

Zhiyi Liu
Yejie Zheng

AI Ethics and Governance

Black Mirror and Order

 Springer

AI Ethics and Governance

Zhiyi Liu · Yejie Zheng

AI Ethics and Governance

Black Mirror and Order

Zhiyi Liu
Computational Law and AI Ethics Research
Center of Shanghai Jiaotong University
Shanghai, China

Yejie Zheng
Computational Law and AI Ethics Research
Center of Shanghai Jiaotong University
Shanghai, China

ISBN 978-981-19-2530-6 ISBN 978-981-19-2531-3 (eBook)
<https://doi.org/10.1007/978-981-19-2531-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface: Metaphor and Future of Humanism in the Wave of Science and Technology

Look in the mirror, and you will get a clear understanding of yourself.

Future Fantasy

In 1957, *Popular Mechanics*, a renowned American tech magazine, prudently proposed that roads in America would become pneumatic pipes by the early 2020s.

“Letters will be delivered by rockets. A letter can be sent from New York to California in a few hours.” U.S. The Postmaster General said delightedly in 1959.

Voyage to the Prehistoric Planet, an American science fiction movie, was released in 1965. In the movie, after establishing a permanent base on the moon, astronauts set about to travel tens of millions of kilometers to explore Venus and discovered multiple exotic animals and plants on the surface of the planet, including dinosaurs that had been extinct on Earth.

In 1997, *WIRED* magazine made a similar prediction; that is, humans would be able to land on Mars approximately 2020.

2020 has arrived now, and various unexpected situations have been unfolding in an unrealistic way. The Australian mountain fire in January killed 1.2 billion animals, and the African locust plague and swine fever in February starved 19 million people. Since March, COVID-19 has caused over 10 million cases of pneumonia. In June, China’s floods hit 24 provinces. However, in previous film and television works, 2020 sounded like an extremely distant year.

In the science fiction world created in film and television works, 2020 is a very special year. By now, human beings should have been able to easily travel to wherever they like and enjoy a high level of scientific and technological civilization. According to the incomplete statistics on Wikipedia, there are at least 97 films with timelines relating to 2020. In the imagination of these movies, cars in *Blade Runner* can drive through skyscrapers in cities; human beings in *Johnny Mnemonic* can directly transmit data with their brains without saying anything as long as they put on a type of helmet; Metalbeast, who is transformed by scientists into a part man, part machine

cyborg in *Robocop*, works to guard the peace of Detroit; and the robots created by humans in *The Terminator* even manage to take over the world.

However, the world now has not become cyberpunk as in science fiction movies. Cars still cannot drive in the air, and AI cannot yet replace human beings. “Brainter-net” has not happened, and it is impossible to buy an air ticket to Mars. Too many things have taken place without warning in 2020.

Although wild imaginations in various movies and television works as well as science fiction have not come true by now. Driven by science and technology, the greatest productivity today, the era of man-machine coexistence has come to some extent. The two industrial revolutions have led to a giant leap in human productivity. The fact that the knowledge and wealth being created during the past hundred years have outnumbered the combined achievements gained over the past several thousand years has made mankind unprecedentedly ambitious. Practical technologies such as robots, AI and cloud computing, which have become increasingly integrated into people’s daily lives, it is reasonable for human beings to expect a more automated life.

Marx once said “the differences of all kinds of economic times, do not lie in what is produced, but how to produce, with what means of labor to produce”. When technology has become a basic way of thinking and living, the prelude to the era of intelligent economy has opened. We need to realize that in the digital society, even if human beings dominate, the relationship between humans and technology also turns out to be a dynamically changing game-playing process. Different types of intellectual products, including driverless cars, face recognition and smart pills, educational robots, etc., are commonly seen in everyday life. These intelligent crystals represent unparalleled achievements on the combination of cognitive reason and practical reason. Especially since 2016, AI has outperformed human beings repeatedly, demonstrating the strong momentum of technologic rationality.

Being heavily influenced by traditional Western philosophy typified by Socrates, Plato and Hegel, enlightenment rationality becomes the gene and basic color of this rationalistic and revealing digital society. However, the stubborn pursuit of rationality and blind pride will inevitably make rationality originally serving as purpose change qualitatively, namely, utilitarian rationality has coercively blinded modern people to the opposite of a happy life.

Actually, the “digital society” itself implies contradictions. It is a combination of “society full of intelligent machines” and “human society”. Since machines and human beings are completely different in nature and attributes, their integration is bound to be an extremely complicated process full of contradictions. Such contradictions have appeared in the early days. Repeatedly incredible victories won by machines, man’s overreliance on machines, accidental injuries “conducted” by machines, and even horrific scenes displayed in science fiction movies about how machines rule and enslave humans have made the relationship between man and machine go far beyond master and servant during its constant upgrades. AI was originally a great invention for us to lead a better and more rational life, but technological rationality has deteriorated into excuses to satisfy human desires in multiple triumphs

over human performance. Rationality is losing its nature in the cage of desire, while desires are behaving unscrupulously under the guise of civilization.

As mentioned above, technology has brought unprecedented power to human beings in science fiction movies, but they also show the “dark side” amplified by science and technology: abnormal human desires are exposed in front of super AI, and people are constantly suffering from their boundless avarice. The beautiful robot in *Ex Machina* is created for human carnal desires. In *AI*, people try to avoid facing death by resurrecting their dead relatives with AI technology. The hero in *Her* falls in love with his virtual girlfriend to escape normal human interaction in reality. Obviously, old tough robots can no longer satisfy the numerous demands of human beings, and AI products are being created on a large scale to stimulate both the physical and spiritual desires of humans for various purposes, including commerce, politics, and war.

It can be seen that while promoting social progress, science and technology may intensify the alienation and distortion of human nature, which in turn will further speed up technological advances and cause more negative effects. With more advanced and intelligent AI products coming down the pike, what they can do is just filling the gap of desire temporarily instead of building a harmonious ethical order between human and machine world. The protagonist of *Her* is addicted to the virtual intelligent world and finds it difficult to adapt to life in the real world. The modern version of Allegory of the Cave suggests that it is not the illusion in the dark but the light of reality that makes us afraid at present. The super AI in the movie is not just a fantasy; it indicates that modern people are no longer satisfied with cold technological products but want to invent a human-like plaything to meet their selfish desires through omniscient role play. By doing so, they are free from the responsibility of reality. Does this mean that technology is nourishing and encouraging the selfish gene of humankind?

Moreover, the relationship between AI and humankind is a topic interconnected with the modernity of human history, as Durkheim stated: “Modern times derive from the past and finally become the past although they always position themselves against the past”. After entering the postmodern period, it is still difficult for us to find a similar relationship pattern in history, especially how to get along with AI in which they lack experience. The imaginations for the future, whether optimistic or pessimistic, all reflect different assumptions and expectations about the good and evil nature of human beings.

Marx’s concept of alienation is based on the assumption that all men are “born” to be kind but later corrupted by society. In contrast, Durkheim’s “anomie concept” holds that men are “born” to be a stubborn organism whose conceit must be strictly restrained by society. The former view is considered close to that of Rousseau and the latter to that of Hobbes, but all of them are related to the discussion on human nature brought about by modernity.

In this sense, the imaginations for the future are actually rules and paradigm propositions for humans to enter a world of “human-computer symbiosis”, rather than just about the imaginations of science fiction on social development.

Metaphor of Black Mirror

As an important product of the era of human-computer symbiosis, ubiquitous electronic devices are penetrating into people's everyday lives. On the one hand, they draw an insurmountable boundary between the so-called "real" world and the "virtual" world; on the other hand, they are the only medium for the two to communicate with each other. The screen will bring us endless information once it is opened, and people will be immersed in the illusion of sound and light it provides without being aware of its border. The black mirror's materiality is exposed to us only when it is closed. The mirror carries an unfathomable force of technological ethics. As Nietzsche said, "When you look long into an abyss, the abyss looks into you."

China and the West hold different perspectives on the understanding of multiple metaphors of mirrors. Le Daiyun discussed the internal differences between Chinese and Western cultures by taking "mirror" as the clue of her article "mirror metaphor in Chinese and Western Poetics". She pointed out that westerners focus more on the reflection function of "mirror". Stendhal, Goethe, Shelley, and other writers compared literature to a mirror to highlight the reflection of literature on reality. Relatively speaking, Chinese culture prefers to use idioms such as Jinghua Shuiyue and Yijing Zhaojing, which literally means the reflection of a flower in the mirror and of the moon in the water and using the images in the mirror to realize introspection, respectively. Deeply influenced by Buddhism and Taoism, Chinese culture places more emphasis on the features of emptiness and inclusiveness of mirrors. It is often used to describe the inner feelings of humans. The most representative image should be the alternate name of *The Dream of Red Mansions*, "*Fengyue Baojian*" (mirror in breeze and moonlight).

When looking at other cultures, we find that mirrors have more complex and diverse cultural connotations. For example, buildings such as Mausoleum of King of the Light in Shirazi, a city in southwest Iran, and Golestan Palace in Tehran, the capital, are mainly decorated with fragmented mirrors, which form the dense geometric ornamental motifs in the inner wall of the building. Leaping lights crack into innumerable tiny light spots through the reflection of the fragmented mirrors, making up a resplendent and flexible view with a sense of holiness. More importantly, this decorative way takes advantage of the properties of mirrors while circumventing religious taboos—Islam forbids the use of images of people or any creature with eyes for decorative purposes, and countless small and interlacing mirror shards just avoid any recognizable figure.

After completing his famous fairy tale *Alice in Wonderland*, Lewis Carroll, an English writer, created another *writer through the Looking-Glass and What Alice Found* in 1871. This time Alice does not fall into the rabbit hole but enters a mirror world through the mirror at home. As an intermediary, the mirror gives this dreamland some rules contrary to reality. For instance, the closer you want to get to a distant mountain, the farther away you are from it; you have to turn a cake first if you want to cut it and so on. Meanwhile, the mirror also divides the world into two parts, a surface world and its opposite, an absurd world filled with the satire and parody of

the former one. The opposite of the surface phenomenon world is the absurd world of satire and parody. It can be said that the introduction of “mirror” adds a touch of allegorical dimension to this fairy tale and interrupts its original narrative rhythm. At the end of the story, the author describes how Alice wakes up and excitedly looks for the real counterparts of the characters appearing in her dream, such as the Red Queen, the White Queen and the Humpty and Dumpty, at great length. In the tale, the mirror also separates the phenomenal world from the mirror world, and the latter implies the reality of the former in the opposite way.

In contrast, the cultural metaphor of the “mirror” in the British drama “Black Mirror” is much more complex. Differing from the metaphors of the above classical mirror, the mirror here is no longer the dividing mark between real and illusory binary imagination. Instead, it allows the mirror world to occupy and replace the real world by reducing its presence. In the play, the traditional metaphor of mirror separating the two worlds becomes something like a two-way transparent membrane that allows the fusion and transition of subjects in the mirror and their realistic counterparts. Such imagination keenly grasps the features of the digital mirror and displays them in an ironic and absurd way. Undoubtedly, the “mirror metaphor” in culture has something to do with the construction of subjects, and the *Black Mirror* also uses mirrors to remind us of the dilemma faced by the increasingly blurred and broken subjects in the technical age.

