School of Informatics

Informatics Research Review
Whatever I Write About

[redacted]

Abstract

Artificial intelligence (AI), especially robots, is increasingly participating in various fields and gradually occupying an increasingly important position, which has aroused the thinking and attention of various industries on artificial intelligence ethics. This article comprehensively reviews the research progress of ethical artificial intelligence. Based on bibliometric analysis, it is found that three different artificial intelligence attitudes: "traditional" and "cautious" and "optimistic" lead to the ethics of the relationship between "human-machine" Conflict, systematically reviewed the ethics of artificial intelligence research results from the perspectives of artificial intelligence, moral algorithms and social ethics. And to look forward to the future research of artificial intelligence ethics.

Date: Wednesday 20th January, 2021

Supervisor: [redacted]

1 Introduction

With the rapid development of artificial intelligence (AI), AI has gradually penetrated into all walks of life, among which artificial intelligence ethics has become one of the core issues of discussion and research in all walks of life (Moor, 2006; Wallach, 2010) . Artificial intelligence ethics has become one of the focuses of people's attention. It includes not only the study of technology itself, but also the study of the relationship between man, machine, and environment under the premise of human value. The uncertainty of the rapid development of artificial intelligence has brought new ethical issues and risks, especially the singularity theory, which makes many people worry that the rapid development of artificial intelligence will lead to some unpredictable disasters, such as car accidents caused by driverless cars. And the death of the nursing robot. At the same time, artificial intelligence also brings discrimination, infringement of personal privacy, changes in employment structure, challenges to the principles of international relations, and other issues, which will have a profound impact on individual rights, government supervision, economic and social development, and even global governance (Wallach, 2010; Bostrom, 2014; Diakopoulos, 2015). Therefore, more people need to pay attention to the ethics of artificial intelligence, prevent ethical problems caused by the rapid development of artificial intelligence, and lay a moral foundation for the future development of artificial intelligence.

Correctly recognize the failure cases of AI in the development process (Yampolskiy Spellchecker,2016) is necessary. Data explosion and depth of the development of the learning algorithm, the assignment of the decision making human make machine tools from passive to active part(intelligent agent) change and don not need human intervention of "perception, thinking, action" reflects the ability of self- learning evolution is in the lead of algorithm for artificial intelligence system errors, as well as the uncertainty of technology lead to consequences hard to forecast and quantitative evaluation. Is it because the bounded rationality of human causes the ethical risk of artificial intelligence, or is artificial intelligence far from having the attribute of ethics? These problems have not yet formed a unified consensus in the academic circle.

Our paper is aimed to use the method of content analysis to try to clarify the development context, current situation and trend of artificial intelligence ethics research, and comment on the existing research from moral philosophy, moral algorithm, design ethics and social ethics, in order to provide reference for future theoretical research and practice.The structure of the paper is as follows:

- Next part conducts a quantitative analysis of the literature related to artificial intelligence ethics.
- The third part analyzes the ethical concerns caused by "man-machine" relationship.
- The fourth part summarizes the research perspective and content of artificial intelligence ethics issues.
- The fifth part is the summary and future research prospects.

2 Quantitative Analysis of Literature Related to Artificial Intelligence Ethics

The research background of artificial intelligence ethics spans a wide range, involving computer science, artificial intelligence, robotics, ethics, philosophy, biology, sociology, religion, etc.

This paper is based on Web of Science (WOS) and CNKI database, CiteSpace as the paper Offering measuring tool, analyze the research context of artificial intelligence ethics in the form of knowledge graphs. Choose AI ethics, machine morality, robert ethics, machine ethics, moral machine, value alignment, artificial ethics Search for keywords such as morality, technology ethics, AI security, friendly AI, etc.

"Artificial intelligence" shows an increasing trend in theoretical research, capital investment and social attention, as shown in below figure.

