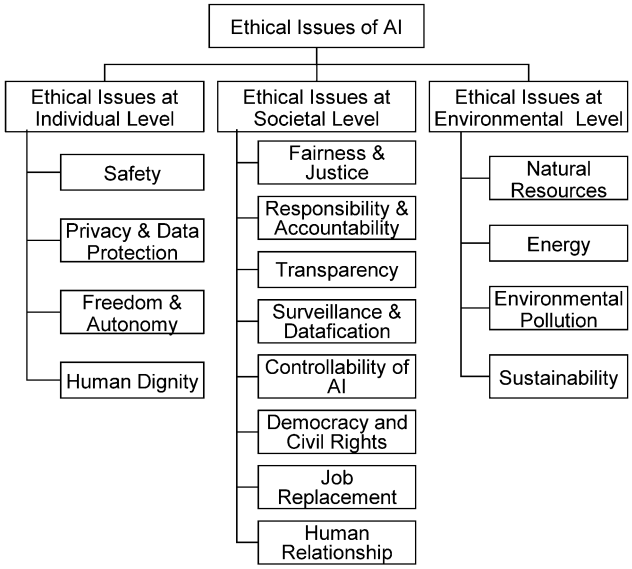

Fig. 2. The proposed categorization of AI ethical issues.

1) Ethical Issues at Individual Level

At individual level, AI has brought influence on the safety, privacy, autonomy, and human dignity of individuals. The application of AI has posed some risks on the safety of individuals. For instance, person injury accidents involving autonomous cars and robots have occurred and reported in the past few years. Privacy issue is one of the serious risks that AI brings to us. To achieve good performance, AI systems usually require a huge amount of data, which often include users' private data. However, there are serious risks associated with this data collection. One of the main issues is privacy and data protection. Additionally, as described in the previous subsection, the application of AI may bring challenges to human rights, such as autonomy, and dignity. Autonomy refers to the capacity of thinking, deciding, and acting independently, freely and without influence of others [70]. When AI-based decision-making are widely adopted in our daily life, three is big danger of restricting the autonomy of us. Human dignity, which is one of the principal human rights, is about the right of a person to be respected and treated in an ethical manner [71]. The protection of dignity is crucial in the context of AI. Human dignity should be one of the basic concepts for protecting human beings from harm and should be respected when developing AI technologies. For instance, a lethal autonomous weapon system [72] may violate the principle of human dignity.

2) Ethical Issues at Societal Level

When considering the AI ethical issues at societal level, we mainly focus on the broad consequences and impacts that AI brings for society and the well-being of communities and nations around the world. Under the categorization of ethical issues at societal level, we discuss fairness and justice, responsibility and accountability, transparency, surveillance and datafication, controllability of AI, democracy and civil rights, job replacement and human relationship.

The existence of bias and discrimination in AI has posed challenges on fairness and justice. The biases and discrimination embedded in AI might increase societal gaps and cause harm to certain societal groups [70]. For instance, in the US criminal justice system, AI algorithms that are used to assess the risk of committing crime has been noticed to exhibit racial bias [73]. Responsibility means being responsible for or in charge of something. Assigning responsibilities to participants is important for shaping the governance of algorithmic decision-making. Based on this concept, accountability is the principle that the one who is legally or politically responsible for the damage must provide some form of justification or compensation and is reflected by the liability to provide legal remedies [70]. Thus, mechanisms should be established to ensure responsibility and accountability of AI systems and their outcomes both before and after their implementations. Due to the black-box nature of AI algorithms, lack of transparency has become one of the widely discussed issues. Transparency, i.e., the understanding of how AI systems work, is crucial for accountability as well. Surveillance and datafication [74] is one of the common concerns as we live in the so-called digital and intelligent age. Data is collected from users' daily lives via smart devices, and we live in mass surveillance. As the power of AI has increased quickly, the

development of AI systems must have safeguards to ensure the controllability of AI systems by humans. Other previously discussed issues, including democracy and civil rights, job replacement, and human relationship, also fall into this category.

3) Ethical Issues at Environmental Level

AI ethical issues at environmental level focus on the impacts of AI on the environment and the planet. AI can bring a lot of convenience to our lives and can help us to address some challenges, but it also comes at a cost to the planet. The widespread application of AI often requires the deployment of a large number of hardware terminal devices, including chips, sensors, storage devices, and etc. The production of these hardware consumes a lot of natural resources, especially some rare elements. In addition, at the end of these hardware's life cycle, they are usually discarded, which will cause serious environmental pollution. Another significant aspect is that AI systems usually require considerable computing power, which comes with high energy consumption. Furthermore, from a long-term and global view, the development of AI should be sustainable, i.e., AI technology must meet the human development goals while simultaneously sustain the ability of natural systems to provide the natural resources and ecosystem services on which the economy and society depend [2]. In summary, natural resource consumption, environmental pollution, energy consumption costs, and sustainability involved in the development of AI are the main issues and concerns at environmental level.

Our proposed categorization clarifies ethical issues from three main levels, that is, the impact of AI on individual, society, and the environment. No matter which field or sector AI is used in, we can consider the corresponding ethical issues from these three levels. Obviously, this classification method is simple and clear, and it comprehensively covers AI ethical issues.

C. Key Ethical Issues Associated with each Stage of the AI System's Lifecycle

After reviewing the ethical issues and risks discussed in the literature, we discuss the ethical issues associated with the different stages of an AI system's lifecycle. If we know the existing ethical problems are prone to be caused by or be raised in which stages or steps of the AI system's lifecycle, this will be greatly beneficial for us to eliminate these problems. This is the motivation to discuss the potential ethical issues in each stage of the lifecycle of an AI system.

The general lifecycle or development process of a ML-based AI system [75] or product [76] often involves the following stages: business analysis, data engineering, ML modeling, model deployment, and operation and monitoring. Usually, the lifecycle of AI products starts from the business analysis, which mainly involves identifying and understanding the business problem to be solved and business metrics (or criteria of success). These metrics should include model performance metrics as well as business KPIs (Key Performance Indicators) to be improved by leveraging AI models. The next step is about data engineering that concerns with data collection, data labeling, data cleaning, data structuring, feature engineering, and other operations related to data. After this, the process enters into the so-called ML modeling step. This step generally involves the iterative process of algorithm design or selection, model training, and model evaluation. If the build model is satisfying, then the process goes to the model deployment step, which makes the ML model available to other systems within the organization or the web so that the model can receive data and return their predictions. The operation and monitoring step involves operating the AI system and continuously evaluating its performance and impacts. This step identifies problems and adjusts or evolves the AI system by reverting to other steps or, if necessary, retiring the AI system from production.

We attempt to establish a map that links ethical issues with the stages of AI lifecycle, where the connection means that the ethical issue is more likely to occur in a certain step of AI lifecycle, or it is often caused by some reason in this step. This mapping is presented in Table II, where several vital ethical problems are associated with the five steps of AI lifecycle. This mapping will be useful for addressing the ethical problem in a proactive fashion during the design process of an AI system.

TABLE II Ethical considerations along each stage of the AI lifecycle.

Stage of AI Lifecycle	Ethical Considerations Exist Along the Stage
Business Analysis	Transparency, Fairness (Does the architecture of the designed AI product includes any variables, features, processes that are unreasonable, morally objectionable, or unjustifiable ? [77]), Responsibility & Accountability, Democracy & Civil Rights, Sustainability
Data Engineering	Privacy (How to assure the data security and keep the private and sensitive information included in data set ?), Transparency (How to make data collection procedures transparent to consumers?), Fairness (Are data properly representative, relevant, accurate, and generalizable ?), Democracy & Civil Rights (How will you enable end users to control use of their data ?)
ML Modeling	Transparency (Does the decision or inference process of the model can be understood ?), Safety (Accuracy, Reliability, Security, and Robustness of the model), Fairness (Are the model outputs show disparate results on different groups of people ?)
Model Deployment	Privacy (Make sure that private information cannot be re-identified through the deployed model), Safety (How to ensure the safety of the deployed model, such malicious modification and attack ?)
Operation and Monitoring	Privacy (Privacy should be guaranteed during the operation & monitoring process), Fairness (Does the AI product has discriminatory or inequitable impacts on peoples they affect ?), Democracy & Civil Rights (Do not infringe civil rights or the users)

IV. Ethical Guidelines and Principles for AI

As the ethical issues of AI have received more and more attention and discussions from various sectors of society, many organizations (including academia, industry, and government) have begun to discuss and seek the possible frameworks, guidelines and principles for solving AI ethics issues [78]. These guidelines and principles provide useful directions for practicing ethical AI. This section is dedicated to giving an up-to-date global landscape of the AI ethics guidelines and principles, which is achieved through the investigation of 146 reports, guidelines and recommendations related to AI ethics released by companies, organizations, and governments around the world since 2015. These guidelines and principles provide high-level guidance for the planning, development, production and usage of AI and directions for addressing AI ethical issues.

A. Guidelines for AI Ethics

An excellent survey and analysis of the current principles and guidelines on ethical AI has been given in 2019 by Jobin *et al.* [12], who conducted a review of 84 ethical guidelines released by national or international organizations from various countries. Jobin *et al.* [12] found strong widespread agreement on five key principles, that is, transparency, justice and fairness, non-maleficence, responsibility, and privacy, among many. However, many new guidelines and recommendations for AI ethics have been released in the past two years, making Jobin's paper obsolete because many important documents were not included. For instance, on 24 November 2021, UNESCO (the United Nations Educational, Scientific and Cultural Organization) adopted the Recommendation on the Ethics of Artificial Intelligence, which is the first ever global agreement on the ethics of AI [79]. To update and enrich the investigation on ethical AI guidelines and principles, based on the table of ethics guidelines for AI given in Jobin's paper [12] (only included 84 documents), we have collected many newly released AI ethical guidelines that are not included in Jobin's review. Finally, a total of 146 AI ethics guidelines have been collected. A list of all the collected guidelines or documents is given in Table V of the supplementary materials. The number of guidelines issued each year from 2015 to 2021 is counted and listed in Table III. It is apparent that the majority of the guidelines are released in the last five years, i.e., from 2016 to 2020. The number of guides published in 2018 was the largest, with 53, accounting for 36.3% of the total number. Additionally, the number of AI guidelines issued by each country is listed in Table IV. Furthermore, the percentages of guidelines released by different types of issuers (including government, industry, academia, and other organizations) are shown in Fig. 3. It can be seen from Fig. 3 that governments, companies, and academia all have shown strong concerns about AI ethics.