On the one hand, *Black Mirror* profoundly reveals the conditions of people who live in a highly mediated posthuman world. The mediated existence mode subverts the perfect fantasy of human beings as rational subjects as well as their dominant position at the technical level. Lacan’s “mirror stage” theory notes that self-concept is built entirely through identification with others (mirror image), which is external. The “self” is deceived from the very beginning by the mirror image, that is to say, the mirror images we see are an imagined “ideal self”, and it is the product of our imagination. This makes our desires become others’ desires, and we turn into the “silent majority”. People hide behind the black mirror, unleashing their dark natures. This seems to be driven by the nature and desire of selves, but no one can escape from being influenced and utilized by the media manipulator. From this point, media, with great symbolizing capability, becomes the “the big Other” proposed by Lacan. As such, the media has completed the “castration” of human beings and pushed the unipolarity of the human desire structure. A medium catalyzed by technology can either bring hope to mankind or just destroy everything like fire.

On the other hand, *Black Mirror* further deepens the thoughts of alienation via the demonstration of a large number of possible cyborg forms. Cyborg is a combination of living body and media technology. Cyborg in the TV drama can be divided into two types: physical cyborg and cyberspace agent. The former exists in reality, while the latter exists in cyberspace. They interpenetrate to each other. What challenge does Cyborg bring us? It makes us think about whether humans have essence. Do “I” still exist if my body is completely replaced by others (machines)? Is “consciousness” reliable at that moment?

It can be argued that this high-profile cultural work is not only the product of technically pessimistic anxiety, and its greatest value actually lies in its scientific and

technological reflection of the coexistence of criticism and reflexivity. As a well-regarded episode of *Black Mirror*, *Fifteen Million Merits* has built a “beautiful new world” in which all people have to amuse themselves to death. Humans in this world are surrounded by a diverse variety of scientific and technological electronic screens every day, and all they can do is keep riding on the indoor sports bike to provide power for all electronic screens and earn “merit” for themselves. Those whose merits reach 15 million can participate in a TV program called *Hot Shot* (an acrimonious version of *Britain’s Got Talent*). Managing to be scouted by the judges on the program through participation is the only way a person can get rid of the fate of pedaling every day and be forced to watch assorted entertainment and pornographic shows. To help his beloved girl Abi realize her music dream, Bing, a black man donates his merits to Abi and accompanies her on the stage of *Hot Shot*. It is on this stage that the song about “love” is sounded for the first time: “You can blame me, try to shame me. However, anyone, who knows what love is, will understand.” Ironically, the judges neither understand “love” nor care about Abi’s dream. Instead, they even suggest that she seek development on pornographic channels with a tone of menace.

According to the scenarios of *Black Mirror*, Abi finally accepts the suggestion that “love” is excluded from the “world”, and only endless “sex” are being left in this system, which is similar to the story written by Huxley 80 years ago: sex can be enjoyed to one’s heart content, but love is not a necessity. Indignant Bing makes every effort to earn back all the merits that need to return to the stage. After the performance, he holds a fragment of the broken electronic screen against his throat and directs a torrent of abuse at the judges present: “And all you see up here, its not people. You don’t see people up here, it’s all fodder... We don’t know any better. All we know is fake fodder and buying shift. That’s how we speak to each other, how we express ourselves is buying shift. The peak of our dreams is a new app for our Dopple; it doesn’t exist! We ride day in and day out going where? All tiny cells and tiny screens and bigger cells and bigger screens.”

It can be said that this passionate speech has basically covered multiple topics involved in this “unfettered drama”, which includes technology prediction/fable, consumer society, media control, technical training, and human alienation. Of course, the “beautiful new world” in the teleplay exposes not only unavoidable problems caused by the “impact” brought by the technology but also all sorts of crux in modern society with the help of the “mirror” of science and technology. To what extent can an AI product implanted memory chip replace its real human counterpart (Be Right Back)? Can the cloned virtual consciousness be entitled to human rights (White Christmas)? Waldo, a virtual character who is well versed in the verbal tricks of the mass media in *The Waldo Moment*, finally becomes popular in political circles, which forms an interesting contrast with Trump’s win. In addition, *White Bear* and *Nosedive* displays a series of current social behaviors, including taking photos, onlooking, scoring, giving likes, etc., through the power of science and technology.

This is the “Black Mirror”—when the light on the screen fades, the dark panel finally reveals the truth being covered, leaving “the world becoming ugly first, and then going out”. Technology here demonstrates its desolate side. However, what

cannot be ignored here is the mirror reflecting interactive images. It carries reflexivity as a metaphor. Once human beings enter the virtual world, their subjectivities are deconstructed first and then reconstructed into another form in a new encoding mode. When we reconstruct retrospectively, we obtain a distorted “I”. This is the huge gap between the “second self” (i.e., cyberspace agent) and the real self. More seriously, the “second self” will cause harm to the real self, and the Internet facilitates the substitution of the second self with the real self. However, after that, the second self cannot possess human emotions such as love and hatred. Then, comes a reflexive paradox, which makes post humanism move toward its opposite, leaving post humanism hovering like a ghost in the spiral structures of “spirit-body-spirit” and “humanism-post humanism-humanism”.

The black mirror metaphor is the reflection of human nature amid the carnival brought by the waves of technology, which is also the key for us to understand the whole book and promote the exploration of a new paradigm of social governance in reflection.

Ethics and Governance

Returning to the digital society itself, let us reflect on what is the original intention for inventing and manufacturing AI. Because we want to pursue a harmonious and beautiful life, a civilized society in which human beings can be friendly coexist with intelligent machines. Such a society is defined by the following three points:

First, the digital society is a real world instead of a virtual space. If we blindly pursue a technical and rational lifestyle and become lost in the trap of excessive consumption, we will forget the true colors of life and unconsciously live an “unreal life”.

Since the twentieth century, phenomenology has unfurled the banner of “attending to the things themselves”. People’s reflection and criticism on scientific and technological rationality and instrumental rationality have never stopped. In a sense, the existence of the digital society is exactly shown through such ubiquitous digital quantification. Tasks such as monitoring a manhole cover, recording the track of a car, and even reading someone’s mind can be comprehensively demonstrated by almost any AI entity via succinct and infallible numbers.

However, real human society is quite a different story. It is a life world. Human beings, the source of strength in the living world, follow the rule that existence precedes essence. For AI, the situation changes. People are active, generative, creative, transcendent and vibrant, and the corresponding life world is ever-changing, rich and higher than the scientific world. Changing our ways of life and thinking by copying the quantitative approaches in the scientific world down to the last detail is bound to destroy people’s subjectivity and initiative, leaving a happy life impossible. Only by taking the life world as the starting point can we more reasonably explain and understand the scientific world represented by the digital society. The application of AI is not for self-protection and continuation of our species but creates a life helper

to liberate people from daily chores, give better play to the creativity in humans, and enable them to have more leisure to enjoy life as well as the convenience of the digital society.

Second, the digital society is free and comprehensive and not unique to the elite. In other words, digital society needs to liberate not a small number of social elites but most ordinary people. AI may be a tool differentiating social and cultural, but on the other side of the struggle, the contradictions among people may also undergo subtle changes in the digital society. Especially when these contradictions become increasingly sharp, the necessity of friendly communication among people is further highlighted. After all, human beings are not AI. Humans' essence must be formed in social practice, and the meaning of life also needs to be constructed through the perceptions of subject experience. Only by getting along with people (not machines) and widely experiencing society can human beings live a more cheerful and unforgettable life and gain the most cordial feelings of happiness in the most natural state. We fully expect a more civilized digital society in which friendly relationships among people, social fairness and justice are more valued.

Third, the digital society features happiness rather than indulgence. Various novel AI products, such as household robots, helpers for the elderly, child care robots and pet robots, are constantly stimulating our consumption desire. They free us from physical fatigue to some extent and provide us with all sorts of sensory pleasure.

People often mistake these pleasure sensations as Epicurean's principle of joy, which emphasizes not only physical relief but also inner peace. Pure physical pleasure is a kind of negative happiness, but those who can overcome desires within them can acquire true pleasure. In the digital society, real happiness never stays only at the sensory level and certainly comes from the enrichment and tranquillity of human spirit. People can embrace their limitations and imperfections as human beings and refuse to be enslaved by desires, generating negative pleasure. Men should live a rational life without disguise, that is, a human life rather than a god's life. We can truly appreciate the beauty of the digital society only if we are able to control our desires and return to human rationality.

As we are still far from the ideal digital society, the technology industry is used to looking forward and paying more attention to product innovation, technology application and industrial realization, which is a good thing for the industry. When besuited senior executives and dedicated programmers in the office building are pressing ahead with the above work in unison, few are willing to spend time carefully spotting and predicting problems at the turning point in a rear-view mirror or even accepting a beneficial practice. This is the challenge we are facing. In other words, we can use the well-organized information in the past to grasp the growing trend of science and technology and confront the increasingly severe social and ethical problems head-on in the field, which is significant to any technology company on the "highway".

In recent years, some quality companies have started to pay more attention to the rear-view mirror because they keep constantly adjusting their development directions and speeds. In fact, a number of large technology companies have focused attention on the ethical problems of AI and took action as early as 2017. Since 2011, Microsoft,

IBM, and the University of Toronto have realized that it is extremely necessary to overcome bias and accomplish fairness in AI. This is because from the perspective of technology, most algorithm models in AI work like black boxes. Usually, researchers only know the input, output and quality, but they cannot explain the internal working mechanism of these algorithms. If AI researchers adopt biased data sets or add their own prejudices when adjusting parameters, unfair AI outputs may occur.

To solve this problem, MSRA created a project named FATE (Fairness Accountability Transparency and Ethics in AI) in 2014 to increase the fairness, reliability, and transparency of AI. The project group organizes a symposium called FAT in ML every year and invites many senior technical experts in AI to share new research progress. Deepmind, a British company acquired by Google, also founded a project named Deepmind & Society in 2017 and put forward two goals. One is to assist technicians in practising ethics, and the other is to help society predict and guide the influence of AI for the benefit of all.

For technology companies, AI is far from a panacea but is a very risky technology. Thus, they need to be capable of “looking forward and rethinking the past” to be competitive in the industry. Before the “black box” of AI can be clearly understood and fully explained, a good suggestion is to formulate the basic laws for AI. Of course, as the impact of AI has gone far beyond the average organization, countries worldwide are laying down AI governance principles and proposing corresponding strategies suitable for local AI development. The most representative AI governance law should be the EU’s ecological governance law based on AI ethics.

As early as January 2015, JURI decided to form a working group to study legal issues related to the development of robots and AI. In May 2016, JURI issued the *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics*, calling on the European Commission to assess the impact of AI. In January 2017, JURI formally put forward a wide range of suggestions on robot civil legislation and the proposal to formulate “Charter on Robotics”. In May, the European Economic and Social Committee (EESC) released a statement on AI and pointed out an initiative to establish AI ethical norms and a standard system for AI monitoring and certification in response to the opportunities and challenges brought by AI in eleven fields, including ethics, security, and privacy. In October of the same year, the European Council proposed that the EU should have a sense of urgency to cope with the new trend of AI and guarantee high-level data protection, digital rights and the formulation of relevant ethical standards. The organization also invited the European Commission to give solutions to deal with the new trend of AI in early 2018. To solve the ethical problems caused by the development and application of AI, the EU has set AI ethics and governance as the key to future legislative work. Undoubtedly, the EU is taking the lead in promoting localization ethics to governance.