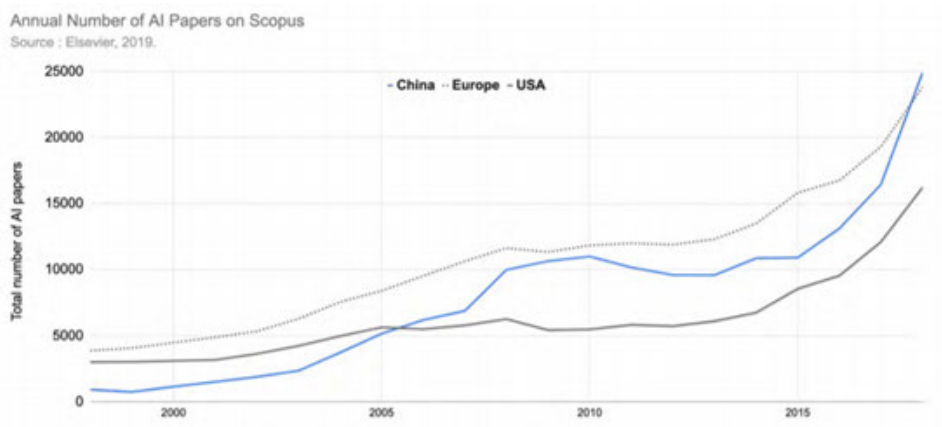

Figure 1: Annually published AI papers on Scopus by region (1998—2019)

Judging from the changing trend of the curve, academic research interest has shown a continuous growth trend. The development trend of artificial intelligence is similar to the emergence and development of new technologies. It can be predicted that artificial intelligence and its related fields will be the focus of research and attention in the future.

Subject headings are the refining and generalization of the core content of the literature, and their frequency also reflects the research hotspots and trends of the literature to a certain extent. Based on the collected artificial intelligence ethics literature, this paper conducts a co-occurrence analysis of the subject terms, as in below figure:

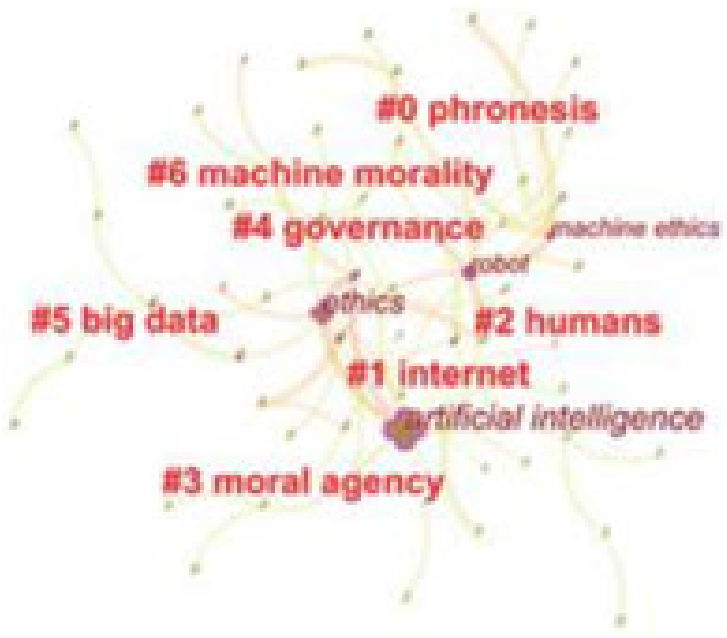

Figure 2: Co-occurrence network analysis of the subject terms of artificial intelligence ethics in WOS database

Machine morality focuses on what moral properties a robot should have and how to achieve these properties (Malle, 2016). Machine ethics focuses on the design, application and practice of robots, and how to get along with humans (Anderson et al., 2005). Robot ethics and the proposal of a road map (Di Robotica and Veruggio, 2006), promote cross-cultural discussion of the influence of scholars on the use of robotics, and understand the generalized machine ethics from the perspective of the degree of implantation of ethical factors (Moor, 2006) .

3 Ethical concerns arising from human-machine relationship

When using computer technology to construct ethical models within the framework of ethical design and apply them to practice, research is innovative and forward-looking (Danielson, 1992). When the "agent" is given the attribute of "artificial", this kind of intelligence belongs to the machine, not the creature. Intelligent machines possess tool rationality through self-deep learning and can make judgments and predictions about certain specific things, thus forming a new "human-machine" relationship. People have different views on this "human-machine" relationship. There are three main positions and viewpoints in the academic and practical fields, as shown in Table 1.