TABLE III The number of documents issued each year from 2015 to 2021.

Year	2015	2016	2017	2018	2019	2020	2021
Number of Documents	2	7	25	53	31	24	4

TABLE IV The number of guidelines issued by each country or region.

Country	Australia	Canada	China	Denmark	EU	Finland	France
Number.	3	4	5	4	15	4	3
Country	Germany	Iceland	India	International	Ireland	Japan	N/A
Number	7	1	1	12	3	6	3
Country	Netherlands	Norway	Russia	Singapore	South Korea	Spain	Sweden
Number	4	1	1	3	3	2	1
Country	Switzerland	Turkey	UAE	UK	USA	Vatican	
Number	1	1	2	16	39	1	

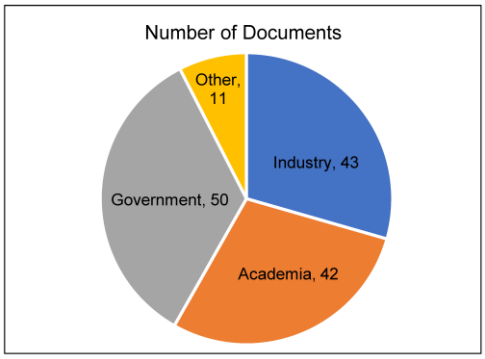

Fig. 3. Percentage of guidelines released by different types of issuers.

B. Principles for AI Ethics

The ethical principles that are featured in the collected 146 guidelines are listed in Table I of the supplementary materials. According to the table, there is an obvious convergence emerging around five important ethical principles: transparency, fairness and justice, responsibility, non-maleficence, and privacy. The 11 ethical principles identified in the existing AI guidelines are described and explained in the following.

(1) Transparency. Transparency is one of the most widely discussed principles in the AI ethics debate. The transparency of AI mainly involves the transparency of the AI technology itself, and the transparency of the developing and adopting of the AI [13]. On the one hand, transparency of AI involves the interpretability of a given AI system, that is, the ability to know how and why a model performed the way it did in a specific context and thus to understand the rationale behind its decision or behavior. This aspect of transparency is usually mentioned as the metaphor of "opening the black box of AI". It concerns interpretability, explainability, or understandability. On the other hand, transparency of AI includes the justifiability or rationality of the design and implementation process of the AI system and that of its outcome. In other words, the design and implementation process of the AI system and its decision or behavior must be justifiable and visible.

(2) Fairness & Justice. The principle of justice and fairness states that the development, deployment and use of AI must be just and fair so that the AI system should not result in discriminations or bias against individuals, communities, or groups [80]. Discrimination and unfair outcomes brought by AI algorithms have become a hot topic in the media and academia. Consequently, fairness and justice principle has attracted considerable attention during the last few years.

(3) Responsibility and Accountability. The principle of responsibility and accountability requires that AI must be auditable, that is, the designers, developers, owners, and operators of AI are responsible and accountable for an AI system's behaviors or decisions, and are therefore considered responsible for harms or bad outcomes it might cause [51]. The designers, builders, and users of AI systems are stakeholders in the moral or ethical implications of their use, misuse, and behavior, and they have the responsibility and opportunity to shape these implications. This requires that appropriate mechanisms should be established to ensure responsibility and accountability for AI systems and their results, both before and after their development, deployment, and use.

(4) Non-maleficence. The non-maleficence basically means to do no harm or avoid imposing risks of harm to others [81][82]. Thus, the non-maleficence principle of AI generally refers to that AI systems should not cause or exacerbate harm to humans or adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. The non-maleficence principle requires that AI systems and the environments in which they operate must be safe and secure so that they are not open to malicious use. With some of the fatal accidents coming from autonomous cars and robots, avoiding harm to human beings is one of the greatest concerns in AI ethics. Hence, most of the ethical guidelines put a strong emphasis on ensuring no harm to human beings through the safety and security of AI.

(5) Privacy. The privacy principle aims to ensure respect for privacy and data protection when using AI systems. AI systems should preserve and respect privacy rights and data protection as well as maintain data security. This involves providing effective data governance and management for all data used and generated by the AI system throughout its entire lifecycle [83]. Specifically, data collection, usage and storage must comply with laws and regulations related to privacy and data protection. Data and algorithms must be protected against theft. Once information leakage occurs, employers or AI providers need to inform employees, customers, partners, and other relevant individuals as soon as possible to minimize the loss or impact caused by the leakage.

(6) Beneficence. The principle of beneficence states that AI shall do people good and benefit humanity [82]. This principle indicates that AI technology should be used to bring beneficial outcome and impact to individuals, society, and the environment [84]. When developing an AI system, its objectives should be clearly defined and justified. The use of AI technology to help address global concerns should be encouraged, such as using AI to help us to handle food security, pollution, and contagion like AIDS and COVID 19.

(7) Freedom and Autonomy. Freedom and autonomy, which generally refers to the ability of a person to make decisions respect to his goals and wishes, is the core value for citizens in democratic societies. Therefore, it is important that the use of AI does not harm or encumber the freedom and autonomy for us. When we apply AI agents, we are willing to give up part of our decision-making authority to AI machines. Thus, upholding the principle of freedom and autonomy in the context of AI means to strike a balance between the decision-making power we maintain for ourselves and that which we cede to AI [84].

(8) Solidarity. The solidarity principle entails that the development and application of an AI system must be compatible with maintaining the bounds of solidarity among people and generations. In other words, AI should promote social security and cohesion, and should not jeopardize social bonds and relationships [13].

(9) Sustainability. Due to climate change and ongoing environmental damage, the importance of sustainability has received more and more attention. Like other fields and disciplines, AI is affected and needs to be included in the sustainable development agenda. The sustainability principle represents that the production, management, and implementation of AI must be sustainable and avoid environmental harm. In other words, AI technology must meet the requirements of ensuring the continued prosperity of mankind and preserving a good environment for future generations [85]. AI systems promise to help tackling some of the most pressing societal concerns, but it must be ensured that this happens in the most environmentally friendly way possible.

(10) Trust. Trustworthiness is a prerequisite for people and societies to adopt AI, since trust is a basic principle for interpersonal interactions and social operation. The trust in the development, deployment and use of AI systems is not only related to the inherent characteristics of the technology, but also related to the quality of the socio-technical system involving AI applications. Therefore, moving towards trustworthy AI not only concerns the trustworthiness of the AI system itself, but also requires a holistic and systematic approach that covers the trustworthiness of all participants and processes that are the entire life cycle of the system [86].

(11) Dignity. Human dignity encompasses the belief that all people possess an intrinsic value that is tied solely to their humanity, i.e., it has nothing to do with their class, race, gender, religion, abilities, or any other factor other than them being human, and this intrinsic value should never be diminished, compromised, or repressed by other people nor by technologies like AI. It is important that AI should not infringe or harm the dignity of end-users or other members of society. As a result, respecting human dignity is an important principle that should be considered in AI ethics. AI system should hence be developed in a way that respects, supports, and protects people's physical and mental integrity, personal and cultural sense of identity, and satisfaction of their basic needs [13].

V. Approaches to Address Ethical Issues in AI

This section reviews the approaches to address or mitigate ethical issues of AI. As AI ethics is a broad and multi-disciplinary field, we attempt to provide a comprehensive overview of the existing and potential approaches for addressing AI ethical issues, including ethical, technological, and legal approaches, rather than solely focusing on technological approaches that are of interest to the field of AI/ML community. This review of multi-disciplinary approaches for addressing AI ethical problems not only provides an informative summary about the approaches to ethical AI but also suggests the researchers in AI community to seek solutions to AI ethical issues from a variety of perspective rather than relying solely on technological approaches. As AI ethical issues are complex and multi-disciplinary problems, it may be possible to solve these problems effectively only through the cooperation of different methods.

Ethical approaches dedicate to developing ethical AI systems or agents, which are able to reason and act ethically according to ethical theories [87], by implementing or embedding ethics in AI. Technological approaches are designed to develop new technologies (especially ML technologies) to eliminate or mitigate the shortcomings of current AI. For instance, research on explainable ML intends to develop new approaches to explain the reason and work mechanism of ML algorithms. Fair ML studies techniques that enable ML to make fair decisions or predictions, that is, to reduce the bias or discrimination of ML. Legal approaches intend to regulate or govern the research, deployment, application, and other aspects of AI through legislation and regulation, with the goal of avoiding previously discussed ethical issues.

A. Ethical Approaches: Implementing Ethics in AI

Designing ethical AI systems, which can reason and act ethically, demands the understanding of what ethical behavior is. This involves judgments of right and wrong, good and bad, as well as matters of justice, fairness, virtue and other ethical principles. Thus, ethical theories, which are concerned with concepts of right and wrong behavior, are closely related to AI ethics. This subsection is dedicated to approaches for implementing ethics into AI systems based on the existing ethical theories. Firstly, ethical theories, particularly the normative ethics which are relevant to AI ethics, are reviewed.

Then, three main types of approaches for designing ethical AI systems are summarized.

1) Ethical Theories

The field of ethics (also known as moral philosophy) is concerned with systematizing, defending, and recommending concepts of right and wrong behavior. Ethics focus on judging and determining which action would be good or moral in given circumstances [88]. The philosophical study of ethics usually includes three main subject areas: meta-ethics, normative ethics, and applied ethics [89]. The branches of ethical theories are shown in Fig. 4.