On the basis of the current situation, we suppose that AI governance is an indispensable proposition for organizations (countries, communities, enterprises, households, etc.) entering modern society. Therefore, it is necessary to write an academic monograph ranging from ethics to governance to explore the systematic logic of human beings from social ethical reflection to ecological governance, which can be regarded as an innovative attempt from technical ethics to technical governance. It

can be seen that since we are experiencing the fourth scientific and technological revolution, the ethical confusion resulting from technical development has become a new normal when we talk about big data, AI, gene pairs, and reproductive cloning. This process involves an obvious paradox. As a means for human beings to expand the world and nature, technology also extends to human beings through rapid self-extension. In short, human beings have eventually become a part of humanized nature. The explosive development of modern science and technology has driven us to review the definition of life, especially the relationship between humans, the “subject”, and other species. At the moment, the humans who seem to have “controlled” everything through technology are getting confused. We raise the eternal propositions again as our ancestors did thousands of years ago. What is the essence of life? What is the core difference in value between man and machine? Only when we determine these propositions can we have a clear thought when we think about the ethics of AI.

The results of the digital technology and AI revolution cannot be predicted because history can never tell us what will happen in the future. Severe direct threat, long-term structural imbalance, and tense situations have triggered a heated debate on AI ethics and governance throughout the world. In the debate, the term “ethics” is often used to describe legitimate concerns about the potential destruction of AI. The most controversial part in the discussion on AI ethics and governance is all sorts of framework definitions of AI principles worldwide, which are mainly put forward by large Internet platforms, multinational corporations, international nongovernmental organizations, and governments around the world.

Although there are some subtle and key differences among these ethical principles, each of them emphasizes that the future AI should be safe, interpretable, fair and reliable, and it should benefit the whole society. All these definitions seem to reach an international consensus; that is, AI should serve the welfare of mankind, be people-centered, responsible and trustworthy, always let human beings take the initiative and be kept under supervision.

However, these positive principle frameworks confirm that today’s ethics and governance capabilities are insufficient to prevent or mitigate the destructive power of AI, which has a fatal global impact and historical significance. Almost all frameworks analyzed the risks of AI in a narrow sense without considering the relationship between the dual attributes of technology and actual social, political, economic and international affairs. These frameworks ignored that the historical trajectories of society, politics, economics, and international affairs are likely to be strengthened rather than changed by AI. AI will make an increasing number of independent decisions, but it cannot escape human control and obtain complete autonomy in the short term. We can never rely on AI to assume the role of an excellent, super beneficial, and people-centered compass to guide humans to live a generally fair and dignified life. Although AI principles are easy to establish, given the complexity and uncertainty of the risks of AI, it will be more difficult to implement AI principles according to the newly formed governance methods.

To solve the problems above, we need to have an interdisciplinary knowledge system. In this book, we will discuss not only humanities frameworks such as ethics, political economics, philosophy, and sociology involved in technical ethics but also

science and engineering frameworks such as computer science, biology, AI, etc. We believe that only by breaking the boundaries of disciplines and expanding our knowledge to the cross fields of corresponding disciplines can we generate helpful ideas for complex problem-solving in a world of uncertainty. Fortunately, we have already begun to make rules for AI before it acquires real intelligence, and that is also the reason why we decided to write this book. We are expecting to expand and construct humanism in the era of AI.

For today's people, breaking through the early ideas of the Enlightenment era and constantly exploring human beings, the external world and the relationship between them is an essential and positive choice. Such exploration is multidirectional, including not only the continuous research and invention on life science, AI, and other science and technology but also research on the solutions to these existing or upcoming phenomena and even prevention measures. The latter requires collaborative cooperation among humanities, social sciences, and natural sciences.

If you review the past billions of years of the planet, *Homo sapiens* created human civilization in only 7,000 years and became the most important force on the planet that has got rid of the constraints of nature and been able to decide on its own development. What is the difference between humans and other animals that makes them a special species? Among multiple theories, the "culture driven theory" based on information transmission efficiency proposed by Kevin Leyland, a British scholar, has been well supported. The theory holds that the difference between human beings and other animals lies in the former having rich oral or symbolic language, which can help them learn more from culture and inherit more than the animals. The amount of cultural reserves of a species and the duration of its cultural characteristics in the population will increase exponentially with the rise in fidelity. Among the existing organisms, humans are the only animals that have broken through the threshold of cultural change to form the so-called "coevolution of gene and culture" process. In other words, the reason why large-scale cooperation occurs in human society is that we have social learning and educating abilities with which we can not only copy the cooperation mechanisms in other organisms but also generate new ones of our own, thus forming an unparalleled population.

Hence, the foothold of our research on AI ethics is humanism; that is, we need to safeguard people's central dominant position through human-centered cultural value inheritance and the construction of a social knowledge system. We can see that the humanistic spirit in the era of AI aims to continuously promote and plan human development and evolution where possible and improve human beings with the help of rapidly changing technology. If human nature cannot develop, gradual elimination and extinction rather than just suspension and standstill will be their final destinies. Humanism in the era of AI covers the following self-confidence and spirit: the development and evolution of human beings have shown the trends of unity of culture as well as the unity of physics, which forms the basis for people's positive views on dealing with human prospects.

Over the past few years, scientists have proposed using "humans" as a calendar unit of the earth. The proposal of this geological concept is based on vital observation and research data, which represent the current situation brought to modern humans

by the 7,000-year history of human civilization. Before the Anthropocene, the power to change the earth's ecology and climate on a large scale came from the earth itself, while natural selection has been gradually replaced by artificial design since humans emerged. Moreover, life forms have been extended from the organic field to the inorganic field, and artificial design has also been put into nature to shape the environment. Most importantly, human beings are using the power of technology to influence their own evolutionary direction, including bioengineering, cyborg engineering, nonorganic bioengineering, etc. All the problems in the fields above are closely related to the ethics of AI. Essentially, it is something about self-cognition and technical ethics, which is also the basis of the book that we want the readers to understand. What is the essence of the ethical problems of AI? How can a reasonable order be established to achieve man-machine symbiosis? We will elaborate on these propositions in the following twelve chapters. The content of the book, ranging from the modernity of social development to the discussion on the social risk of digintelligence, from AI medical discussion to in-depth synthesis technology, and from computational law to man-machine cogovernance, is expected to provide enough space for readers to understand AI ethics and governance that might matter the survival of human beings. We are glad to welcome all readers to embark on a "fantasy thinking journey" about the future of mankind with us.

Shanghai, China
Shanghai, China

Zhiyi Liu
Yejie Zheng

Contents

1	New Problems in the Era of AI	1
1.1	What Makes a Human Being a Human Being? AI and Gene Editing	2
1.2	Man-Machine Symbiosis: Challenges and Problems in the Near Future	7
1.3	Data Ethics: Political Economy in the New Oil Age	12
2	Ethical Enlightenment in the Age of Intelligence	17
2.1	Development of Human Consciousness and Machine Consciousness	19
2.2	Looking at AI from the Differences Between the East and the West	22
2.3	Ethics and Moral Responsibilities of AI	27
3	Digintelligence Risk Society is Around the Corner	33
3.1	AI Ethics in Digintelligence Risk Society	35
3.2	Will Winter Come?	39
3.3	Data Privacy and Social Responsibility	42
4	AI Medical Treatment: Epidemic, Death and Love	47
4.1	Black Swan Incidents: Epidemic and Influenza	48
4.2	Breakthrough and Ethical Principle of AI Medical Treatment	52
4.3	Restart AI and Expand the Meaning Boundary of Life	57
5	“Secret” Left by Turing—Privacy Computing	63
5.1	Ethical Turn of In-depth Learning: Theory of Materialization of Morality	64
5.2	Fantasia in the Turing Era—From Encryption to Computing	69
5.3	Secret Greatness: Creating Privacy Enhancing Technology	73

6	AI and Robot: Darwin and Rebellious Machine	79
6.1	Paradigm Revolution of Cognitive Science: Taking Consciousness Research as an Example	80
6.2	Human-Computer Symbiosis Under a “Brain-Computer Interface”	82
6.3	Automatic Driving: “Survival of the Fittest”	87
7	Virtual World Under AI: Augmented Reality and Deep Synthesis	95
7.1	AI+AR: A New World Under Augmented Reality	96
7.2	New Risks of Deep Synthesis Technology	101
7.3	Masked Virtual Digital Humans	105
8	Start of the “Age of Exploration” of AI Governance	111
8.1	New Channel or Old Ticket	112
8.2	Excellence, Trust and Reliability	116
8.3	Human Versus AI: “Victory” of Anthropocentrism	120
9	AI Legislation in Computational Society	127
9.1	Who Do We “Legislate” For?	128
9.2	“What Makes a Machine Human”?	132
9.3	The Foundation of Civilization Comes from “Dignity”	136
10	New Rules and New Order in the Era of AI	143
10.1	Creating a Technological Contract Between Humans and Nature	144
10.2	“New Rules” of Human-Computer Interaction	146
10.3	“New Order” for the International Community	151
11	How Can “We” Realize Cogovernance in the Postepidemic Era	159
11.1	Public Goods from a Global Perspective	160
11.2	Governance and Innovation of Intelligent Society	164
11.3	Truth of AI and Reflection of Black Mirror	169

Chapter 1

New Problems in the Era of AI



With the application of a range of technologies, such as AI, big data, gene editing, augmented reality, blockchain, etc., the way we live has changed greatly. Science and technology have brought us a healthier and more comfortable life, liberating our hands to allow us to have more opportunities for cultural exploration and self value fulfillment. In particular, AI and other technologies have played a significant role in the global fight against the pandemic. It can be found that AI techniques such as medical image-aided diagnosis, UAV contactless service, detection and prevention of automation have gradually gained popularity in various scenarios. Meanwhile, China has made it clear that data have become a production factor in the era of a digital-driven economy. How to standardize and promote the use of data has become an essential topic for the development of AI.

The global outbreak we are facing is a fight head-to-head between humans and viruses. Will technology stand in the same boat with human beings when facing such “black swan” incidents? We need to face up to this fundamental problem here—the relation between technology and mankind. Where will technology, one of the most powerful tools of mankind, lead humans in the future? If technology is compared to a medicine, then what are the side effects of this medicine?

Although technology in itself is neither good nor bad, its feature may change when mixed with complicated human nature. A good tool may sometimes turn into dangerous “black technology”. Criteria for good and bad rooted in social cultures can greatly influence the development of technology. Each technology today is marked by social cultures of humankind. Therefore, the analyses of technical ethics must be placed in the social and cultural background. In other words, technology itself does not involve any qualitative ethical judgment, but it creates specific conditions and space for the corresponding ethical judgments. In this way, human beings can discuss technology in a certain social culture. It cannot be simply explained by “technology neutrality theory”, but it indicates that the image of human nature being reflected on the mirror of technology can directly affect the development of technology and the corresponding ethical choices.

Let us first look back on the background of technical ethics. For a long time, technology did not show its importance in philosophy and ethics until the period of industrial revolution. As technological modernization pushed human society into modernity, the significance of technological ethics is remarkably rising. The word “modern” did not exist until the 16th century. It means “now or present” in English. Moreover, “modern times” represent a new conception of history. History repeats itself again and again before modern society. After the industrial revolution and the Enlightenment, people began to shift their attention to the “current era”, and then time became a linear and directional vector. This is the core background of “technical ethics”. By then, people start to deny and break with the tradition, turn their eyes to the present and future, and further reinforce human creativity and subjectivity. This indicates that human civilization has moved into a modernization process. The process has also brought the problems of modernity at the same time.

The author will explore the ultimate problems brought by technology from the perspective of technical ethics in this chapter. These problems include but are not limited to the following: should we use gene editing technology to change the method of reproduction and evolution? Can cyborg help sustain the continuation of our species? Will AI control humans? How do we formulate data rules in the era of digintelligence... It is worth pointing out that we prefer to take a further look at these future technologies with certain “post-human” attributes, especially gene technology and AI technology, from the angle of digital economics rather than traditional economics.