Attitudes towards AI	Representative personage	Viewpoint	Evaluation of ethical issues
traditional	Neumann Turing Dreyfus Searle	Human intelligence is the ultimate state of artificial intelligence,which cannot be surpassed	(1) AI is only a means and tool, which cannot distinguish between "good and evil" and "good and bad". The key lies in the evaluation of good and evil of the application consequences; (2) AI development is still in its infancy and cannot ensure that AI will necessarily obey human settings (3) Research on the relationship between weak artificial intelligence, strong artificial intelligence, super artificial intelligence, and the relationship between human intelligence and artificial intelligence should be strengthened
cautious	Hawking BillGates Mask	The development of AI will threaten the survival of human beings, and there is the possibility of "doing evil"	
optimistic	Herbert A.Simon Ray Kurzweil Markram	AI will eventually reach and surpass the level of human intelligence, "singularity theory", "man-machine coexistence"	(1) Unilateral and isolated view of the positive aspects of AI, ignoring or covering up the negative effects of AI;(2) Take a cautious development attitude and formulate machine ethics and value system through government supervision

Table 1: Three main attitude towards AI and evaluation of ethical issues

One is the "traditional school" represented by von Neumann and Alan Turing. They believe that human beings are always in a dominant position and that artificial intelligence will never surpass human intelligence. This view stems from the belief that "the creator must be better than the created thing." In the 1970s, it was concluded from the biological and psychological level that artificial intelligence will fail (Dreyfus, 1972), who believed that "the current risk facing humans is not the advent of super-intelligent machines, but the emergence of low-intelligence people.". The "Chinese room" model of the 1980s confirmed this view (Searle, 1980).

The second is "caution" represented by Hawking, Bill Gates and Musk (Brooks, 2017). They believe that humans should be in awe of the rise of artificial intelligence. When a machine has self-awareness and can even make decisions autonomously through its own "nervous" and "consciousness" and connects with other artificial intelligences, artificial intelligence will eventually pose a real threat. The development of thorough artificial intelligence may even lead to the destruction of mankind. The machine may update itself, redesign itself at an ever-increasing speed, and constantly upgrade and update itself. The human biological evolution process is too slow to compete with intelligent machines at all, and may eventually be surpassed by machines.

However, Zuckerberg, Kai-Fu Lee, Enda Wu and others put forward opposite views on the "artificial intelligence threat theory". The point of contention is whether and when powerful artificial intelligence will emerge. Musk's background mainly refers to "powerful artificial intelligence"

[Ng GW, Leung WC, 2020], that is, the ability to handle multiple types of tasks and adapt to unexpected situations. It represents the public's expression of strong artificial intelligence and super artificial intelligence that may be out of control and threaten the future of humanity. The "artificial intelligence" mentioned by Zuckerberg is a narrow professional field of artificial intelligence capabilities. It is the industry's continuous exploration of artificial intelligence research, development and application from a functional and commercial perspective. He believes that any technology has its advantages and disadvantages. The key lies in whether developers are cautious about new technologies and how they are used, rather than wanting to slow down the development of new technologies in the face of technical threats. These two typical views stem from different understandings of "artificial intelligence".

The third is the "optimist" represented by Herbert A. Simon, who believes that artificial intelligence will definitely reach and surpass human intelligence. The "Singularity Theory" predicts that machine intelligence will surpass human intelligence in 2045 (Kurzweil, 2014). In 2009, the scientists in charge of the Blue Brain Project declared that they are expected to create the first "thinking" machine in the history of science in the future, which may have feelings, pain, desires and even fears. However, some people questioned its feasibility. They believe that the development of artificial intelligence is still in the initial stage and is far from reaching the stage where it can be matched with human intelligence. The "optimists" believe that in the future, artificial intelligence and human intelligence will be two forms of social development, and "human-machine coexistence" will become a relatively ideal state of human society.