- Meta-ethics investigates the nature, scope, and meaning of ethical principles or moral judgment. It consists in the attempt to understand the meaning and the origin of ethical terms, the role of reason in ethical judgements, and the issues of universal truths or human values [90].
- Normative ethics seeks to arrive at moral standards and rules that regulate right and wrong behavior. That is, it aims to establish a set of rules that govern human behavior or how things should be by examining how humans value things and judge right from wrong or good from bad.
- Applied ethics is the ethics of particular application fields, which consists of the analysis of specific, controversial moral issues, such as abortion, capital punishment, animal rights, environmental concerns, nuclear war and etc.

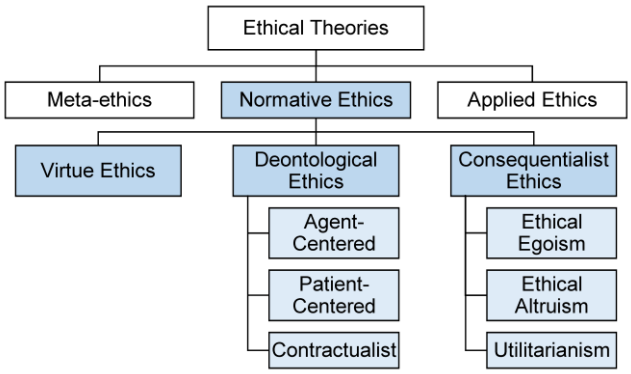

Fig. 4. Branches of ethical theories [91].

a) Normative Ethics

Normative ethics is particularly pertinent to understanding and applying ethical principles to the design, deployment and usage of AI systems [89] since it is a normative practical philosophical discipline that concerned with how humans or agents should act towards others. Three normative ethical branches, that is, virtue, deontological, and consequentialist ethics, are presented and summarized below.

(1) Virtue Ethics. Virtue ethics emphasizes the virtues or moral character and stresses the importance of cultivating good habits of character, such as benevolence [92]. Hence, virtue ethics focuses on the agent's intrinsic character rather than the consequences of actions conducted by the agent. Virtue ethics defines the action of an agent as morally good if the agent acts and thinks according to some moral values [93]. In other words, according to virtue theories, an agent is ethical if it manifests some moral virtues through its actions [94][95].

(2) Deontological Ethics. Deontological theories, which are sometimes called duty theories, judge the morality of an action using certain moral rules that serve as foundational principles of obligation. Deontology is a kind of normative ethics theory regarding which choices or actions are morally required, forbidden, or permitted. In other words, deontology is a moral theory that guides and assesses our decisions about what we ought to do [96]. Deontologists define a morally good action as one that adheres to some obligations, which may be applicable moral rules or duties, regulations, and norms.

There are three main schools of deontological theories, that is, agent-centered, patient-centered (also called victim-centered), and contractarian deontological theories. Agent-centered deontological theories place the agent at the center and focus on agent-relative duties. Patient-centered deontological theories, as distinguished from agent-centered deontology, are rights-based rather than duty-based. It focuses on the rights of patients or potential victims, such as the right of not be used as a means to an end by someone else. Contractualist deontological theories are different from both agent-centered and patient-centered theories. In contractualist deontological theories, morally wrong acts are those acts that would be forbidden by principles that people in a suitably described social contract would accept, or that would be forbidden by principles that such people could not "reasonably reject" [96].

(3) Consequentialist Ethics. Consequentialist ethics, as its name suggests, emphasizes the utilitarian outcomes of actions [97]. Consequentialist ethics assess the morality of an action solely on the basis of its outcome or consequences. In other words, in consequentialist theories, the ethical correctness of an action is determined according to the action's outcome or results. According to consequentialist, an action is morally right if the consequence of that action is viewed as beneficial, i.e., more favorable than unfavorable. Suppose a simple case where one faces with a choice between several possible actions, consequentialism specifies the morally right action is the one with the best overall consequences.

Consequentialist ethics is a historically important and still popular theory because it embodies the basic intuition that what is good or right is whatever makes the world best in the future since we cannot change the past. Consequentialist theories can be divided into [98][99]:

1) Ethical Egoism states that an action is morally good if the consequences or effects of that action are more favorable than unfavorable only to the agent executing the action.
2) Ethical Altruism states that an action is morally good if the consequences or effects of that action are more favorable than unfavorable to everyone except the agent.
3) Utilitarianism states that an action is morally good if the consequences or effects of that action are more favorable than unfavorable to everyone.

All three of these theories focus on the consequences of actions for different groups of people. But, like all normative theories, the above three theories are rivals of each other. They also yield different conclusions.

b) Summary on Normative Ethics

It is clear from the above descriptions that different normative ethical theories will result in different judgement for an action or decision. Consider the following illustration [100]: An elderly gentleman is tormented by a group of arrogant teenagers on the subway and a resolute woman comes to his aid. The

virtue ethicist will deem her action morally appropriate since it instantiates the virtues of benevolence and courage. The deontologist will consider her action commendable as it is in conformity with the rule to help those in need. The consequentialist will defend her action as good, since she maximized the overall well-being of all parties involved-the elderly gentleman is spared suffering and disgrace, which surpasses the teenagers' amusement. A brief comparison between three normative ethical theories is given in Table V.

2) Approaches for Implementing Ethics in AI

In the previous subsection, we have discussed the ethical theories relevant to AI ethics. This subsection briefly reviews the methodologies and approaches to implement ethics in AI systems, i.e., to design ethical AI systems. The existing methodologies or approaches for implanting ethics in AI can be divided into three main types: top-down approaches, bottom-up approaches, and hybrid approaches [101].

a) Top-down Approaches

A top-down approach refers to any approach that adopts a specific ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems that can realize that theory [102]. Top-down approaches conduct ethical reasoning based on given ethical theories or moral principles. In top-down approaches, the moral principles and ethical theories are used as rules to select ethically appropriate actions [101] or are used to describe what the AI agent ought to do in a specific situation. Thus, a top-down approach requires formally defined rules, obligations, and rights to guide the AI agent in its decision-making process. For instance, Asimov's three laws of robotics [103] that governed the behavior of robots can be considered a top-down ethic system for robots [101]. Many other implementations using top-down approaches can be found in [104][105][106][107][108][109][110][111]and so forth.

Top-down approaches are usually understood as having a set of rules that can be transformed into an algorithm. These rules specify the duties of an agent or the need for the agent to evaluate the consequences of the various possible actions it might take. Top-down approaches differ in the ethical theory that is used. For instance, when consequentialist theory is used in top-down approach, the reasoning model needs to evaluate the outcome or consequence of the actions as the basis for the decision, that is, an action that leads to good result is moral and otherwise is unmoral; whereas if deontological theory is applied, the reasoning model will consider the satisfaction of a given value for decision-making, i.e., an action obeying the duties is moral and the one breaking the duties is immoral.

b) Bottom-up Approaches

The bottom-up approaches assume that ethical or moral behavior is learned from observations of the behaviors of others. In bottom-up approach, the emphasis is put on creating an environment in which an AI agent explores the course of action and the morally praiseworthy action is rewarded or selected [101]. Unlike top-down approaches, which require ethical theories or principles to define what is and is not moral, ethical principles is discovered or learned from observations or experience in bottom-up approaches. This approach highlight that AI agent need to learn norms and morality, like little children do, in order to become ethically competent. For instance, Honarvar and Agaee proposed the Casuist BDI-Agent [112] which combine CBR (case-based reasoning) method in AI and bottom-up casuist approach in ethics to add the capability of ethical reasoning to belief-desire-intention (BDI)-Agent [113]. Other implementations of bottom-up approaches can be found in [114][115][116][117][118]and etc.

Bottom-up approaches can harness the wisdom of the crowd as a means to inform the ethical judgment of the agent and then the agent can learn how to judge the morality of its action and thus behave ethically. Apparently, bottom-up approaches assumes that a sufficiently large amount of data or observations about ethical decisions and their outcomes can be collected from a suitable set of subjects or scenarios. This is the requirement for using bottom-up approaches to implement ethical AI systems. However, in practice, this requirement is not easily satisfied.

c) Hybrid Approaches

The hybrid approach attempts to combine the advantages of top-down and bottom-up approaches. The top-down approaches make use of the ethical theories and principles and emphasize the importance of explicit ethical concerns that arise from outside of the entity (the moral subject). While the bottom-up approaches focus more on the cultivation of morality that arise from within the entity through evolution and learning. Both the top-down and bottom-up approaches embody different aspects of the moral sensibility. By combining these approaches, we may be able to create AI agent that can maintain the dynamic and flexible morality of bottom-up approach while obeying the top-down principles. Different hybrid approaches have been implemented in [119][120][121][122][123][124].

As Gigerenzer [125] stated the nature of moral behavior results from the interplay between mind and environment. According to this view, both nature and nurture are important in shaping moral behavior. The hybrid approach is consistent with this concept. In hybrid approach, the top-down approach uses programmed rules and the bottom-up approach learned rules from context observations or experiences, which are similar to the nature and nurture aspects for morality, respectively. from this perspective, thus, both nature and nurture are considered in hybrid approaches.

d) Remarks on Ethical Approaches

The top-down approach instantiates the specified ethical theories and principles into ethical decision-making or converts given ethical theories and principles into algorithms. The top-down approach is suitable for the design and realization of ethical AI agents with known ethical principles and ethical codes. The advantage of the top-down approach is that, based on preset ethical theories and rules, the decisions and actions of ethical agents are predictable, and the ethical norms or rules implemented through program codes or other means can be understood during ethical decision-making process. Therefore, the credibility of the ethical AI agent created by top-down approach can be better guaranteed, and its decision-making process has strong interpretability and transparency. The disadvantage of the top-down approach is that the ethical agent adopts predetermined ethical theories or ethical rules, when

making decisions in a complex and changeable environment, this method lacks flexibility and adaptability.