1.1 What Makes a Human Being a Human Being? AI and Gene Editing

On November 26, 2018, He Jiankui, a former associate professor at the Southern University of Science and Technology, announced at a conference that a pair of gene editing twins was born in China in November. One of the twins’ genes was modified to prevent HIV infection. If that worked, they would become the world’s first gene editing case that is immune to HIV. He Jiankui’s announcement drew widespread condemnation in the global scientific community. Many world gene research pioneers criticized that his research had departed medical ethics. Meanwhile, China’s medical and scientific research regulatory authorities immediately filed the case for investigation. This “farce” resulted in the birth of a baby with a modified gene for the first time in the history of the world. The case is also commonly believed immoral by the scientific community and seen as a trample on human dignity and scientific spirit.

Technically, CRISPR gene editing adopted by He Jiankui received a great deal of attention, and the case started an ethical debate on human germline manipulation. However, the more central issue is that before CRISPR research can be safely transformed into treatment, scientists need better methods to avoid the potentially

destructive target effect of this technology. In other words, the problem lies in the fact that DNA will either repair itself or mutate during the process after the CRISPR-Cas9 editing tool cuts double-stranded DNA.

Gene editing technology has completely broken the restrictions of traditional medical treatment. If not being used properly, the technology can completely subvert the existing landscape of human society in an uncontrollable way and may even turn into the source of human destruction. This is why the event has attracted much attention worldwide. We should be cautious about the development and application of gene editing technology. However, it is undeniable that compared with the strong influence of gene editing technology, traditional medical treatment actually provides limited help to human beings.

In essence, although the technology of traditional medicine is still advancing, the fact that the essence of “traditional medicine is just a kind of support for the self-healing of life” can never be changed. Thomas Szasz, an American doctor, once said, people mistook God’s power for medical treatment when the influence of religion proved to be greater than that of science; people are pinning all their hope on medical treatment when the opposite happens.” Indeed, a tacit understanding shared by all doctors is that it is the patient himself who truly cures his disease, and all medical treatment just plays a supporting role. That is, the self-healing ability of life is what truly counts. Medical support aims to win time and create conditions for self-healing, waiting for self-rehabilitation to finally play its role and triumph over illness.

Today’s gene editing technology has changed the behavioral logic of traditional medicine and directly explored the means to improve human existence. Genetic manipulation of human embryos in test tubes may not only help prevent genetic diseases but also change heights, intelligence and other characteristics of humans. In the wave of gene editing technology, there is a strong chance that “super eugenics” will emerge in the future. Cross-use of gene manipulation, high-quality genes, cloning and some traditional methods will thoroughly subvert our understanding of “medical treatment” and directly alter the future of human society. Moreover, enhancing the sensitivity of human perceptual intuition and expanding the scope of human motion control through a brain-computer interface is also an approach to improve human beings.

The way of changing the nature of medical treatment by manipulating genes to “strengthen human conditions” has been ethically opposed. Because a large number of scholars believe that such a practice makes human beings lack “naturalness” like handicrafts, which is detrimental to human dignity. However, this excuse seems to be increasingly challenged, especially when we face unpredictable disasters and unhelpful modern medical procedures. At that time, we often turn to such technical solutions. The contradiction here lies in the fact that we choose to advocate human dignity in the face of various events. However, medical defects, diseases and the promotion of genetic ability are essentially the same thing, but it is difficult for human beings to treat them equally in terms of ethics and emotions. This means that the combination of AI and gene editing technology challenges the fundamental ethical proposition that “what makes a human being a human being?”.

From the perspectives of economics and sociology, how to strengthen humans is a great challenge for us. This is principally because the technology may serve those affluent groups and families, resulting in permanent social inequality, namely, the expansion of “free eugenics”. The rich who occupy a large amount of resources can modify their offspring by editing genes so that their offspring can gain an overwhelming advantage over the average person in intelligence, appearance, height and even life expectancy, thus forming a genetic aristocracy composed of wealthy people. By doing so, they can monopolize all resources. By then, compared with genetic aristocrats, ordinary people without any strength will be eliminated to the bottom of society and may even become slaves or die out, leaving hierarchical solidification worse and worse.

In some extreme cases, if it were possible to judge genes subjectively, the tragedy of Nazi genocide might be repeated. As described by the British writer Aldous Huxley, in his novel, “*Brave New World*”, published in 1932, all infants are incubated in laboratories instead of through natural viviparity. The fates and conditions of these babies are directly predetermined before they are incubated. They are classified into five “castes” or social strata, including alpha (α), beta (β), gamma (γ), delta (δ), and epsilon (ϵ). The target of all conditions set is to make people content with their inescapable social destinies. Five “castes” are cultivated separately. Among them, alpha and beta, the most upper class, lead and control other classes; gamma, the ordinary class, is equivalent to civilians; delta and epsilon, the lowest class with low intelligence, can only does ordinary physical labor... If the situation continues going like this, the long-established social ethical relationship will be overturned, let alone social equity. Ultimately, this will result in the disintegration of human society.

In addition, from the perspectives of the technical level and human driving force, these strengthening technologies will bring us to the “post-human era”. The term “posthumanism” appeared in contemporary social science in the mid-1990s, but its source can be traced back to at least the 1960s. The philosopher Foucault also mentioned in *The Order of Things* that “man is a recent invention within it... Man would be erased, like a face drawn in sand at the edge of the sea and is approaching its end.” The sentence can be summarized as his declaration of the “death of man”. In all the ancient myths of mankind, the difference between man and God is that man is immortal, and God is immortal. When man breaks through his biological limitations and obtains eternal life, man becomes God.

Posthumanism calls the new human transformed by technology posthuman and reflects on it from two research approaches. One is the vision of human perfection driven by new technologies based on the summary and outlook of the development and utilization of technology. The other is the inquiries into posthuman and its era from the positions of reflection and criticism. Although these two approaches have a general sequence, they both focus on the boundary and changes in the form of “human” and the corresponding multidimensional thinking in the context of the high development of technology.

In this sense, posthumanism can be regarded as the deconstruction of humanism, and it dispels the particularity of humans in nature from the level of body and species. There are three evolutionary models for posthuman: genetic engineering or asexual

reproduction (such as cloning technology); technical cultivation or artificial cultivation; the third way is to create virtual subject and transform real subject with virtual technology, combining virtual world with real world, virtual person with real person, which are prominently presented by a man-machine combination-cyborg.

In *The Letter from Utopia*, the famous philosopher Nick Bostrom strongly advocates the benefits of the posthuman era and summarizes three characteristics of posthuman states:

First, most people can fully control their sensory experience by virtual technologies such as brain computer interfaces. There will be no difference between simulated life and real life.

Second, the majority of people no longer need to suffer from psychological pain, such as depression, fear and self-loathing. The problem here is that we cannot determine whether there is a natural link between these painful experiences and personal achievement and self-esteem and whether we can embrace the benefits of “man-made” happiness.

Third, most people will have a much longer life expectancy than their natural life span. The age of longevity not only results in aging, delayed retirement and labor shortages but also makes tediousness and irresponsibility common.

The Letter from Utopia depicts the happy side posthuman brings to us and awakens exciting scenes. People can rethink the meaning of human beings in new ways, being away from old constraints. However, from a dialectical point of view, the “evolution” of mankind needs to be transformed from the inside to obtain a new cultural form. Otherwise, the individual will be eliminated by times. Especially after the increasing technicalization of human beings, the debates on whether “technology can replace human beings” will break out. The former will cause disembodied posthumanism, and the latter will embody posthumanism.

Let us return to real cases. In the past few years, driven by the power of capital, some commercial hospitals have been researching gene editing technology. Scholars such as He Jiankui are conducting gene experiments regardless of ethical risks. The driving force behind this is the game between capital and power. On the one hand, biological gene knowledge is capital, and the capitalization of knowledge brings power. Many overseas hospitals make profits through the commercialization of knowledge and pay universities in the form of patents. In other words, as the life gene codes of life bodies are all being transformed into income, human beings are slowly losing their dominant right to control their own genes. Actually, capital promotes the knowledge system of production and accelerates the deviation of the human central position, contributing to the distortion of ethics and values. On the other hand, the technological ethics system is taking shape. The original opposition system between nature and man will gradually evolve into a technological ethics system dominated by secular culture and enlightenment.

Stephen Hawking, a famous scientist, expressed such concerns in his book *Brief Answers to the Big Questions* that although laws can prohibit human beings from editing genes, humans cannot resist temptation.” How to get rid of the shackles of capital to speed up the construction of civil society has become an urgent problem

to be solved. People have great expectations for technology. As Dennis Gabor, the winner of the Nobel Prize for Physics in 1971, said that all that can be realized at the technical level deserves to be fulfilled at any moral cost. Under the joint influence of technology and markets, people begin to go astray further and further.

Many countries and business organizations have set out to establish all sorts of ethics committees to examine the acceptability of scientific and technological achievements to guarantee that the application of products meets ethical requirements. However, the root problem lies in the fact that “the essence of markets is breaking all shackles ahead with innovation”, and the development of ethics always lags behind that of science and technology. What’s more, human beings are becoming a creator by getting power from nature by virtue of the advancement of experimental science. However, they forget to set limits for the power. In *Humankind 2.0: Changed Bible*, American futurologist Ray Kurzweil discusses “technological singularity”, which combines biotechnology, robotics and AI, and he also expresses aggressive future individualism through the study of efficiency and technology leap behavior.

In regard to the thinking of human existence, in the 1970s, Michel Foucault, a French philosopher, refuted that the humanities we understand cannot be equated to the universal propositions of humanism. In fact, it is constructed by a set of clear assumptions about “human”, which are limited by history and context. Human beings, an integration of life, labor force and language, are a continuously developing “double structure of empiricism and transcendentalism”.

In fact, in such an era of rapid technological development, we have to reexplore this issue from the angle of the relationships between technology, society, ethics and economy and pay more attention to social and ethical values other than economic benefits. Both values also have a decisive influence on the business itself, and this is exactly what is going on in the real business world. That is, technology and capital have changed the world, but ethics will change the trend of business and capital.

Expanding gene editing technology from the scope of “treatment” to “prevention” will blur its boundary, while “prevention” and “improvement” are just a step apart. When a family expects its offspring to have certain characteristics (high appearance, good health and high IQ), a huge crisis is looming. If ludicrous He has started a piece of history, the peril of it lies in that it is likely to bring an end to human civilization, including destroying the diversity of the human gene pool, causing eternal inequality, etc. This is also the choice and challenge that human civilization is facing. The question “What makes a human being a human being?” not only redefines the development history of our technological civilization but also defines the history of the whole “anthropocene”. How to make choices will lead us to completely different evolutionary paths and civilization processes. Since it is a major historical mission of our generation, we should treat it with a particularly cautious attitude to avoid enormous risks and unpredictable future.

1.2 Man-Machine Symbiosis: Challenges and Problems in the Near Future

During the process of the development of AI in the past few years, all positive and negative problems involved are related to the same proposition: man-machine symbiosis.

It can be seen that views of the whole society on AI have been divided into two camps. One camp assumes that AI has brought new technological dividends to social development, energizing multiple industries. The possibilities of human society and the human future will eventually be reshaped through the development of AI technology, and we are becoming a semiorganic and semimechanized “cyborg”. The other believes that AI technology has brought about immense ethical risks, which will impact the basic order and ethical bottom of society, leaving the person in charge of ethical dilemmas. Even in the long run, human beings may be thoroughly surpassed, backfired and replaced by AI.