Although they are optimistic about the development of artificial intelligence, human society is not yet ready for the era of artificial intelligence in terms of philosophy and rationality. For the rapid development of artificial intelligence, more and more ethical issues are now appearing in front of the crowd. We need to call on humans to use artificial intelligence correctly, and to formulate ethics and value systems related to robots as soon as possible. Establish standards, seek the "precautionary principle" of risk assessment, and then assess hazards and risks on this basis (Wallach & Allen, 2010).

4 Artificial intelligence ethics issues research perspective and content

The research field of artificial intelligence mainly focuses on the social ethics, law and safety of robots, involving technical unemployment, lethal autonomous weapons, algorithm fairness, moral judgment, value consistency and other issues. According to the status and trend of artificial intelligence ethics research, the related attributes of artificial intelligence and various fields of society, this article understands the research of artificial intelligence ethics from four aspects: artificial intelligence moral algorithms and social ethics.

4.1 Ethical algorithms for artificial intelligence

- **Algorithmic discrimination**

  An algorithm is a series of operations of a computer to solve a specific problem or complete a certain result (Diakopoulos, 2015). Humans have begun to transfer the decision-making power of some things to highly intelligent algorithms. However, the algorithm lacks advanced intelligence such as learning, inference, and association, and it is difficult to explain the decision logic behind the behavior. Because human behavior is subjectively motivated

and intentional, the ethics of the algorithm needs to be analyzed, which is also the focus and difficulty in the algorithm design process

Algorithm is a relatively objective mathematical expression, and many people would think that algorithmic decision-making tendencies are fair. However, human values and ethics are not included in the digital system, and the management of data is more difficult to control in the context of global interconnection. The opacity of data and algorithms reflects the imbalance between data rights, which will inevitably lead to algorithm discrimination (AlgorithmsBias). The United States has revealed the law of algorithm discrimination in computer (Joseph et al., 2016), criminology (Goel et al., 2016), political science (Veale and Binns, 2017), human resource management and other fields, and discovered relevant Data and evidence, the research level is deeper. The subjective tendency of discrimination is difficult to judge, which has also led foreign scholars to use empirical and case research methods to make suggestions in their research (Edelman and Luca, 2014)

- **The root cause of algorithmic discrimination**

  The reason for algorithm discrimination is mainly attributed to the two aspects of algorithm technology and data input (Executive Office of the President, 2016). The structure of the machine completely designed by humans and the data received in the learning process form a "cognition" of the world, leading to Wrong decisions occur in certain scenarios. Algorithm design is the subjective choice and judgment of technicians. It is questionable whether they write moral rules and laws into the program fairly and impartially. This makes the algorithm inherit all the prejudices of human decision-makers.

  Algorithms are inseparable from the support of big data, and the validity and accuracy of data will affect the accuracy of algorithm decisions and predictions. Biased defects of the data itself, the status disparity of large and small samples, and sensitive attributes all lead to the inevitability of algorithm discrimination. Discriminatory data will also have a discriminatory tendency after calculation. Innovation will also make discrimination and inequality of opportunity continue.

- **The solution to algorithmic discrimination**

  The logic basis of algorithm is a multi-context system, which adopts interdisciplinary principles from design, implementation, testing to promotion, so as to be more inclusive. It is proposed to design a more "fair" algorithm to prevent discrimination from the source of the algorithm design, so that the actions of the artificial intelligence body conform to relevant ethical norms. Maintain the transparency of the algorithm, formulate the design principles of equal opportunities, and make the power of the algorithm work, especially the interests of the disadvantaged and easily overlooked. Platforms such as OpenAI have launched some artificial intelligence open source campaigns to ensure the transparency and fairness of algorithms. Google advocates "equality of opportunity" in practice to avoid discrimination based on a set of sensitive attributes. Of course, the relationship between algorithm transparency and trade secrets and national security needs to be further balanced.

  More and more countries, regions, and organizations have introduced regulatory measures, including design standards, performance standards, responsibility standards, etc., and even extended to core content involving artificial intelligence algorithms and data. Some scholars suggested that the data processor's collection and processing of health-related information should be explained in written form; the competition order of the big data market should be protected under the premise of fairness, justice and economic efficiency (Townley et al., 2017).