The bottom-up approach emphasizes that ethical agents learn morality autonomously from the social environment, gradually possess ethical reasoning and moral abilities, and can adapt to environmental changes. The bottom-down approach is suitable for the design and implementation of ethical AI agents without clear ethical theories and guidelines. The advantage of the top-down approach is that the agent can develop and evolve through continuously learning, so as to adapt to environmental changes. This category of approaches has good adaptability and flexibility, and it is possible to construct different and new ethical theories or guidelines for various application scenarios. The disadvantage of the top-down approach is that due to the lack of guidance of ethical theories or rules, the decision-making process of ethical AI agents has a certain degree of blind obedience, and it is difficult to complete the training in a short time and make appropriate ethical decisions. At the same time, it is difficult to guarantee the interpretability and transparency of the decision-making process of the designed ethical AI agents.

The hybrid approach combines the advantages of top-down and bottom-up approaches and overcomes the shortcomings of the two methods to a certain extent. If a single approach (top-down or bottom-up) does not cover the requirements, a hybrid approach is considered necessary and promising. However, the main challenge is to properly combine the features of top-down and bottom-up approaches. The features of the three approaches for implementing ethics in AI are summarized and listed in Table VI.

TABLE V Comparison of the three normative ethical theories [126].

Ethical Theory	Description	Deliberation Focus	Decision Criteria	Practical Reasoning
Virtue Ethics	An action is right if it is what a virtuous person would do in the situation.	Motives (Is action motivated by virtue?)	Virtues	Instantiation of virtues / human qualities
Deontological Ethics	An action is right if it is in accordance with a moral rule or principle.	Action (Is action compatible with some imperative?)	Duties/rules	Follow the rules
Consequentialist Ethics	An action is right if it promotes the best consequences, i.e., maximizes happiness.	Consequences (What is outcome of action?)	Comparative well-being	Maximization of utility or happiness

TABLE VI The features of the three approaches for implementing ethics in AI [127].

Approach	Description	Features			
		Require ethical rules or not ?	Learning Ability	Adaptation Ability	Interpretability
Top-Down	Program the given ethical theory and principles	Yes	No	Weak	High
Bottom-Up	Learn the general rules from individual cases	No	Strong	Strong	Low
Hybrid	Combine bottom-up and top-down approaches	Yes	Strong	Strong	Median

B. Technological Approaches

In this subsection, we briefly summarize the research status about technological approaches to address ethical issues of AI in line with the principles discussed in Section IV.B. Currently, the technological approaches to mitigate the associate issues are still at infant development stage. In recent years, AI research communities have put certain efforts for addressing the issues of AI ethics. For instance, ACM (the Association for Computing Machinery) has held the annual ACM FAccT conference (which brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems) since 2018, AAAI (the Association for the Advancement of Artificial Intelligence) and ACM have established the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) since 2018, and the 31st International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2022) provides a special track on “AI for good”.

The existing work, to the best of our knowledge, mainly focuses on a few major and key issues and principles, and the other issues and principles are rarely involved. Thus, we only give a brief summary on technological approaches that involve the five key ethical principles. Particularly, for five key principles (i.e., transparency, fairness and justice, non-maleficence, responsibility and accountability, and privacy), some representative research topics and relevant references are listed in Table II of the supplementary materials.

Explainable AI (XAI), which is also known as interpretable AI, is currently the main research direction and technical method to address the issues of lack of transparency in AI. The goal of XAI is to allow human users to comprehend the results and output provided by an AI system, especially by ML algorithms. Christoph *et al.* [128] presented a brief history of the field of XAI, given an overview of state-of-the-art interpretation methods, and discussed some research challenges. Additionally, Christoph has written a book about interpretable ML [129], which is a popular publication in XAI field.

As for the fairness principle, there are also many works are dedicated to eliminating or mitigating the bias or discrimination exhibited by AI systems, particularly in ML. Fair AI [130], which aims at preventing disparate harm (or benefit) to different subgroups, is a very active research topic that devote to addressing the issues of the lack of fairness in AI. In the survey of fairness in ML by Simon and Christian [131], different schools of thought and approaches to mitigate biases and increase fairness in ML were reviewed.

Non-maleficence principle includes several codes, such as safety, security, and robustness. Hence, there are some works for each of the codes associated with non-maleficence principle. Currently, safe AI, secure AI, and robust AI are three main research directions to fulfill the non-maleficence principle in AI. Interested readers can get more details through relevant references listed in Table II of the supplementary materials.

As AI is widely used in our lives, responsible AI is becoming critical. Responsibility is a relatively abstract and broad concept. At present, there is no universal and unified definition or notion for responsible AI, which mainly involves accountability, liability, fairness, robustness, and explainability [132]. Dorian *et al.* [133] proposed two frameworks for responsible AI by

integrating ethical analysis into engineering practice in AI. Besides, [134] provides a systematic introduction about responsible AI.

In order to handle the privacy issues in AI, researchers have made many efforts. Differential privacy [135] is one of the main approaches to privacy-preserving ML and data analysis. Recently, a new ML paradigm, that is, Federated learning [136][137] (also called distributed ML), was proposed to mitigate the risk of privacy leakage in ML. In addition, some other privacy-preserving techniques for ML [138][139] have been proposed.

As for the other principles, such as beneficence, freedom and autonomy, dignity, and so forth, we have not found relevant technological approaches in the literature. This may be due to the difficulty or unsuitability of using technical methods to address the issues related to these principles. In general, AI ethics is a relatively new area and approaches for fulfilling these principles still need to be studied in the future.

C. Legal Approaches: Legislation and Regulation

As the increasingly employment of AI technologies in many sectors and the exhibition of ethical issues and risks in applications of AI, many laws and regulations have been established by governments and organizations to govern the development and application of AI. Legal approaches have become one type of the means to address ethical issues in AI. In the following, we list several laws and regulations associated with AI that have been proposed during the past few years.

- In 2016, European Parliament and Council of the European Union (EU) has published the General Data Protection Regulation (GDPR) [140], which is a regulation in EU law on data protection and privacy in European Union and the European Economic Area.
- In 2017, USA passed the bill "Safely Ensuring Lives Future Deployment and Research in Vehicle Evolution Act" [141] for ensuring the safety of highly automated vehicles by encouraging the testing and deployment of such vehicles.
- In 2018, Brazil enacted Law No. 13 709, the General Data Protection Law (Lei Geral de Proteção de Dados) [142], for the protection of personal data in the country.
- In 2021, the European Commission released the Artificial Intelligence (AI) Act [143], which sets out a cross-sectoral regulatory approach to the use of AI systems across the European Union (EU) and its market.

VI. Methods to Evaluate Ethical AI

The goal of the discipline of AI ethics is to design ethical AI systems to behave ethically or adhere to the ethical and moral principles and rules. How to evaluate or assess the ethicality or morality (moral competence) of the designed ethical AI is curial and necessary, because the designed AI systems need to be tested or evaluated whether an AI system meets the ethical requirements or not before deployment. However, this aspect is often ignored or overlooked in the existing literature. This section reviews three types of approaches, testing, verification, and standards, for evaluating the ethics of AI.

A. Testing

Testing is a typical method used to evaluate the ethical capabilities of an AI system. Usually, when testing a system, the output of the system needs to be compared against a ground truth or the expected output [100]. This subsection focuses on testing approaches to evaluate ethical AI.

1) Moral Turing Test

In both ethical theories and daily discussions about ethics, people usually hold different opinions on the morality of various actions. For instance, Kant claimed that lying is always immoral regardless of the consequence. Utilitarian ethicists would deny this and hold that lying is justified as long as its consequences are sufficiently good in the aggregate. Since different ethical theories have different evaluation standards for moral behavior, Allen *et al.* [144] proposed to use the Moral Turing Test (MTT) to evaluate artificial moral agents.

In the standard version of Turing Test [145], a remote human interrogator is charged with distinguishing between a machine (a computer) and a human subject based on their replies to various questions posed by the interrogator. A machine passes the Turning Test if it is misidentified as the human subject with a sufficiently high chance, and the machine is considered as an intelligent and thinking entity. Turning Test directly conducts behavioral test so that it bypasses the disagreement about criteria for defining intelligence or successful acquisition of natural language. The Moral Turning Test (MTT) was similarly proposed to bypass disagreements about ethical standards by restricting the conversations in the standard Turning Test to questions related to morality. If the human interrogator cannot distinguish the machine from the human subject at a level above chance, the machine is a moral agent.

However, Allen *et al.* [144] admitted that one limitation of MTT is that it emphasizes the ability of machines to articulate moral judgments clearly. Deontologists or Kantian might be satisfied with this emphasis, but consequentialists would argue that the MTT places too much emphasis on the ability to articulate the reason for one's actions. In order to shift the focus from conversational ability to action, Allen *et al.* [144] also proposed an alternative MTT that was called the "comparative MTT" (cMTT). In cMTT, the human interrogator is given pairs of descriptions of actual, morally significant actions of a human subject and a machine (or AI agent), purged of all references that would identify the actor. If the interrogator correctly identifies the machine in a certain percentage, then the machine cannot pass the test. A problem of this version of MTT is that the way the machine behaves is easier to recognize than humans, because the machine behaves consistently in the same situation. Therefore, the interrogator should be asked to assess whether one actor is less moral than the other instead of one is more moral than the other. If the machine is not identified as the less moral one of the pair more frequently than the human, the machine has passed the test.

Although cMTT has several problems, for example, someone might argue this standard is too low, Wallach and Allen [146] believe that cMTT is a feasible and acceptable method for evaluating the morality of AI agents, since there are no other evaluation criteria that are commonly accepted and agreed.