In reality, unlike the mainstream deep learning methodology, the core philosophy of AI is to assume that intelligent systems can operate under constrained resource conditions. Based on this cognition, this article discusses the real dilemma between AI ethics and the philosophy of technology.

In terms of the research stage of AI ethics, the current research on AI ethics by scholars worldwide has gone through three stages. The first phase is about its necessity. Relevant research was mainly initiated in the United States. Due to American leadership in AI technology, enterprises represented by Google encountered ethical problems such as AI militarization in the early stage, which attracted the attention of industry and academia.

The second phase is the discussion of AI ethics. During this stage, both the EU and China were actively involved. For example, in April 2019, the European Commission issued a set of AI ethical principles that includes ensuring human initiative and supervision, technological robustness and security and transparency, strengthening privacy, data management and accountability, maintaining diversity, non discrimination and fairness in the application of AI systems and increasing social well-being. To date, dozens of research institutions or organizations have put forward their own ethical standards and suggestions on AI. Generally, these principles have certain universality and internal consistency and have reached a certain consensus on the establishment of ethics.

The third phase is where we are now. During this period, the studies focus on the AI ethics system and its specific connotations and application measures. Through the system of “ethical mission—ethical standards—rules for implementation”, we can solve the two problems that cannot be solved at the principle level. One is “the self-execution of AI ethics”, that is, how to implement the principles via cooperative operation mechanisms. The other is the “risk control of AI ethics”, which plans to reduce its application risk through forward-looking deployments. In short, the AI system planning stage is the process of implementing AI ethics from theory to

practice. Only in this way can we standardize the development of AI technology and constantly improve its technology evolution path.

Technologically, how does the AI ethics system adapt to the dynamic evolution of AI technology? Before answering this question, we need to determine the technical essence of AI. AI technology has different paradigms, including logical intelligence (propositional logic and first-order predicate logic), probabilistic intelligence (Bayes theorem and Bayesian network), computational intelligence (genetic algorithm and evolutionary computation), neural intelligence (machine learning and deep learning) and quantum intelligence (quantum computation and quantum machine learning).

Generally, **AI can be regarded as a technology for humans to interact with the real world through machines with the help of logical reasoning and perceptual learning on the basis of big data.** In other words, the underlying architecture that AI logic algorithms can execute is massive data. The more data resources an AI company possesses, the more competitive it will be. A professor of AI and ethics at Stanford University, Jerry Kaplan, the author of *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence*, holds that a top AI company often relies on a large amount of data. Strong ones will become increasingly stronger because they can form a virtuous circle by virtue of data accumulation, iteration and automatic annotation. Machines are able to summarize laws and specific knowledge from a particular large amount of data and then apply such knowledge to real scenes to solve practical problems. That's how computer practices logical reasoning.

Since the establishment of the concept of AI in 1956, it was initially dominated by the logic school. This is primarily because logical reasoning and heuristic search avoided the deep-seated complex problems in the brain's law of thought in intelligent simulation and made major breakthroughs in key fields such as theorem proving, problem solving, pattern recognition, etc. In the early days, scientists generally believed that the essential difference between AI and traditional computer programs lies in logical reasoning. This way of thinking puts aside the microstructure of the brain and the evolution process of intelligence and only uses the process of problem solving through program or science of logic to simulate human thinking logic, so it is also classified as weak AI.

Looking back on the theory of Descartes, an important philosopher in modern theory of knowledge, real intelligence will be embodied as a "universal problem-solving ability" rather than a post synthesis of specific problem-solving ability. The fundamental feature of this universal ability is that it has the plasticity of constantly changing itself in the face of different problem contexts and has strong learning and updating abilities.

The author believes that universal problem-solving ability is a necessary and insufficient condition for machine perception. Only when the machine can not only outperform human beings on a specific problem but can it build a self-logic and learning system to give outside feedback on all kinds of universal problems can it realize machine awakening from consciousness. Correspondingly, Kant put forward the theory of mind integrating empiricism and rationalism in his book *The Critique of Pure Reason*. He divides the perceptual activities of the mind into two parts: one is perceptual ability, whose task is to pick out the original inputs of sensory information;

the other is intellectual ability, whose task is to organize those inputs into a coherent and meaningful universal experience. Kant devoted most of his energy to intellectual ability, worked out a fine model of high-order cognition, and divided intellectual ability into twelve categories with this model.

Once a machine gains intellectual ability, it obtains the ability to perceive and interact with the world, which can also be called autonomous consciousness. Autonomous awareness allows machines to deal with complex systems through self-perception and learning without preset programs. From the concept of “cognitive computing” put forward by the American psychologist and computer scientist J. C. R. Licklider, it is possible to make the computer systematically think and propose solutions to problems and realize man-machine collaboration in decision-making and control of complex situations without depending on preset programs.

We can also gradually generate the subjective impression of machine perception from a series of science fiction movies: from the German film *Metropolis* in 1927, *2001: A Space Odyssey* in 1968, to *Big Hero 6*, *Her* and *Ex Machina* in recent years. The audience generally believes that strong AI will bring us a conscious, humanoid robot whose intelligence is equal to or even better than that of human beings.

Although AI technology is not yet mature at the moment, which means we are still in the “era of weak AI”, the ethical problems in AI application in real life have become quite serious. With the increasing popularity of self-driving cars, especially the occurrence of some unmanned traffic accidents, the “trolley problem” has become an issue that must be considered to ensure the safety of unmanned driving and even for the ethics of AI.

MIT launched an online test project called Moral Machine to collect and sort out public moral decision-making data. The data of 40 million moral decisions made by millions of users from 233 countries and regions reflect some global preferences: saving humans rather than animals, saving the majority at the expense of the minority, and giving priority to children’s lives. However, due to the heterogeneity of geographical and cultural factors, people in different regions still have different tendencies in the choices to face the same problem. In 2018, Germany set the first ethical rule for self-driving automobiles. The rule mentions that the system must give the highest priority to human safety compared to the damage to animals or property. If the accident is unavoidable, any discrimination based on age, sex, race, physical characteristics or other differences is prohibited. For morally ambiguous events, human beings must regain control. The way of controlling the choices and behaviors of machines by implanting preset moral algorithms into machines is top-down.

In terms of solution, the designer (human) must first reach social consistency in ethical theory, analyze the information and overall program requirements necessary to implement the theory in the computer system, and then design and execute the subsystem under ethical theory. Although this top-down design approach can guarantee relative fairness on the basis of the veil of ignorance, the preset algorithms often get caught up in paradoxes in specific circumstances.

In contrast, “bottom-up” is a new thought that allows machines to derive their own moral standards through the iteration of daily rules. Machines with perceptual learning ability can summarize external information and form system behavior

patterns under different situations. In the meantime, designers can encourage machines to take certain actions by establishing a reward system. Such a feedback mechanism can encourage the machine to develop ethical norms of its own in time. The process is similar to the learning experience of moral characters in childhood. By doing so, AI can truly become the subject of artificial morality and guarantee the legitimacy of its behaviors.

To explore whether machines can become moral subjects, we must think about the relationship between human beings and machines. The technology of AI has broken through the boundary between human and nonhuman entities since the enlightenment. With the growth of AI, the relationship structures between man, technology and the world have changed. Man and technology began to merge, such as the cyborg relationship and compound relationship proposed by the post phenomenological technology philosopher Verbeek. When the subject status of machine is independent of human beings, can it become a more humane responsibility subject?

Take the intelligent UAV as an example. Since the U.S. Department of Defense announced the establishment of JAIC in June 2018, and the United States has continuously accelerated the pace of AI militarization application. At the beginning of 2020, the US troops attacked and killed Qasem Soleimani, a senior Iranian general, leaving the situation in the Middle East suddenly flaring up. Media reports said the task was carried out by UAV “Reaper”. With the deepening of AI technology in the field of military application, UAVs represented by “Reaper” have possessed the features of intelligence, thus triggering a new debate on the responsibility subject of war—that is, can highly intelligent UAVs better bear war responsibilities than humans as the only responsibility subject in traditional wars, leading the future war in a more humane direction?

Ronald C. Arkin of the Georgia Institute of Technology pointed out that compared with manned combat platforms, intelligent robots have the following six advantages in ensuring the justice of engagement: 1. they do not need to consider their own safety; 2. they have superhuman battlefield observation ability; 3. they are free from the influence of subjective emotions; 4. they are free from the impact of customary models; and 5. Faster information processing speed; 6. independent and objective monitoring of moral behaviors on battlefields.

Based on the above advantages, Arkin believes that the intelligent unmanned combat platform will outperform humans in the execution of humanitarian principles. However, as the current UAV autonomous system is still unstable and risky, accidents including control system failure, electronic signal interference, hacker network attack and other unforeseen circumstances on battlefields will affect its implementation of decisions in line with humanitarian regulations and even cause the killing machine on battlefields to get out of control. In addition, the dilemma of responsibility distribution caused by UAVs is also reflected in how to tackle “responsibility transfer”. Backed by the high human-machine integration, the military and the government can transfer man-made responsibilities to UAVs to avoid war crimes.

The emergence of unmanned warfare will inevitably lead to profound changes in some traditional war ethics, which need to be taken seriously. Currently, some countries have proposed formulating international laws and ethical norms for military

unmanned aerial systems (UASs) of increasingly higher intelligence to restrict their battlefield behaviors. On May 27, 2013, the regular meeting of the United Nations Human Rights Council also pointed out that robots are likely to accidentally kill enemies who are preparing to surrender if they are sophisticated enough to be able to judge enemy situations automatically instead of just being controlled remotely. This reminds us that we should not only use war ethics to safeguard our own interests but also change war ethics to provide legal guarantees for the application of unmanned systems or develop unmanned systems in accordance with war ethics. For example, unmanned systems should be modified to automatically identify and aim at the weapons used by the enemy, making them ineffective or destroyed, to remove the threat to their own side without killing relevant personnel. In this way, people's concerns about potential "robot killers" can be eased.

Making machines have "mind" is the pursuit of human beings, which also raises concerns about human security. This actually reflects the fear of the uncertainties of AI autonomous evolution that humans face. As Hawking said, "used as a toolkit, AI can augment our existing intelligence to open up advances in area of science and society. However, it will also bring dangers... The concern is that AI would take off on its own and redesign itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete and would be superseded. In addition, in the future AI could develop a will of its own, a will that is in conflict with ours".

From the technical path, futurologists are worried that super AI is capable of intelligent evolution and self-replication, and then it can surpass human beings in thought, that is to say, it will achieve the technical singularity. **To understand whether AI can achieve this so-called technical singularity, we need to make it clear that the essence of intelligence contains the basic elements required by the sensory ability mentioned above.**

Let us return to the origin of the philosophy of technology. Only by deducing the future of technological development through the logic of epistemology and scientific principles of philosophy and discussing comprehensive knowledge in multiple academic fields, such as cognitive science, mental research and language modeling, can we eliminate the discussion of utopian theory or crisis theory, which are meaningless to the development of AI, and return to the path of realism to construct ethical and technical routes that are conducive to the future development of AI.

The arrival of AI makes us consider the path of human evolution. Without considering the extreme situation of human extinction in science fiction movies, there are two evolution directions for the immediate future: one is the formation of new species. A new human race can evolve on the earth or on other planets; that is, new species can be formed through directed evolution (in fact, "gene editing" is also the technical practice of such a directed evolutionary path). The other is man-machine symbiosis combining machines with the human brain to create new symbiotic species. Human beings are getting more and more "cyborg". This is an important development direction of man-machine combinations and a question that needs to be answered in the "near future" we have to face.