4.2 Artificial Intelligence Social Ethics

At present, no better social contract has been reached on the application of artificial intelligence. People suspect that the rapid development of artificial intelligence technology will bring great danger to human society. Among them, responsibility and security, and privacy protection are the main concerns for the rapid development of artificial intelligence.

- **Responsibility and safety**

  One of the core issues of the robot ethics debate is: If a robot causes an accident, who should be held responsible for the robot's behavior? UNESCO and the World Commission for Science and Technology (2015) proposed two countermeasures. One is that intelligent robots should take responsibility. Some scholars have long used the specific APACHE system as an example to explain that machines may bear some moral responsibilities (Friedman and Kahn, 1992), because no one can bear the responsibility for robot failures (Matthias, 2004). However, some scholars believe that robots cannot bear all moral responsibilities, because the autonomous consciousness of computers is still far behind human expectations, and robots cannot achieve the ability of independent thinking like humans.

  People's basic ethical consensus is that people are the core. Under such an ethical framework, it is difficult to effectively define the behavior of robots, and it is also difficult to judge who should bear the moral responsibility. Therefore, an open method to ensure algorithm transparency should be promoted in the design of robots to help humans maintain a clear understanding of the moral responsibility of machines.

  Another countermeasure is to share responsibility for everyone involved in the process of robot invention, authorization, and distribution. Due to the limitations of the algorithm, the algorithm cannot foresee all the possible behaviors of the robot in the process of getting along with humans. It is also difficult to accurately assess the danger of the robot, and it cannot fully control the future behavior of the robot. Therefore, when a problem occurs, the programmer should not take full responsibility, but it cannot be completely exempted from responsibility. In order to avoid no one taking responsibility when an accident occurs, an insurance system should be implemented so that everyone involved in the accident can share the responsibility. And set up and fill in perfect laws and regulations as the basis for judgment, fill the "responsibility gap", and find ways to solve the problem of liability attribution.

- **Social employment problem**

  The widespread application of artificial intelligence liberates mankind from dangerous, boring and difficult tasks. While sharing the huge material wealth and convenience of life brought by AI, mankind also bears huge psychological pressure, such as whether AI will cause labor-intensive work Large-scale unemployment in areas such as type, repetitiveness, and proceduralization affects people's income and welfare, creates new gaps between the rich and the poor and social divisions, which will cause social crises, reduce social security and trigger turbulence (Bruun & Duka, 2018). The McKinsey report[25] predicts that by 2030, 390 million people in the world will change jobs and 800 million people will be unemployed due to advances in technology and the popularization of AI. Therefore, the social stability problems caused by the development of AI in the future are also serious.

The rapid development of artificial intelligence is not just a simple machine replacing all people's work, but also allowing machines to complete dangerous, replaceable, and meaningless tasks to facilitate people's lives. People should use machines flexibly in their work and production, rather than being replaced by machines, but should achieve a realm of cooperation and coordination between man and machine. Therefore, artificial intelligence will not fundamentally affect human employment and employment. In the first industrial revolution, the appearance of the steam engine replaced many people's jobs, but created more new job opportunities, and the overall unemployment rate in society stabilized. Of course, not everyone has the ability to overcome technical and social obstacles. This will reduce the bargaining power of labor in negotiations with capital. In the future, some people's jobs may be completely replaced by machines, causing new social problems.

- **privacy problem**

  Algorithms require a large amount of high-quality data support. The data exchange between various activities has become a new source of value creation. Personal data is easily collected and used by organizations and is relatively passive, which will weaken individuals' personal data. Control and management. If these data are disclosed by institutions, it will have an impact on personal privacy. Therefore, it is necessary to strengthen the protection of personal privacy in the deep learning process.