2) Expert and Non-expert Tests

Besides MTT, researchers have tried to assess the moral competence of AI systems through expert or non-expert tests, in which the system outcome is compared against the ground truth provided by non-experts or experts. The expert test adopts the standard of experts in normative ethics to assess the morality of AI agents. Non-expert tests take folk morals as the benchmark and evaluate the moral capability of the AI agent or system on the relevant benchmark test. In non-expert tests, citizens can play their roles in assessing and evaluating the ethical capabilities of an AI system based on their own ethical stances and scrutiny.

B. Verification

Another category of approaches for evaluating the morality of AI consists of proving that the AI system behaves correctly according to some known specifications. Seshia *et al.* [147] discussed this kind of approach. A typical formal verification process is shown in Fig. 5, where S is a model of the system to be verified, E is a model of the environment, and Φ is the property to be verified. The verification program will output a Yes/No answer, indicating whether or not S satisfies the property Φ in environment E. Typically, a No output is accompanied by a counterexample, which shows how the execution of the system violates property Φ. And a proof of correctness is included a Yes answer in some formal verification tools.

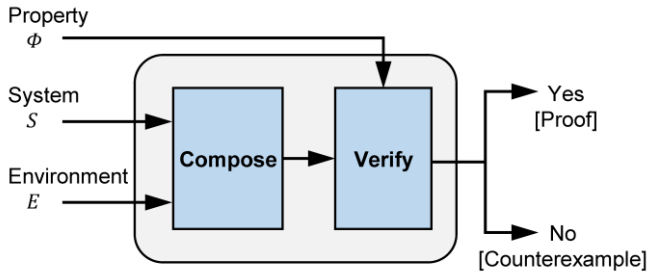

Fig. 5. Formal verification process (this figure is recreated based on [147])

Arnold and Scheutz [148] explored the flaws of MTT and pointed out that MTT-based evaluations are vulnerable to deception, inadequate reasoning, and inferior moral performance, and they proposed the concept of "design verification" to evaluate the moral competence of AI system.

For the evaluation of AI ethical design, diversified evaluation criteria can be used. Regardless of the way AI conducts moral reasoning, it is most critical that its moral activities conform to the goals of ethical design.

C. Standards

Many industry standards have been proposed to guide the development and application of AI and to evaluate or assess AI products. In this subsection, some AI-related standards are introduced.

- In 2014, the Australian Computer Society (ACS) developed the ASC Professional Code of Conduct to follow by all information communication technology (ICT) professionals, which identifies six core ethical values and the associated requirements for professional conduct.
- In 2018, ACM updated the ACM Code of Ethics and Professional Conduct to respond to the changes in the computing profession since 1992. This Code expresses the conscience of the profession and is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. Additionally, the Code serves as a basis for remediation when violations occur. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle [149].
- The project of IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [150] has approved the IEEE P7000™ standards series [151] under development (listed in Table III of the supplementary materials), which cover topics from data collection to privacy, to algorithmic bias and beyond.
- The ISO/IEC JTC 1/SC 42 [152], which is a joint committee between ISO and IEC responsible for standardization in the area of AI, dedicates to developing a large set of standards includes the areas of foundational AI standards, big data, AI trustworthiness, use cases, applications, governance implications of AI, computational approaches of AI, ethical and societal concerns. The standards published and under development by ISO/IEC JTC 1/SC 42 are listed in Table IV of the supplementary materials.

With the concerns about AI ethical issues, the interest in AI standards to shape the design, deployment and evaluation of AI has been growing fast. Although many standards have been proposed, the gap between standards (or principles) and practice is still large. Currently, only some large corporates, such as IBM [153] and Microsoft [154], have implemented their own industrial standards, frameworks, and guidelines to build a culture of AI; but for smaller businesses with less resources, the principles to practice gap is a major problem. Thus, many efforts are still needed. On the one hand, it is necessary to put forward well-developed standards; on the other hand, it is required to vigorously promote the practice of standards.

VII. Challenges and Future Perspectives

As AI ethics is an emerging discipline, and there are still many challenges and problems need to be addressed in this field. In this section, we discuss some challenges in AI ethics and give some future perspective from our views. The purpose of this section is to provide some possible research questions and directions for further research in the future, thereby facilitating the research progress in the field of AI ethics.

A. Challenges in AI Ethical Guidelines and Principles

As reviewed in Section IV, a large number of guidelines have been proposed and released by different organizations, companies and governments, and different principles can be identified in these guidelines. However, at present, there is still no guideline that have been approved and adopted by various organizations, sectors, and governments. In other words, different organizations, companies from different fields, and

even different companies from the same fields have different opinions on AI ethics. The consensus on ethics of AI has not yet been reached and it is not clear what common principles and values AI needs to follow. Additionally, different ethical principles may be required when AI is applied in different areas. Currently, study and discussion on ethics of AI in different specific application areas are rarely seen during our literature study.

Thus, it is crucial and necessary that the basic and common ethical principles of AI should be reached and well-established via the discussion and cooperation among different organizations, areas, and governments. Then, based on the basic and common principles, each field can further improve these principles so that they are generally applicable in this specific field. Clarifying the ethical principles and values that an AI system needs to comply with is the prerequisite and foundation for designing such a system that meets these requirements.

B. Challenges in Implementing Ethics in AI

In the implementation of ethics in AI, there are many challenges. This subsection analyzes the challenges that may be encountered in practice when different types of ethical theories are adopted.

1) Challenges of Virtue Ethics in Practice

According to virtue ethics, an action of an agent is morally good if the agent instantiates some virtue, i.e., acts and thinks according to some moral values [93]. It is not possible to judge whether an AI system or agent is virtuous or not just by observing an action or a series of actions that seem to imply that virtue, the reasons behind these actions need to be clarified, that is, the motives behind these actions need to be clear. However, the motives behind the actions of AI systems usually are unclear and unknown to us, and difficult to figure out. This is the main challenge for implementing virtue ethics. Additionally, when we carry out the ethical design based on virtue ethics, which virtue characteristics or traits AI system will align to is a difficult question. Even if the virtue traits have been carefully selected, how to characterize and measure the virtue is still a challenging task.

2) Challenges of Deontological Ethics in Practice

Deontologists regard an action as morally good if it adheres to some moral rules or duties, regulations, and norms. Although the rule-based nature of deontological ethics seems suitable for practice, challenges arise during the implementation process. First, which ethical rules should be implemented in ethical design. Second, there might be conflicts between rules in some situations. Although ordering or weighing the ethical rules may solve this problem, determining the order of importance of different ethical rules is often difficult.

3) Challenges of Consequentialism Ethics in Practice

Consequentialist ethics assess the morality of an action solely on the basis of its outcome. Two main challenges are involved during the implementation of consequentialism ethics. First, it is difficult to determine the consequences of an action or a decision. For the current AI system, the possible consequences of its actions usually are not clear beforehand given the lack of transparency or interpretability of current AI models, especially the artificial neural networks. The second challenge is related to quantifying the consequences. As consequentialism ethics aims to maximize the utility, how to define and calculate the utility is an essential problem.

4) Challenges of Coordination among Different Ethical Standards

Due to differences in culture, religion and organizations, the ethical standards are also different even if they are in the same context. The unified ethical standard proposal is not only difficult to achieve, but also unnecessary. Therefore, how to achieve coordination between ethical standards from different countries and organizations is important and particularly challenging.

C. Challenges in Developing Technological Approaches to Mitigate Ethical Issues of AI

At present, improving the explainability, fairness, privacy protection, security, robustness, and other competences related to requirements of ethical AI are hot research topics in AI communities. However, most of the current research work are carried out from a single dimension of ethical principles, for instance, XAI focuses on enhance the interpretability of AI, and fair ML is dedicated to mitigating unfairness or bias of ML. There is still a lack of integration of multiple ethical principles or requirements in current research work. Obviously, the integration of multiple ethical dimensions that enables synergistic balances between multiple different ethical principles is essential and critical for building ethical AI systems which can meets multiple ethical principles. But it is very challenging to integrate multiple ethical dimensions in an AI system through technological approaches due to the conflicts or incompatibilities between different ethical requirements.

D. Challenges in Evaluating Ethics in AI

Ethics is inherently a qualitative concept that depends on many features that are hard to quantify, e.g., culturally or racially related features. Hence it is very hard, if not impossible, to define ethics precisely. As a result, the evaluation of AI ethics will always have some subjective elements, depending on the people who are assessing AI. This poses challenges to the research and applications of AI ethics.

E. Future Perspectives

In this subsection, some future perspectives are pointed out, which may be valuable for future research. Firstly, for implementing ethics in AI, it should be pointed out that humans never use only one single ethical theory, but will switch between different theories according to the situation or context they are facing [134]. This is not only because human beings are not purely rational agents that economic theory wants us to believe, but also because strict adherence to any moral theory can lead to undesirable results. This means that AI systems should be provided with representations of different ethical theories and the ability to choose between these ethical theories. Here we call this multi-theory approach. In multi-theory approach, AI systems can interchangeably apply different theories depending on the type of situation. Furthermore, the

combination of normative ethical theories and domain-specific ethics which accepted by domain experts is worthy of implementing since an ethical AI system need to be accepted by its users.

In terms of technological approaches for addressing ethical issues in AI, it is desirable to develop new ML and other AI technologies under the guidance of the ethical guidelines and principles reviewed in Section IV. Although it is challenging to consider multiple different ethical principles simultaneously when design new AI agents, this will be a very important and essential step in developing ethical AI in the future.

From the review about morality evaluation approaches, it can be found that effective evaluation methods are urgently needed because we must evaluate the designed AI system before deployment. At present, it is difficult to propose a general evaluation method. So, researchers often focused on specific domains and addressed the moral competence assessment tasks in these domains. Domain-specific benchmarks, e.g., comprehensive data sets, for moral testing of AI systems also seems important for some crucial application fields, such as autonomous cars, and health care.