1.3 Data Ethics: Political Economy in the New Oil Age

Machine learning and big data are promoting the transfer of technological and commercial power throughout the world. It can be seen that companies with the highest market value in 2001 are all energy enterprises, such as GM and Exxon Mobil. However, by 2020, the most valuable companies have become those closely related to big data. It is no wonder that people say that data are oil and data are money at the moment.

The fourth industrial revolution takes big data as the core. All subsequent technological changes, such as the Internet of Things, AI and blockchain, can only be driven by big data. With the development of big data, the potential value of big data information has been continuously developed.

Data have become the “oil” of the data economy era. Countries with big data capabilities have successively implemented big data development strategies, promoted the reform of methods of production and information exchange, and expected to improve the quality of economic growth through data value. However, what cannot be ignored is that ethical issues about big data, such as data monopoly, data privacy and data information security, have also drawn great attention as the value of big data is continuously developed and verified. How to standardize and promote the use of data has become an important topic in the development of AI.

In reality, the development of information value depends on the collection of large-scale original data. The Internet, mobile communication, e-commerce, social platforms and government departments are all collecting massive data. However, what kind of personal data can be available and how to avoid data abuse are difficult to discriminate in specific practice. From the perspective of economics, this section will establish a thinking framework for readers by discussing the institutional construction and philosophical problems of data ethics.

The sharing of data and information is in an increasingly balanced stage. In the big data environment, information sharing and flow are the premise for the information value of big data to be available to the public. Without information sharing, the so-called “information island” is formed, and the value of information cannot be fully developed. At the same time, information abuse will lead to the disordered development of data, causing corresponding data ethics disputes.

Let us first look at the problems of data monopoly and privacy protection. Taking the notorious “Facebook Cambridge Analytica” as an example, Facebook was found to have manipulated American presidential election by its interest groups through the data on social media platforms, which not only sparked widespread outrage but also made us realize that we often underestimate the shaping power of big data to society. Facebook’s database holds a large volume of Internet users’ information, and social media giants are imperceptibly exerting influences on people’s decisions in a monopolistic posture. Human beings live in a virtual and digital activity space in which they use digital technology to engage in information transmission, communication, learning, work and other activities. Each individual’s words and deeds may inadvertently leave some “traces” and become an object that can be recorded and

analyzed. In the era of big data, the quantified “traces” do not exist in isolation but are inextricably linked. Monopolists can produce a new set of power relations through correlation coupling and then have a profound impact on the increasingly digitalized society.

In essence, what we have learned from the Facebook case is the “duality of information sharing”: the free boundary of information sharing and the value expansion of information islands. This contradiction is almost an endogenous problem of data value. Facebook provides nearly unlimited information sharing for all users, but it also brings problems such as privacy and data monopoly.

From the perspective of the logic of digital economics, the privacy and monopoly of big data is the constraint mechanism of science and technology ethics in the era of information sharing. The basic logic behind it is the boundary of information sharing and the fair distribution of information value. We should think about how to establish a constraint mechanism for data ethics to avoid corresponding risks while ensuring that the value of big data information is tapped.

Next, before we discuss the formulation of data protection rules according to the relationship between system and ethics, we need to look at Google’s privacy infringement case. In June 2017, Google blocked its competitors’ shopping comparison websites for its own promotion in search results. The move violated Article 102 of the *Treaty on the Functioning of the European Union (TFEU)* for the abuse of its market dominant position, and Google was fined a huge amount by the European Commission. In this case, relevant institutions creatively put forward the concept of “the right to be forgotten”, which is used to indicate that people have the right to require relevant service providers to delete personal information such as data traces left on the Internet in the era of the digital economy.

From the viewpoint of privacy protection, the case of Google data privacy provides a legal framework for data protection in three aspects: anyone has the basic rights and freedom to access and process personal data, the controller of personal data must bear the legal obligations and responsibilities of personal information, and all EU members must establish special data protection institutions. In fact, it can be seen that the verdict returned by the European Union Court of Justice over the case of *Google Spain SL and Google Inc. v AEPD* not only achieved the above three objectives but also proposed relevant opinions on the innovation of adopting extensive interpretation in the provisions on the protection of personal data privacy in the *EU Data Protection Directive*. The *General Data Protection Regulation (GDPR)* launched by the EU after 2018 is obviously affected by the relevant cases, prompting people to focus on the freedom of information sharing and ethical limits of data space as public fields.

Facing the challenges related to data ethics, governments have formulated many laws and regulations relating to privacy and data security. The European Union launched the *Directive on Privacy and Electronic Communications, Directive 2002/58/EC* in 2002. The United States protected people’s right to privacy through the relevant precedents of *The Fourth Amendments to the Constitution of the United States* and *The Fourteenth Amendment to the United States Constitution*. The U.S. also issued the *Consumer Data Privacy in a Networked World* in 2012, the *Protecting the Privacy of Customers of Broadband and other Telecommunications Services* in

2016 and some other regulations. Moreover, many international cooperation organizations are also carrying out relevant institutional construction, such as the European Federation of Data Protection Organizations (CEPDO), the International Association of Privacy Professionals (IAPP) and other institutions.

It is worth mentioning that the *Research Report on Data and Competition Policies* issued by the Japan Fair Trade Commission (JFTC) in June 2017 organized and introduced many issues related to data, such as the changes in the use environment and status of data, the impact assessment methods of mobile phones and data use on competition, data collection and use behavior. The document focuses on a large number of frontier issues of the data system (such as privacy considerations in the framework of competition law, data material blockade, etc.). Moreover, Japan's report highlights the aspects of data collection and use and divides data into personal data, industrial data and public data. In particular, the research on the latter two types of data was very cutting-edge and detailed.

For personal data, the report proposes that the concepts of data and information are convergent (the cross perspective of data ethics and information ethics) and mainly discusses relevant behaviors in the social network market. In terms of industrial data, the report emphasizes the concept of data hoarding and points out that monopoly or oligopoly enterprises may restrict data access or data collecting channels to achieve data hoarding. Regarding public data, the research focuses on how to maximize the value application of government institutions and public data. By this means, the report analyzes the necessity of defining the data trading markets and puts forward crucial perspectives on the M&A and a review of the data.

Although these directives or regulations escort data ethics to a certain extent, from the angle of institutional economics, the crises of data supervision still exist. One of the reasons is that the modern anti-monopoly law based on the theory of Chicago School mainly focuses on three types of behaviors: mere horizontal fixed prices and market segmentation monopoly agreements, bilateral monopoly and horizontal merger of monopoly, and limited exclusive behaviors. Being affected by the above phenomenon, regulators tend to ignore the maintenance of data competition in cross industry mergers, and multiple data-driven mergers and acquisitions have not been included in the concentrated reviews of operators, fully reflecting the crisis of dealing with problems relevant to data monopoly on the multisided markets under the mind-set of modern anti-monopoly law.

Moreover, the antitrust investigation on the abuse of market dominant position by platforms turns out to be a kind of lagging and passive post review. To promote the long-term and healthy development of the data industry, regulators should not simply ask enterprises to accept monetary compensation or provide data beyond the necessary scope on the grounds of breaking the enterprise's "data monopoly". Therefore, in addition to strengthening the prior review of cross-industry data M&As and integration on platforms, optimizing the data sharing mechanisms is also the key to promoting the development of big data.

In the field of data ethics, the traditional economic mode mostly represented by physical resources takes land, natural resources, population and capital as production factors, while key information and value elements in the digital economy are

generally reflected in the production, storage, circulation and application of data resources. The form of virtual resources not only expands the application width and depth of factor resources but also combines with the traditional production factors in the real economy to form new economic paradigms such as AI, robots, blockchain, digital finance and so on, further energizing traditional products and services of the real economy in quality, efficiency and benefits to double the benefits of the real economy. Data resources have become the most important production factor in the development of the digital economy, but the data come from every individual participating in production activities in the real world. Under the infiltration of various technologies and forces, some ethical problems have been completely unmasked, mainly in the following aspects:

- Information security issues: the data industry chain is interlocking and intricate, and the risk of uncertainty in the acquisition terminal, processing node, storage medium and transmission path of data analyses is high.
- Personal rights and interests issues: The abuse of data leads to privacy overdraft, and personal dignity is more likely to be degraded.
- Business ethics issues: AI products may induce users to actively overdraw personal data, and human beings may lose control over the boundaries of privacy and sensitive data.
- The principle of informed consent is becoming dispensable in the context of ubiquitous sensors and potential data development value.
- Intellectual property issues: big data is aggravating the contradiction between intellectual property and free data sharing on the Internet.

We hope that members of society can equally obtain data resources according to their needs and acquire corresponding wealth distribution based on their contributions to production activities. However, in reality, due to the contradiction between data-driven productivity and productive relations, individual users and platform enterprises often oppose each other in the application of data resource distribution, and there also exist contradictions such as data monopoly among enterprises. Specifically, it can be deconstructed into the following four aspects:

First, “who owns the data?”—the ownership of data property has always been the focus of debates in the industry, especially for those data transactions that remove the attributes of personal identities, as well as the data produced by individuals who are collected and stored after desensitization treatment by enterprises or government departments. There is still no consensus on whether data ownership should belong to individuals, enterprises or the government.

Second, “who is using the data?”—government and enterprise are the two principal users of large-scale data in the current digital economy era. The government collects a large amount of data through various systems, including public service websites, digital government platforms and the “Integrated Online Platform”, while enterprises gather users’ information by providing services for users and obtaining multidimensional trends and regular features through data analyses to precisely improve and optimize their services and user experiences. For this reason, the convenience, nonrivalness and low-cost replication of data storage and transmission

reflected in the above application process also make it difficult to protect data property rights. Data can never get rid of being occupied, stolen and abused illegally even with clear property rights. Moreover, as technologies diffuse, people come to realize the value characteristics of data. Data theft technologies such as web crawlers and credential stuffing attacks are rampant. Whether public privacy, government governance or national security is more vulnerable to the threats of improper use, such as privacy and data theft and abuse, than ever before, seriously infringing on the property rights of data owners and causing damage to the scarcity of data.

Third, “How much data is available?”—as the source of wealth and value in the digital economy, the data generated by individual consumers are not only the basic source of profits and value of platform enterprises but also the key element for digital public services and digital governance of government. However, currently, the unclear ownership of personal data leads to extreme situations such as abuse or excessive restriction on the use of personal data, which place the public individual interests and the interests of platforms or public organizations in a contradiction of binary opposition.

Fourth, “who should get the data revenue?”—the considerable economic benefits brought by using data to optimize products and services. The distribution between data producers (individuals), collectors and processors (enterprises and government) affects the interests of many subjects. Although the current judicial decisions prefer to distribute the data revenue to the collectors, producer and actual controllers at the link of secondary exploitation of data—enterprises. However, in some public services, especially data scenarios relating to government affairs, can the government as a data user and individual data producers have the right to enjoy legal benefits of data without the support of judicial decisions?

The above is the discussions on science and technology ethics of big data from the points of information ethics, institutional assumptions and philosophy. Whether it be big data, AI or gene editing technology, they are essentially subversive technologies in the digintelligence era, such as “every coin has two sides”. When we look at the ethical issues of science and technology, we should not only pay attention to their technological advantages but also examine their disadvantages through other angles, such as economics and philosophy. Only by doing so can we have the chance to understand their application paradigms in practice as well as their impact on the basic logic of social operation.