  In order to prevent the expansion and deterioration of privacy issues, it is necessary to start with both ethical and technical aspects. Establishing data ethics centered on human rights, emphasizing the transparency of algorithms, eliminating data islands, advocating standardized data sharing, preventing data abuse, and aiming to eliminate dataism's worship of data freedom. Technically, there are already applications of tools such as anonymization[Bayardo and Agrawal 2005], differentiated privacy, and decision matrices to implement privacy protection. "Artificial Intelligence + Blockchain Technology + Quantum Technology" may also provide more effective solutions in the future, but before that, relevant entities such as governments, enterprises, and civil organizations are still required to work together to prevent large-scale data leaks. The issue of privacy is to rebuild the dominant position of people in the era of big data, construct the free relationship between people and technology, and between people and data, and eliminate the adverse effects of the mechanical world outlook. These ideas have their pros and cons, and at present it is still an open issue that needs to be explored.

5 Summary & future research prospective

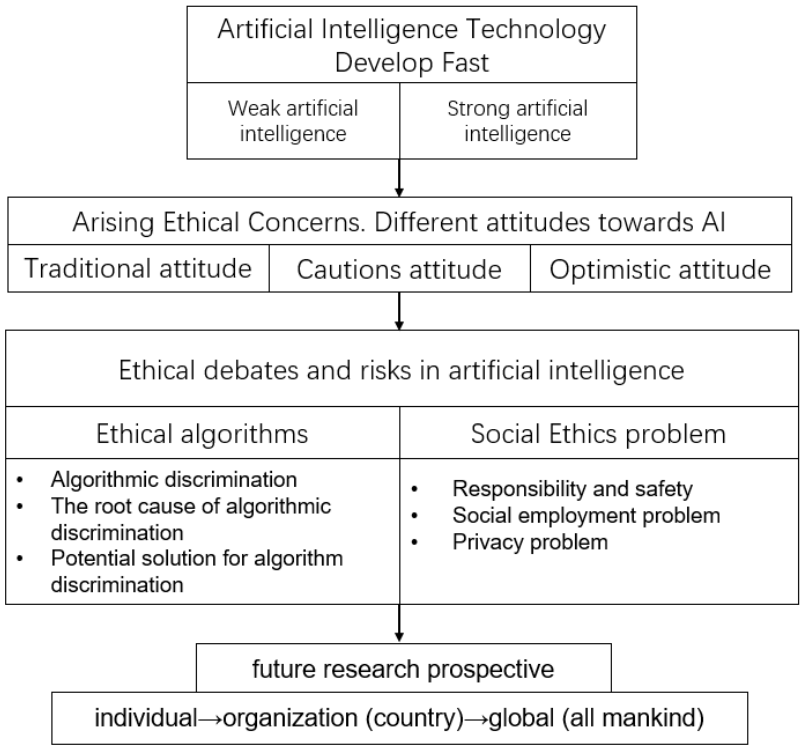

Figure 3: The main research framework of artificial intelligence ethics

This paper selects journals with large impact factors and reviews the research literature related to artificial intelligence ethics. Whether it is problems arising from the application of artificial intelligence on the basis of existing ethics, or new ethical problems affecting human behavior, the different manifestations of weak artificial intelligence and strong artificial intelligence have triggered the necessity and urgency of ethical research. These are two stages in the development of artificial intelligence. Humans should have a certain sense of anxiety and risk awareness about artificial intelligence.

Research on artificial intelligence ethics needs to start from the general issues of artificial intelligence ethics (moral philosophy, moral algorithms), rather than focusing on the benefits of technology itself to humans, and the harm that technology may cause to human society. Therefore, it is very necessary to establish a new ethical code, conduct in-depth research and discussion in different fields, analyze the ethical issues that may be caused by artificial intelligence, and propose corresponding effective solutions to different AI ethical issues.

The development of artificial intelligence is a global process, not just a certain organization, a certain country's affairs, and the ethical issues that it may produce should also be of common concern to the world. Therefore, the solution of artificial intelligence ethical problems requires the joint efforts of the whole world. Taking into account the different national conditions of each country, the ethics and ethics are not the same, so we should discuss together and negotiate the problem-solving standards and norms that all countries agree. Scholars represented by Walleck put forward the ethical necessity of studying artificial intelligence from a personal perspective, which has attracted widespread attention from the international community. Different organizations, companies, and countries have also proposed their own artificial intelligence ethics, and

the recommendations of these standards have basically reached a consensus. In the future, it is necessary to establish an artificial intelligence ethics system from a global level, that is, to carry out the research of "awareness→standards→system" of artificial intelligence ethics along the path of "individual→organization (country)→global (all mankind)".