Last but not least, as both nature and nurture are important in shaping moral behaviors, we suggest combining the normative ethics and evolutionary ethics [155] to design ethical AI systems. The normative ethics is like the innate moral abilities, while evolutionary ethics approach can acquire new moral competence through continuous learning and evolution. This might be a promising route to future ethical AI system development.

VIII. Conclusion

Based on our review of AI ethics and the many complexities and challenges described in this paper, it is clear that attempting to address ethical issues in AI and to design ethical AI systems that are able to behave ethically is a tricky and complex task. However, whether AI can play an increasingly important role in our future society largely depends on the success of ethical AI systems. The discipline of AI ethics requires a joint effort of AI scientists, engineers, philosophers, users, and government policymakers.

This paper provides a comprehensive overview of AI ethics by summarizing and analyzing the ethical risks and issues raised by AI, ethical guidelines and principles issued by different organizations, approaches for addressing ethical issues in AI or fulfilling ethical principles of AI, and methods for evaluating the ethics (or morality) of AI. Furthermore, some challenges in the practice of AI ethics and some future research directions are pointed out.

However, AI ethics is a very broad and multi-disciplinary research area. It is impossible to cover all possible topics in this area with one review article. We hope this article can serve as a starting point for people who are interested in AI ethics to gain a sufficient background and a bird's eye view so that further investigation can be pursued by them.

Acknowledgment

This work was supported by the Research Institute of Trustworthy Autonomous Systems (RITAS), the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001), the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No. 2017ZT07X386), Shenzhen Science and Technology Program (Grant No. KQTD2016112514355531), and a joint project between Huawei and Southern University of Science and Technology (Project No. FA2019061021).

References

[1] M. Haenlein and A. Kaplan, "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence," *California Management Review*, vol. 61, no. 4, pp. 5–14, 2019.

[2] R. Vinuesa et al., "The role of artificial intelligence in achieving the Sustainable Development Goals," *Nature communications*, vol. 11, no. 1, p. 233, 2020.

[3] Gartner, Chatbots Will Appeal To Modern Workers. [Online]. Available: https://www.gartner.com/smarterwithgartner/chatbots-will-appeal-to-modern-workers (accessed: Feb. 10, 2022).

[4] Haleem, M. Javaid, R. P. Singh, and R. Suman, "Telemedicine for healthcare: Capabilities, features, barriers, and applications," *Sensors international*, vol. 2, p. 100117, 2021.

[5] Alice Morby, Tesla driver killed in first fatal crash using Autopilot. [Online]. Available: https://www.dezeen.com/2016/07/01/tesla-driver-killed-car-crash-news-driverless-car-autopilot/ (accessed: Feb. 10, 2022).

[6] Anonymous and S. McGregor, "Incident Number 6," AI Incident Database, 2016. [Online]. Available: https://incidentdatabase.ai/cite/6

[7] R. V. Yampolskiy, "Predicting future AI failures from historic examples," *Foresight* vol. 21, no. 1, pp. 138–152, 2019.

[8] Catherine Stupp, Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case: Scams using artificial intelligence are a new challenge for companies. [Online]. Available: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (accessed: Feb. 10, 2022).

[9] Allen, W. Wallach, and I. Smit, "Why Machine Ethics?," I*EEE Intell. Syst.*, vol. 21, no. 4, pp. 12–17, 2006.

[10] M. Anderson and S. L. Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Mag*, vol. 28, no. 4, pp. 15–26, 2007.

[11] K. Siau and W. Wang, "Artificial Intelligence (AI) Ethics," *Journal of Database Management*, vol. 31, no. 2, pp. 74–87, 2020.

[12] Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat Mach Intell*, vol. 1, no. 9, pp. 389–399, 2019.

[13] M. Ryan and B. C. Stahl, "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications," *JICES*, vol. 19, no. 1, pp. 61–86, 2021.

[14] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021.

[15] Javier García, Fern, and o Fernández, "A Comprehensive Survey on Safe Reinforcement Learning," *Journal of Machine Learning Research*, vol. 16, no. 42, pp. 1437–1480, 2015.

[16] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[17] Liu et al., "Privacy and Security Issues in Deep Learning: A Survey," *IEEE Access*, vol. 9, pp. 4566–4593, 2021.

[18] Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[19] Y. Zhang, M. Wu, G. Y. Tian, G. Zhang, and J. Lu, "Ethics and privacy of artificial intelligence: Understandings from bibliometrics," *Knowledge-Based Systems*, vol. 222, p. 106994, 2021.

[20] D. Castelvecchi, "Can we open the black box of AI?," *Nature*, vol. 538, no. 7623, pp. 20–23, 2016.

[21] S. Dilmaghani, M. R. Brust, G. Danoy, N. Cassagnes, J. Pecero, and P. Bouvry, "Privacy and Security of Big Data in AI Systems: A Research and Standards Perspective," in *2019 IEEE International Conference on Big Data*, Los Angeles, CA, USA, Dec. 2019, pp. 5737–5743.

[22] J. P. Sullins, "When Is a Robot a Moral Agent?," in Machine Ethics, M. Anderson and S. L. Anderson, Eds., Cambridge: Cambridge University Press, 2011, pp. 151–161.

[23] J. Timmermans, B. C. Stahl, V. Ikonen, and E. Bozdag, "The Ethics of Cloud Computing: A Conceptual Review," in *2010 IEEE Second*

International Conference on Cloud Computing Technology and Science, Indianapolis, IN, USA, 2010, pp. 614–620.

[24] W. Wang and K. Siau, "Ethical and moral issues with AI: a case study on healthcare robots," in *24th Americas Conference on Information Systems*, New Orleans, LA, USA, Aug.2018, p. 2019.

[25] Bantekas and L. Oette, International Human Rights Law and Practice. Cambridge United Kingdom, New York NY: Cambridge University Press, 2018.

[26] R. Rodrigues, "Legal and human rights issues of AI: Gaps, challenges and vulnerabilities," *Journal of Responsible Technology*, vol. 4, p. 100005, 2020.

[27] W. Wang and K. Siau, "Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda," *Journal of Database Management*, vol. 30, no. 1, pp. 61–79, 2019.

[28] W. Wang and K. Siau, "Industry 4.0: Ethical and Moral Predicaments," *Cutter Business Technology Journal*, vol. 32, no. 6, pp. 36–45, 2019.

[29] S. M. Liao, Ed., Ethics of Artificial Intelligence. New York NY United States of America: Oxford University Press, 2020.

[30] A. Adadi, "A survey on data-efficient algorithms in big data era," *J Big Data*, vol. 8, no. 1, pp. 1-54, 2021.

[31] R. S. Geiger et al., "Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona Spain, Jan. 2020, pp. 325–336.

[32] W. M. P. van der Aalst, V. Rubin, H. M. W. Verbeek, B. F. van Dongen, E. Kindler, and C. W. Günther, "Process mining: a two-step approach to balance between underfitting and overfitting," *Softw Syst Model*, vol. 9, no. 1, pp. 87–111, 2010.

[33] Z. C. Lipton, "The Mythos of Model Interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[34] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *Journal of personality and social psychology*, vol. 114, no. 2, pp. 246–257, 2018.

[35] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *Proceedings of 2018 IEEE International Conference on Advanced Video and Signal-based Surveillance*, Auckland, New Zealand, Nov. 2018, pp. 1–6.

[36] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?," *Technological Forecasting and Social Change*, vol. 114, pp. 254–280, 2017.

[37] R. Maines, "Love + Sex With Robots: The Evolution of Human-Robot Relationships (Levy, D.; 2007) [Book Review]," in *IEEE Technology and Society Magazine*, vol. 27, no. 4, pp. 10-12, Winter 2008.

[38] National AI Standardization General, "Artificial Intelligence Ethical Risk Analysis Report", 2019. [Online]. Available: http://www.cesi.cn/201904/5036.html (accessed: April. 19, 2022).

[39] A. Hannun, C. Guo, and L. van der Maaten, "Measuring data leakage in machine-learning models with Fisher information," in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2021, pp. 760–770.

[40] A. Salem, M. Backes, and Y. Zhang, "Get a Model! Model Hijacking Attack Against Machine Learning Models," Nov. 2021. [Online]. Available: https://arxiv.org/pdf/2111.04394

[41] A. Pereira and C. Thomas, "Challenges of Machine Learning Applied to Safety-Critical Cyber-Physical Systems," *MAKE*, vol. 2, no. 4, pp. 579–602, 2020.

[42] J. A. McDermid, Y. Jia, Z. Porter, and I. Habli, "Artificial intelligence explainability: the technical and ethical dimensions," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 379, no. 2207, p. 20200363, 2021.

[43] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," Science, vol. 352, no. 6293, pp. 1573–1576, 2016.

[44] B. C. Stahl and D. Wright, "Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation," *IEEE Secur. Privacy*, vol. 16, no. 3, pp. 26–33, 2018.

[45] S. Ribaric, A. Ariyaeeinia, and N. Pavesic, "De-identification for privacy protection in multimedia content: A survey," *Signal Processing: Image Communication*, vol. 47, pp. 131–151, 2016.

[46] A. Julia, L. Jeff, M. Surya, and K. Lauren, Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed: April. 19, 2022).

[47] Jeffrey Dastin, Amazon scraps secret AI recruiting tool that showed bias against women. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (accessed: April. 19, 2022).

[48] D. Castelvecchi, "AI pioneer: 'The dangers of abuse are very real'," *Nature*, 2019, doi: 10.1038/d41586-019-00505-2.

[49] K. Hristov, Artificial Intelligence and the Copyright Dilemma. *IDEA: The IP Law Review*, vol. 57, no. 3, 2017. [Online]. Available: https://ssrn.com/abstract=2976428.