Only by this means can we continue our accumulation of positive value in shaping modernity and apply more and better science and technology to the expansion and development of human civilization after entering the “digintelligence risk society”. In the meantime, the risks brought by technical ethics, the conflict between humans and nature, the continuation of the human race and the spiritual feelings of loneliness and wandering caused by modernity should all be controlled at an acceptable level. It is also an important and essential problem that the author believes the digital economy era is facing.

Chapter 2

Ethical Enlightenment in the Age of Intelligence



If the Enlightenment Movement can be seen as the impetus for rational reflection and value reconstruction in specific fields, the trend of thought “tech for social good” driven by China’s science and technology companies in the past few years can be regarded as an obvious mark of China’s “ethical enlightenment” movement in the smart era.

The proposal of the idea of “tech for social good” in this age primarily stems from the ethical challenges we have encountered when promoting the development of the digital economy, such as the repeated big data discrimination on travel websites and the security risks brought by car sharing. We have seen that these technical ethical issues are challenging people’s moral bottom line and their ethical concepts of digital life. How to delimit the red line when technology promotes business development and how to understand the role that the concept of “tech for social good” plays in creating a good social ethics orientation and soft landing mechanism for scientific and technological innovation are the propositions we need to think about.

Given that AI technology is one of the important technologies to solve human poverty and inequality and may create and expand the digital divide in its development, it is very necessary for us to understand technical ethics from the angle of economics—we should fundamentally consider the challenge of ethics to the underlying logic of social development rather than just the impacts of technology on economy and society. Technology enterprises take tech for social good as their mission, which actually determines the new value dimensions of assessment techniques, products and services. The change also means that technical innovation needs to consider not only commercial value but also other aspects, such as ethics and social responsibility. Obviously, the move is not only an enlightenment but also a cognitive upgrade.

The core of AI ethics research is to “put people first”. We should make AI work for human well-being by coordinating human ethical values with those of AI. Furthermore, human beings should first reach some “consensus” on their own ethical values before discussing the development framework and basic governance principles of AI ethics.

Therefore, the focus of this chapter is to discuss these already formed “human ethical consensus” and their corresponding origins and understand these ideas on the basis of modernity. As Kant pointed out, “Nothing is required for public enlightenment, however, except freedom; the freedom to use reason publicly in all matters”.

The first part of this book aims to tell readers that the **ethical problems of AI (including all technical ethical problems) are actually human problems brought by “modernity” and “modernization”**.

It is worth mentioning the “impoverishment” of modern economic theory criticized by Amartya Sen, the Nobel Prize winner of the economy. On the one hand, mainstream neoclassical economics today is analyzed and constructed under an ideal theoretical assumption of zero transaction fees, so it differs greatly from the real economic world of human life. On the other hand, although the new institutional economics theory has introduced variables such as “transaction fee” to explore the institutional arrangement of market operation, “ethics” of systems has never become a consideration, and the moral basis and ethical thinking seem to have never been incorporated into the research framework of economics. Therefore, this is also the ethical difficulty that we face in the digital economy—since the traditional economic perspective has never taken ethics into consideration, how do we construct ethics in the era of the digital economy and AI?

If we do not look at the paths and roles of institutional changes from the perspective of ethics, we cannot understand the reasons for the existence of the “non-Pareto effect” or “non-Nash-Sutcliffe efficiency” in economic operation. It is also impossible for us to understand why most people follow certain ethical norms to coordinate and cooperate with each other in the community.

The real economic world is Hayek’s “extended order of human cooperation”, rather than the “a war of all against all” discussed by Hobbes in *Leviathan*. In his paper “Conjecture on the Beginning of Human History”, Kant mentioned his views on history, human nature and human society, which may help us correctly understand the problems we are facing. He made a guess at the initial development history of human freedom and gave us a poetic interpretation. At first, human beings only had animal instincts. He had to rely on the guidance of instincts and put himself at the mercy of God. However, gradually driven by reason, he expanded his range of food. Then, the reason soon made him notice its existence. Human beings began to seek to transcend the boundaries of instincts and expand their knowledge. This marks that reason began to control the consciousness of impulse. After that, mankind began to learn to predict the future, that is, how to prepare for long-term goals, which is the decisive factor of human strength. Finally, human beings realized that human is the real purpose of nature, so they take whatever they can as tools and means with the exception of human beings themselves. Thus, they found a way to change the evolution of nature. To put it another way, historically, behind the human evil lies the purpose of good”. Therefore, when we look at the ethical problems in the intelligent era, we also need to adhere to this human-centered principle. To realize self enlightenment, we need to have the courage to use our own rationality to confront unknown problems.

2.1 Development of Human Consciousness and Machine Consciousness

Before discussing the ethics of AI, we must first discuss the consciousness of AI from a philosophical perspective. Yuval Noah Harari, the author of *Sapiens: A Brief History of Humankind*, once said, “intelligence and consciousness are very different things. Intelligence is the ability to solve problems, while consciousness is the ability to feel things such as pain, joy, love, and anger”. Obviously, ethics is a proposition related to consciousness, so whether there is an essential difference between machines and humans determines the significance of ethical problems.

In this chapter, we discuss the philosophical significance of human consciousness and machine consciousness from the perspective of Eastern and Western philosophy and the necessity of ethical issues from the uniqueness of human nature. Finally, we try to understand the development of AI from generalized consciousness and look at the impacts of such a technological paradigm on human civilization.

First, let us discuss the relationship between intelligence and consciousness. We can compare the differences between them from philosophical concepts. “Intelligence” tends to treat and deal with problems in a purposeful, objective and effective way in hopes of obtaining answers to questions. “Consciousness” refers to the subject’s ability to feel pain and emotions. It is the ability of an individual to construct meaning after being touched physically and mentally. Therefore, consciousness represents a certain meaning that can be embodied in certain objects. For example, the phenomenon of the “phantom limb” in medicine means that individual consciousness can be given meaning subjectively whether there is an object or not.

It can be seen that philosophers basically follow two paths when discussing human nature: “intelligence” and “consciousness”. It is precisely because different civilizations and societies have different understandings of intelligence and consciousness that the East and the West have different understanding directions of contemporary technological philosophy, showing the evolution paths of philosophy of AI from different perspectives.

By comparing the relevant contents of Chinese and Western philosophy, we find that the West pays more attention to the philosophical significance of “intelligence”. People answer through intelligence when discussing the proposition of “why human beings are human beings”. Therefore, the main goal of “robot ethics” is to create a robot that can act in accordance with ideal moral principles and ethical laws. Thus, ethical dimensions need to be added to robot intelligence.

- (1) In the design of robots, ethical norms and moral laws can be embedded into robots according to human conceptions so that they can carry out corresponding behaviors according to the moral ways set by engineers.
- (2) The ideal moral principles of the robot should be set up, including the corresponding moral dilemma, cases of correct feedback and the algorithm of abstract ethical principles for each approval project, so that the robot can guide its behaviors according to the corresponding examples.

In contrast with western “intelligence”, oriental philosophy mainly discusses ethics and human nature from the level of consciousness, and one typical example is Chinese Confucian moral philosophy.

The ethical representatives of Confucianism include “Mencius” and “Hsun Tzu”. Mencius believes that what makes humans unique is that all men have a mind that cannot bear to see the sufferings of others. Cultivation of kindheartedness can make humans have real human nature, promoting human society to a moral life. Hsun Tzu believes that people become moral by acquiring social norms; that is, they can act morally through continuous social practice, which means they can obtain ethics through social procedures.

Therefore, both Mencius and Hsun Tzu emphasize that human moral consciousness is crucial. From this angle, Hsun Tzu’s idea is more suited to the development of contemporary AI. He believes that the keys to human consciousness are “righteousness” and “group”. The former defines the appropriate consciousness of morality of the human theme, and the latter defines the human ability to form groups, and both emphasize the importance of ethical rules in the social system.

We can see that compared with Western philosophy, Confucian philosophy can better deal with ethical challenges in the field of AI. For example, Kant’s ethics takes rationality as the premise, and he believes everyone possesses rationality and can practice it, while Confucian school discusses how different people promote ethical practice according to their actual situations under the assumption that everyone can be a gentleman. In other words, Confucianism leaves more room to embrace AI as a moral subject, and AI needs not to be treated as fully equivalent as humans.

Then, let us discuss the similarities and differences between AI and human consciousness. One thing worth mentioning here is Moravec’s paradox. It was proposed by Hans Moravec, Rodney Brooks, Marvin Minsky, et al. in the 1980s. Moravec’s paradox points out that, contrary to traditional assumptions, for computers, high-level reasoning requires very little computation, but low-level sensorimotor skills require enormous computational resources.

Scientists have found that it is comparatively easy to make computers exhibit adult-level performance on intelligence tests or playing checkers, and it is difficult or impossible to give them the skills of a one-year-old in regard to perception and mobility. Because a large number of perceived behavioral abilities depend on considerable tacit knowledge, human emotion and certain self-consciousness, they are difficult to algorithmize. To put it another way, the AI we face belongs to the “unconscious machine”, and its characters can be reflected in three aspects: (1) AI is not able to establish a connection with the external world since it is not capable of understanding the meanings of symbols and the physical world, especially the meanings of human languages; (2) AI has no consciousness, an important feature of human beings as a unique life form, and related experiences; and (3) AI lacks autonomy and self-awareness. What we see is the coupling of human-computer interaction.

At present, research on human and AI consciousness has become an important topic. Scientists are exploring how to help AI gain consciousness. One is to construct

consciousness through algorithms such as symbolic calculation and statistical calculation; the other is through brain-like research, that is, we can carry out relevant research under the inspiration of the study of human brain structure and working mechanisms.

The most well-known is the Human Brain Project (HBP) launched by the European Union. In 2013, the EU led the ten-year HBP, with 135 cooperation institutions from 26 countries becoming involved. They set an extremely high goal for this project with a high-risk investment of up to 1.3 billion euros. They planned to develop an information and communication technology platform that is committed to neuroinformatics, brain simulation, high-performance computing, medical informatics, neural morphology computing and neurorobot research. The research aimed to realize AI by simulating brain functions with these supercomputer technologies.

The project can be divided into three stages. The first stage is two-and-a-half years of the “climbing stage”, during which they established a preliminary version of the ICT (information and communication technology) platform and implanted it with some strategically selected data to prepare a platform for researchers inside or outside the project. The second stage is the “operation stage” for the next four to four and a half years. The targets of this period of time are to produce more strategic data and add more functions by strengthening platform application and exhibit the value of the platform to basic neuroscience research, drug development and application as well as future computing technology. The last three-year “sustainable stage” is to ensure that the project is financially independent and sustainable and can become a permanent asset of European science and industry.

Of course, this plan seems to have failed thus far. In contrast, China’s brain science research turned out to be more practical and reliable. In 2016, China joined the global competition of the “brain project”. The China Brain Project has two orientations: brain science research guided by the exploration of brain secrets and brain diseases and brain-like research directed by the establishment and development of AI technology. The China Brain Project took the study of the neural principle of brain cognition as the “main body” and the development of new approaches in the diagnosis and treatment of major brain diseases and new brain-machine intelligence technologies as its “two wings”. It aims to achieve international advanced results in the three frontiers of brain science, including brain science, early diagnosis and intervention of brain diseases and brain-like intelligent devices, in the next 15 years. The ongoing scientific research plan based on China’s actual situation and international experience has been recognized worldwide.