References

[1]Moor J H. The nature, importance, and difficulty of machine ethics[J]. IEEE Intelligent Systems,2006, 21(4):18-21.

[2]]Wallach W, Allen C. Moral machines: Teaching robots right from wrong[M]. Oxford: Oxford University Press, 2010.

[3] Bostrom N. Superintelligence: Paths, dangers, strategies[M]. Oxford: Oxford University Press, 2014.

[4] Diakopoulos N. Algorithmic accountability: Journalistic investigation of computational power structures[J]. Digital Journalism,2015,3(3):398-415.

[5] Roman V. Yampolskiy,M. S. Spellchecker. Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures 2016

[6] Malle B F. Integrating robot ethics and machine morality: The study and design of moral competence in robots[J]. Ethics and Information Technology,2016, 18(4): 243-256.

[7] Anderson M,Anderson S L. Machine ethics: Creating an ethical intelligent agent[J]. AI Magazine,2007, 28(4): 15-26

[8] Danielson P. Artificial morality: Virtuous robots for virtual games[M]. London: Routledge Press, 1992.

[9] Searle J R. Minds, brains, and programs[J]. Behavioral and Brain Sciences,1980, 3(3): 417-424.

[10] Dreyfus, H.L. 1979. What computers can't do—the limits of artificial intelligence. New York: Harper and Row.

[11] Brooks, R. [2017] Artificial Intelligence is not as smart as you (or Elon Musk) think. Available: https://techcrunch.com/2017/07/25/artificial-intelligence-is-not-as-smart-as-you-or-elon-musk-think/

[12] Ng GW, Leung WC, 2020. Strong artificial intelligence and consciousness. Journal of Artificial Intelligence and Consciousness, 7(1), pp. 63-72.

[13] Kurzweil, R. 2014. "The Singularity is Near." In Ethics and Emerging Technologies

[14] Boden M, Bryson J, Caldwell D, et al. Principles of robotics: Regulating robots in the real world[J]. Connection Science,2017, 29(2): 124-129.

[15] Diakopoulos N. Algorithmic accountability: Journalistic investigation of computational power structures[J]. Digital Journalism,2015, 3(3): 398-415.

[16] Veale M, Binns R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data[J]. Big Data & Society,2017, 4(2): 1-17. [35]

[17] Matthew Joseph,et al. 2016b. Fair Algorithms for Infinite and Contextual Bandits. arXiv:1610.09559 [cs] (Oct. 2016). http://arxiv.org/abs/1610.09559 arXiv: 1610.09559.

[18] Sharad Goel, Justin M Rao, and Ravi Shroff. 2016. Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. Annals of Applied Statistics 10, 1 (2016), 365–394.

[19] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2):2053951717743530, 2017.

[20]]Edelman B, Luca M. Digital discrimination: The case of Airbnb.com[R]. Working Papers 14-054, 2014.

[21] Townley C, Morrison E, Yeung K. Big data and personalized price discrimination in EU competition law[J]. Yearbook of European Law,2017, 36: 683-748

[22] Friedman, B. and Kahn, P.H. Jr. (1992). Human agency and responsible computing: implications for computer system design. Journal of Systems Software, 17, 7–14.

[23] Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology 6(3): 175–183.

[24] Bruun, E. P. G., & Duka, A. (2018). Artificial intelligence, jobs and the future of work: Racing with the machines. Basic Income Studies, 13(2), 1–15. https://doi.org/10.1515/bis2018-0018

[25] McKinsey Global Institute. Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages(November, 2017 — Report)

[26] BAYARDO, R. J. AND AGRAWAL, R. 2005. Data privacy through optimal k-anonymization. In Proceedings of the International Conference on Data Engineering (ICDE'05).