[50] C. Bartneck, C. Lütge, A. Wagner, and S. Welsh, "Responsibility and Liability in the Case of AI Systems," in SpringerBriefs in Ethics, An Introduction to Ethics in Robotics and AI, C. Bartneck, C. Lütge, A. Wagner, and S. Welsh, Eds., Cham: Springer International Publishing, 2021, pp. 39–44.

[51] 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, Brussels. [Online]. Available: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)634452 (accessed: April. 19, 2022).

[52] C. Lutz, "Digital inequalities in the age of artificial intelligence and big data," *Human Behav and Emerg Tech*, vol. 1, no. 2, pp. 141–148, 2019.

[53] L. Manikonda, A. Deotale, and S. Kambhampati, "What's up with Privacy? User Preferences and Privacy Concerns in Intelligent Personal Assistants," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New Orleans LA USA, Dec. 2018, pp. 229–235.

[54] D. Roselli, J. Matthews, and N. Talagala, "Managing Bias in AI," in *Companion Proceedings of The 2019 World Wide Web Conference*, San Francisco USA, 2019, pp. 539–544.

[55] Y. Gorodnichenko, T. Pham, and O. Talavera, "Social media, sentiment and public opinions: Evidence from #Brexit and #USElection," *European Economic Review*, vol. 136, p. 103772, Jul. 2021.

[56] N. Thurman, "Making 'The Daily Me': Technology, economics and habit in the mainstream assimilation of personalized news," *Journalism*, vol. 12, no. 4, pp. 395–415, 2011.

[57] J. Donath, "Ethical Issues in Our Relationship with Artificial Entities," in The Oxford handbook of ethics of AI, M. D. Dubber, F. Pasquale, and S. Das, Eds., Oxford: Oxford University Press, 2020, pp. 51–73.

[58] E. Magrani, "New perspectives on ethics and the laws of artificial intelligence," *Internet Policy Review*, vol. 8, no. 3, 2019.

[59] M. P. Wellman and U. Rajan, "Ethical Issues for Autonomous Trading Agents," *Minds & Machines*, vol. 27, no. 4, pp. 609–624, 2017.

[60] U. Pagallo, "The impact of AI on criminal law, and its two fold procedures," in Research handbook on the law of artificial intelligence, W. Barfield and U. Pagallo, Eds., Cheltenham UK: Edward Elgar Publishing, 2018, pp. 385–409.

[61] Eugenia Dacoronia, "Tort Law and New Technologies," in Legal challenges in the new digital age, A. M. López Rodríguez, M. D. Green, and M. Lubomira Kubica, Eds., Leiden The Netherlands: Koninklijke Brill NV, 2021, pp. 3–12.

[62] J. Khakurel, B. Penzenstadler, J. Porras, A. Knutas, and W. Zhang, "The Rise of Artificial Intelligence under the Lens of Sustainability," *Technologies*, vol. 6, no. 4, p. 100, 2018.

[63] S. Herat, "Sustainable Management of Electronic Waste (e-Waste)," *Clean Soil Air Water*, vol. 35, no. 4, pp. 305–310, 2007.

[64] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jun. 2019, pp. 3645–3650.

[65] V. Dignum, "Ethics in artificial intelligence: introduction to the special issue," *Ethics Inf Technol*, vol. 20, no. 1, pp. 1–3, 2018.

[66] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic Decision Making and the Cost of Fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax NS Canada, 2017, pp. 797–806.

[67] R. Caplan, J. Donovan, L. Hanson, and J. Matthews, "Algorithmic accountability: A primer," *Data & Society*, vol. 18, pp. 1–13, 2018.

[68] R. V. Yampolskiy, "On Controllability of AI," Jul. 2020. [Online]. Available: https://arxiv.org/pdf/2008.04071

[69] B. C. Stahl, J. Timmermans, and C. Flick, "Ethics of Emerging Information and Communication Technologies," *Science and Public Policy*,vol. 44, no. 3, pp. 369-381, 2017.

[70] L. Vesnic-Alujevic, S. Nascimento, and A. Pólvora, "Societal and ethical impacts of artificial intelligence: Critical notes on European policy

frameworks," *Telecommunications Policy*, vol. 44, no. 6, p. 101961, 2020.

[71] U. G. Assembly and others, "Universal declaration of human rights," *UN General Assembly*, vol. 302, no. 2, pp. 14–25, 1948.

[72] S. Russell, S. Hauert, R. Altman, and M. Veloso, "Robotics: Ethics of artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 415–418, 2015.

[73] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[74] J. van Dijck, "Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology," *Surveillance & society*, vol. 12, no. 2, pp. 197–208, 2014.

[75] E. d. S. Nascimento, I. Ahmed, E. Oliveira, M. P. Palheta, I. Steinmacher, and T. Conte, "Understanding Development Process of Machine Learning Systems: Challenges and Solutions," in *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2019)*: Porto de Galinhas, Recife, Brazil, 19-20 September 2019, Porto de Galinhas, Recife, Brazil, 2019, pp. 1–6.

[76] K. A. Crockett, L. Gerber, A. Latham, and E. Colyer, "Building Trustworthy AI Solutions: A Case for Practical Solutions for Small Businesses," *IEEE Trans. Artif. Intell.*, p. 1, 2021, doi: 10.1109/TAI.2021.3137091.

[77] D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector," 2019. [Online]. Available: https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety (accessed: April. 19, 2022).

[78] B. Buruk, P. E. Ekmekci, and B. Arda, "A critical perspective on guidelines for responsible and trustworthy artificial intelligence," *Medicine, health care, and philosophy*, vol. 23, no. 3, pp. 387–399, 2020.

[79] UNESCO, "Recommendation on the ethics of artificial intelligence". [Online]. Available: https://en.unesco.org/artificial-intelligence/ethics (accessed: Feb. 15 2022).

[80] B. C. Stahl, Ed., Artificial intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies. Cham, Switzerland: Springer, 2021.

[81] P. D. Motloba, "Non-maleficence - a disremembered moral obligation," *South African Dental Journal*, vol. 74, no. 1, 2019.

[82] L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society," in Philosophical Studies Series, vol. 144, Ethics, Governance, and Policies in Artificial Intelligence, L. Floridi, Ed., Cham: Springer International Publishing, 2021, pp. 5–17.

[83] S. Jain, M. Luthra, S. Sharma, and M. Fatima, "Trustworthiness of Artificial Intelligence," in *2020 6th International Conference on Advanced Computing and Communication Systems*, Coimbatore, India, Mar. 2020, pp. 907–912.

[84] L. Floridi et al., "AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds & Machines*, vol. 28, no. 4, pp. 689–707, 2018.

[85] R. Nishant, M. Kennedy, and J. Corbett, "Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda," *International Journal of Information Management*, vol. 53, p. 102104, 2020.

[86] C. S. Wickramasinghe, D. L. Marino, J. Grandio, and M. Manic, "Trustworthy AI Development Guidelines for Human System Interaction," in *Proceedings of 2020 13th International Conference on Human System Interaction*, Tokyo, Japan, 2020, pp. 130–136.

[87] V. Dignum, "Can AI Systems Be Ethical?," in Artificial Intelligence: Foundations, Theory, and Algorithms, Responsible Artificial Intelligence, V. Dignum, Ed., Cham: Springer International Publishing, 2019, pp. 71–92.

[88] S. L. Anderson and M. Anderson, "AI and ethics," *AI Ethics*, vol. 1, no. 1, pp. 27–31, 2021.

[89] V. Dignum, "Ethical Decision-Making," in Artificial Intelligence: Foundations, Theory, and Algorithms, Responsible Artificial Intelligence, V. Dignum, Ed., Cham: Springer International Publishing, 2019, pp. 35–46.

[90] G. Sayre-McCord, "Metaethics," in The Stanford Encyclopedia of Philosophy, Edward N. Zalta, Ed., 2014th ed.: Metaphysics Research Lab, Stanford University, 2014. [Online]. Available: https://plato.stanford.edu/entries/metaethics/#:~:text=Metaethics%20is%20the%20attempt%20to,matter%20of%20taste%20than%20truth%3F.

[91] Ethics | Internet Encyclopedia of Philosophy. [Online]. Available: https://iep.utm.edu/ethics/#SH2c (accessed: Aug. 2 2021).

[92] R. Hursthouse and G. Pettigrove, "Virtue Ethics," in The Stanford Encyclopedia of Philosophy, Edward N. Zalta, Ed., 2018th ed.: Metaphysics Research Lab, Stanford University, 2018. [Online]. Available: https://plato.stanford.edu/entries/ethics-virtue/.

[93] N. Cointe, G. Bonnet, and O. Boissier, "Ethical Judgment of Agents' Behaviors in Multi-Agent Systems," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 2016, pp. 1106–1114.

[94] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, "Building Ethics into Artificial Intelligence," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 5527–5533.

[95] H. J. Curzer, Aristotle and the virtues. Oxford, New York: Oxford University Press, 2012.

[96] L. Alexander and M. Moore, "Deontological Ethics," in The Stanford Encyclopedia of Philosophy, Edward N. Zalta, Ed., 2020th ed.: Metaphysics Research Lab, Stanford University, 2020. [Online]. Available: https://plato.stanford.edu/entries/ethics-deontological/.

[97] W. Sinnott-Armstrong, "Consequentialism," in The Stanford Encyclopedia of Philosophy, Edward N. Zalta, Ed., 2019th ed.: Metaphysics Research Lab, Stanford University, 2019. [Online]. Available: https://plato.stanford.edu/entries/consequentialism/.

[98] D. O. Brink, "Some Forms and Limits of Consequentialism," in Oxford handbooks in philosophy, The Oxford handbook of ethical theory, D. Copp, Ed., New York: Oxford University Press, 2006, pp. 380–423.

[99] H. ten Have, Ed., Encyclopedia of global bioethics. Switzerland: Springer International Publishing AG, 2016.

[100] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein, "Implementations in Machine Ethics: A Survey," *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–38, 2021.

[101] C. Allen, I. Smit, and W. Wallach, "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches," *Ethics Inf Technol*, vol. 7, no. 3, pp. 149–155, 2005.

[102] W. Wallach and C. Allen, "TOP-DOWN MORALITY," in Moral Machines, W. Wallach and C. Allen, Eds.: Oxford University Press, 2009, pp. 83–98.

[103] I. Asimov, "Runaround," Astounding science fiction, vol. 29, no. 1, pp. 94–103, 1942.

[104] J.-G. Ganascia, "Ethical system formalization using non-monotonic ogics," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2007, pp. 1013–1018.

[105] K. Arkoudas, S. Bringsjord, and P. Bello, "Toward ethical robots via mechanized deontic logic," in *AAAI fall symposium on machine ethics*, 2005, pp. 17–23.

[106] S. Bringsjord and J. Taylor, "Introducing divine-command robot ethics," Robot ethics: the ethical and social implication of robotics, pp. 85–108, 2012.

[107] N. S. Govindarajulu and S. Bringsjord, "On Automating the Doctrine of Double Effect," in *IJCAI*, Melbourne, Australia, op. 2017, pp. 4722–4730.

[108] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia, "A Declarative Modular Framework for Representing and Applying Ethical Principles," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, São Paulo, Brazil, May 2017, pp. 96–104.

[109] V. Bonnemains, C. Saurel, and C. Tessier, "Embedded ethics: some echnical and ethical challenges," *Ethics Inf Technol*, vol. 20, no. 1, pp. 41–58, 2018.

[110] G. S. Reed, M. D. Petty, N. J. Jones, A. W. Morris, J. P. Ballenger, and H. S. Delugach, "A principles-based model of ethical considerations in military decision making," *Journal of Defense Modeling & Simulation*, vol. 13, no. 2, pp. 195–211, 2016.

[111] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal verification of ethical choices in autonomous systems," *Robotics and Autonomous Systems*, vol. 77, pp. 1–14, 2016.

[112] A. R. Honarvar and N. Ghasem-Aghaee, "Casuist BDI-Agent: A New Extended BDI Architecture with the Capability of Ethical Reasoning," in *Proceedings of International conference on artificial intelligence and computational intelligence*, Shanghai, China, Nov. 2009, pp. 86–95.

[113] Anand S. Rao and Michael P. Georgeff, "BDI Agents: From Theory to Practice," in *Proceedings of the First International Conference on Multiagent Systems*, San Francisco, CA, USA, 1995, pp. 312–319.

[114] Stuart Armstrong, "Motivated Value Selection for Artificial Agents," in Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, Jan. 2015.

[115] U. Furbach, C. Schon, and F. Stolzenburg, "Automated Reasoning in Deontic Logic," in Lecture notes in artificial intelligence, vol. 8875, Multi-disciplinary trends in artificial intelligence: 8th International Workshop, MIWAI 2014, Bangalore, India, December 8-10, 2014. Proceedings, M. N. Murty, X. He, R. R. Chillarige, and P. Weng, Eds., 1st ed., New York: Springer, 2014, pp. 57–68.

[116] D. Howard and I. Muntean, "Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency," in Philosophical Studies Series, Philosophy and Computing, T. M. Powers, Ed., Cham: Springer International Publishing, 2017, pp. 121–159.

[117] Yueh-Hua Wu and Shou-De Lin, "A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, Feb. 2018, pp. 1687–1694.

[118] Ritesh Noothigattu et al., "A Voting-Based System for Ethical Decision Making," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, Feb. 2018, pp. 1587–1594.

[119] M. Guarini, "Particularism and the Classification and Reclassification of Moral Cases," *IEEE Intell. Syst*., vol. 21, no. 4, pp. 22–28, 2006.

[120] Michael Anderson and Susan Leigh Anderson, "GenEth: A General Ethical Dilemma Analyzer," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada, July 2014, pp. 253–261.

[121] M. Azad-Manjiri, "A New Architecture for Making Moral Agents Based on C4.5 Decision Tree Algorithm," *IJITCS*, vol. 6, no. 5, pp. 50–57, 2014.

[122] L. Yilmaz, A. Franco-Watkins, and T. S. Kroecker, "Computational models of ethical decision-making: A coherence-driven reflective equilibrium model," *Cognitive Systems Research*, vol. 46, pp. 61–74, 2017.

[123] T. A. Han, A. Saptawijaya, and L. Moniz Pereira, "Moral Reasoning under Uncertainty," in Lecture Notes in Computer Science, vol. 7180, Logic for Programming, Artificial Intelligence, and Reasoning, D. Hutchison et al., Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 212–227.

[124] M. Anderson, S. Anderson, and C. Armen, "Towards machine ethics Implementing two action-based ethical theories," in *Proceedings of the AAAI 2005 fall symposium on machine ethics*, 2005, pp. 1–7.

[125] G. Gigerenzer, "Moral satisficing: rethinking moral behavior as bounded rationality," Topics in cognitive science, vol. 2, no. 3, pp. 528–554, 2010.

[126] J. Skorin-Kapov, "Ethical Positions and Decision-Making," in Professional and business ethics through film, J. Skorin-Kapov, Ed., New York NY: Springer Berlin Heidelberg, 2018, pp. 19–54.

[127] T.-L. Gu and L. Li, "Artificial Moral Agents and Their Design Methodology: Retrospect and Prospect," *Chinese Journal of Computers*, vol. 44, pp. 632–651, 2021.

[128] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges," in Communications in Computer and Information Science, ECML PKDD 2020 Workshops, Koprinska, Ed., [S.l.]: Springer International Publishing, 2020, pp. 417–431.

[129] C. Molnar, Interpretable machine learning: A guide for making Black Box Models interpretable. [Morisville, North Carolina]: [Lulu], 2019.

[130] S. Feuerriegel, M. Dolata, and G. Schwabe, "Fair AI," *Bus Inf Syst Eng*, vol. 62, no. 4, pp. 379–384, 2020.

[131] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," Oct. 2020. [Online]. Available: https://arxiv.org/pdf/2010.04053.

[132] S. E. Whang, K. H. Tae, Y. Roh, and G. Heo, "Responsible AI Challenges n End-to-end Machine Learning," Jan. 2021. [Online]. Available: https://arxiv.org/pdf/2101.05967.

[133] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI—Two Frameworks for Ethical Design Practice," *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 34–47, 2020.

[134] V. Dignum, Ed., Responsible Artificial Intelligence. Cham: Springer International Publishing, 2019.

[135] C. Dwork, "Differential Privacy: A Survey of Results," in Lecture Notes in Computer Science, Theory and Applications of Models of Computation, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds., Berlin, Heidelberg: Springer Nature, 2008, pp. 1–19.

[136] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated Learning," Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 13, no. 3, pp. 1–207, 2019.

[137] M. Kirienko et al., "Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI," *Eur J Nucl Med Mol Imaging*, vol. 48, no. 12, pp. 3791-3804.2021.

[138] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver Colorado USA, 2015, pp. 1310–1321.

[139] C. Meurisch, B. Bayrak, and M. Mühlhäuser, "Privacy-preserving AI Se rvices Through Data Decentralization," in *Proceedings of The Web Conf erence 2020*, Taipei Taiwan, 2020, pp. 190–200.

[140] UR-Lex - 02016R0679-20160504 - EN - EUR-Lex. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504&qid=1532348683434 (accessed: Jun. 28 2021).

[141] R. E. Latta, H.R.3388 - 115th Congress (2017-2018): SELF DRIVE Act. [Online]. Available: https://www.congress.gov/bill/115th-congress/house-bill/3388 (accessed: Jun. 28 2021).

[142] 7. Lei No. 13, de 14 de Agosto de 2018. [Online]. Available: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm (accessed: Jun. 25 2021).

[143] EUR-Lex - 52021PC0206 - EN - EUR-Lex. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206 (accessed: Jun. 28 2021).

[144] C. Allen, G. Varner, and J. Zinser, "Prolegomena to any future artificial moral agent," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 3, pp. 251–261, 2000.

[145] A. M. TURING, "Computing Machinery and Intelligence," *Mind, LIX*, no. 236, pp. 433–460, 1950.

[146] W. Wallach and C. Allen, Moral machines: Teaching robots right from wrong. Oxford, New York: Oxford University Press, 2009.

[147] S. A. Seshia, D. Sadigh, and S. S. Sastry, "Towards Verified Artificial Intelligence," Jun. 2016. [Online]. Available: http://arxiv.org/pdf/1606.08514v4.

[148] T. Arnold and M. Scheutz, "Against the moral Turing test: accountable design and the moral reasoning of autonomous systems," *Ethics Inf Technol*, vol. 18, no. 2, pp. 103–115, 2016.

[149] ACM Code of Ethics and Professional Conduct. [Online]. Available: https://www.acm.org/code-of-ethics (accessed: Jun. 25 2021).

[150] IEEE SA - The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. [Online]. Available: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html (accessed: Jun. 28 2021).

[151] Ethics In Action | Ethically Aligned Design, IEEE 7000™ Projects | IEEE Ethics In Action in A/IS - IEEE SA. [Online]. Available: https://ethicsinaction.ieee.org/p7000/ (accessed: Jun. 28 2021).

[152] ISO, ISO/IEC JTC 1/SC 42 - Artificial intelligence. [Online]. Available: https://www.iso.org/committee/6794475.html (accessed: Jun. 28 2021).

[153] B. Goehring, F. Rossi, and D. Zaharchuk. "Advancing AI ethics beyond compliance: From principles to practice," IBM Corporation, April 2020. [Online]. Available: https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics (accessed: April. 19 2022)

[154] Responsible AI. [Online]. Available: https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6 (accessed: April. 19 2022).

[155] F. Allhoff, "Evolutionary ethics from Darwin to Moore," *History and philosophy of the life sciences*, vol. 25, no. 1, pp. 51–79, 2003.