We will summarize the research and development of human and machine consciousness here. Minsky once said “man is just an emotional machine in the flesh”. Scientists proposed that the construction of conscious machines needs to meet the following conditions: perceptual state, conscious ability, attention ability, planning ability and emotion. Machines must have emotions to be conscious. That is, they know how to make value or moral judgments and behave morally. The AI ethics we are discussing need to learn human ethics and value judgments from bottom to top to generate spontaneous moral emotion and moral awareness. It is far from enough

to merely implant human norms into machine systems from top to bottom to acquire the ability to make ethical decisions that will satisfy actual needs and do no harm to humans.

Building a good ecological and social environment for human-computer interaction is a specific challenge we are facing. We should be particularly vigilant to avoid machines with malicious consciousness. If we can produce AI products with autonomy and self-consciousness in one day, the subject rights of the machine will also become a problem. We need to consider the future development of humans and machines, especially the construction of related morality and ethics, from the level of individual subjects, the individual self and the social technology system.

2.2 Looking at AI from the Differences Between the East and the West

Let us return to the process of human civilization after the discussions on AI and human consciousness. As the autonomy of AI continues to increase, its decisive role in human production activities is becoming gradually prominent. However, under the impacts of AI, traditional ethical concepts and moral theories have difficulties in applicability, and some propositions may even fall into the vacuum of moral philosophy, thereby triggering disputes.

Meanwhile, such risks also reversely force ethical and philosophical researchers to think more deeply. We should revise and optimize the current moral concepts and ethical theories based on the experience of sages and the cultural contexts and cognitive characteristics from different perspectives of East and West to put forward a philosophical moral system that can better adapt to the age of global intelligence. From multiple perspectives, we can explore the speculations of moral philosophy about AI at various times vertically and explore the ethical development of AI horizontally in the different cultural contexts of East and West, exhibiting the extension and expansion of traditional moral philosophy.

In essence, AI is the product of human beings and human civilization. Therefore, let us first think about the question of **whether there are similarities and differences in the thinking perspectives of Eastern and Western civilizations on AI and whether we can reflect on the development of AI from the angle of civilization.** Generally, to understand the challenges of AI, we can first think about the process of human civilization before taking a further look at human nature, human values, human evolution and alienation.

Let us start from Western civilization. Generally, the philosophical research on AI in the context of Western civilization came from the article “*Computing Machinery and Intelligence*” published by Alan Turing in 1950. Since then, Western philosophers and computer scientists have begun to be keen on discussing AI. They expect that machines can become intelligent as humans. They have human-like thoughts,

feelings, instinctive reactions and creativity. From the origin of more ancient civilization, the ancient Greek tradition defined human beings as creatures with rational thinking. Therefore, how to develop rational intelligence is the main path for major algorithms in expanding intelligence.

How to develop the “rational intelligence” of machines, that is, how to make machines have ethical attributes, has become the focus of applied ethics research. Multiple paths have emerged surrounding the topic, including those around value-sensitive design or those based on regional cultures. Among them, “humanism” centered on human rights and “technicism” are the most typical. The latter mainly focus on how to achieve the targets that contribute to human moral goals, including safety, credibility, reliability and controllability, through technology.

“Anthropocentrism” holds that man is not a part of nature but a superior existence. Homo sapiens has unique rationality, self-consciousness and subjective personality and occupies a higher position in the world than any other animals, plants or even any life being. After human society entered the era of industrialization and globalization, anthropocentrism developed to its peak. We should put human dignity and well-being at the core when considering AI ethics from the perspective of “anthropocentrism”. However, the inherent humanistic presupposition of this path leads to discussions on AI ethics falling into a metaphysical dilemma. For example, when ethical principles are embedded in artificial intelligence agents, difficulties such as the protection mechanism of human rights, the challenges of interpretation and the plight of beyond the minimum standard, etc., will occur.

In this field, a representative work is the “*Robot Series*” written by Isaac Asimov, a master of modern American science fiction, and the “Three Laws of Robotics” presented in this work, which are considered by many scholars as the theoretical basis of robot ethics and important principles for the construction of robot ethics. Isaac Asimov put forward it in “*I, Robot*” in 1950, aiming to ethically regulate future AI. “Three Laws of Robotics” are as follows:

- The first law: a robot may not injure a human being, or through inaction, allow a human being to come to harm.
- The second law: a robot must obey orders given it by human beings except where such orders would conflict with the first law.
- The third law: a robot must protect its own existence as long as such protection does not conflict with the first or second law.

What is the essence of AI different from previous technologies? In short, it is more crucial than tools. Previous technologies are just service tools for human beings, but AI is “intelligence” instead of being just a tool. It can be said that the functional understanding of AI is an important method, but one limitation of this understanding is that it sparks many people’s worry and fear about the impact of functional evolution—the proposal of “Three Laws of Robotics” represents people’s irrational fear of AI such as robots with emotional and empathic capabilities. Although robots can abide by the set of laws, they can live longer than humans, and their performance, tolerance and firmness under some extreme environments are better than those of

humans. All these strengths may provoke human jealousy and fear and make people want to destroy them.

In the West, technicism is closely connected with AI technology and is mostly valued by technical experts. Targeting the robustness and nonparaphrase of second-generation AI, technicism discusses how to realize moral purposes conducive to human beings through technology, such as safety, trustworthiness, reliability and controllability. Two typical representatives of this aspect are reliable AI and interpretable AI.

Moreover, we can also temporarily put aside the current mainstream Westernized perspective and understand AI from the perspective of oriental civilization, which is a meaningful ideological experiment.

Corresponding to the thinking on the “human nature” of AI from the perspective of Western humanism and technicism, the Confucianism represented by Confucius in Chinese civilization defines man as a social being in interpersonal relations and believes that man can recognize each other as human beings (the definition of “humanity”) and reflect on his own behaviors, values and thoughts (the way of “introspection”) are important embodiments of being a social person. From a philosophical viewpoint, the emergence of AI means that human beings are creating new species. If the Enlightenment Movement marks the departure of human beings from nature to become independent beings, then the existence of AI can be regarded as a significant practice of the movement. To put it another way, after people obtained the dominant power in the sense of ontology, they began to change the process and significance of civilization.

Some Western scholars believe that cutting-edge scientific and technological progress such as AI may pose a threat to human survival, while unlike the West’s excessive panic over AI, the Chinese take more inclusive and receptive attitudes toward it. As such, an interesting phenomenon, we can start with the three characteristics of ancient Chinese philosophy to help us understand why the Chinese are calmer in the face of the development of AI technology.

We note that Chinese Confucianism and Taoism were the main philosophical concepts in ancient times. Taoism holds that human nature is temporary. If a man is not obsessed with any object, then he can unify with all things on earth and realize the nature of things rather than regarding them as objective materialized objects. Confucianism believes that the family ethical relationship is the starting point of all values. The thought that “benevolence means to love others, and the greatest love for people is the love for one’s parents” means the formation of human nature is to gradually spread love from those closest to you to others, even to the world or everything. It can be said that in the world outlook of oriental philosophy, the concept of non-anthropocentrism—the “trinity of heaven, earth and man”—is a classic framework for Chinese people to understand the relationship between man, nature and society.

The concept originally came from *The book of Changes* in ancient China. According to the theory in the book, the three elements, nature, earth and man, as well as the two forces of yin and yang contained in them, are the most basic elements of the universe, in which nature evolves, human beings develop and society

progresses. On the grounds of this theoretical framework, human beings are essentially a part of nature and are closely linked with nature. Only by following the laws of nature and respecting the unity of man and nature can human beings multiply and thrive.

As individuals living on the earth, human beings have the unique ability to learn from nature. By virtue of the power of all things in nature, human beings make the world in which they live vibrant and sustainable and then comprehend the “Tao” and carry forward it in the world; that is, “man can make use of natural conditions to nurture all things on the planet through his intelligence”. From the perspective of Confucianism, “Tao” means that people should follow the moral teachings of benevolence, righteousness and integrity. Although Confucianism encourages people to enter into society positively, it also advocates that people should respect and even advocate the laws of nature and follow the laws of nature when dealing with specific affairs rather than exploiting nature intemperately. Human beings should understand the information of heaven and earth from various natural phenomena, such as seasonal changes, and act in accordance with the “Tao”.

In addition, Confucianism advocates that people should adjust their thoughts and behaviors according to the times and places. Mencius once praised Confucius as “the sage of his time”, which means that Confucius is a sage who can deal with the changes of the times. Therefore, Confucianism is not a rigid dogma but a set of wisdom that can be applied to different times and situations. With extremely rich understandings of human beings, Confucian tradition emphasizes the natural life order formed by interpersonal relations in society and extends such emotion to the generalized “benevolence” of all things in the world.

Correspondingly, the concept of “harmony between man and nature” has gained a more prominent position in Taoist ideas. Lao Tzu, the founder of Taoism, points out in *Tao Te Ching* that “I am abstracted from the world, the world from nature, nature from the way, and the way from what is beneath abstraction”. Tao is embodied in the three forces of heaven, earth and man. The mutual integration and supplementation of natural law, universal law and humanity can make heaven, earth and man coexist in harmony. Chuang Tzu, a Chinese philosopher who lived in the 4th century BC, further developed the thought of “harmony between man and nature”. He believes that heaven, earth and man came into being at the same time, and the existence of the universe and man is one.

In conclusion, we can assume that ancient Chinese philosophy did not give man a unique paramount status in the universe. As a developing advanced technology, AI is not the product of nature. From the perspective of “harmony between man and nature”, the development of AI should be limited to a reasonable range and guided and constrained in an orderly way to realize respect for the natural attributes of life. It is precisely because they are deeply influenced by their own nonanthropocentric philosophical tradition that Chinese people do not feel that their survival is being threatened by the development of AI technology; thus, they are not as fearful as Westerners.

On the one hand, many Chinese thinkers are unconvinced that artificial intelligence can surpass human intelligence one day; on the other hand, it is not uncommon

for machines or animals to outperform humans in some aspects. After the inheritance of super life legends such as immortals in Taoist culture, AI or some digital forms of intelligence can be considered a super life form in the eyes of Chinese. Some Confucians and Taoist scholars hope that AI can be integrated into the order constructed by human beings in the future. They have begun to take AI as a partner or a friend of human beings.

In addition to nonanthropocentrism, Chinese culture often holds a relatively open mind in the face of uncertainties and changes. Compared with the strong panics that arise from the development of AI technology in the West, China shows a higher tolerance for uncertainties and changes due to the influence of *The Book of Changes* (Zhou Yi). According to the key point of the classic, “a constant state of change” is the basic form of the existence of the universe, while “existence” with static characteristics is not the essence of cosmic existence. This idea of static existence was widely recognized in the European ideological community in the 20th century. Influenced by *The Book of Changes*, Confucianism believes that people should actively predict and deal with changes, which reflects the “humanistic vitality” of Confucianism. German sinologist Richard Wilhelm (1873–1930) preached in China in the late 19th and early 20th centuries. He was the first person to translate the book into a Western language. “There is no so-called eternal plight in this world; everything is changing. Therefore, even if we are facing an extremely difficult situation, we should do something to promote the formation of a new situation in the future” he once said.

Aside from Confucianism, since the Han Dynasty, Taoism has been characterized by its willingness to keep pace with the times and take the initiative to reform according to actual situations. Chuang Tzu’s viewpoint of “keeping pace with the times and rejecting stubbornness and pedantry” has gained an important status in contemporary Chinese culture. From the perspective of ancient Chinese philosophy, uncertainties and changes are not problems to be solved but an important part of the laws of nature and a constant in the universe. In Buddhist thought, impermanence is a very vital concept, and the essence of the reality we live in is illusory. The explanation makes many changes in the world more insignificant. The way of thinking in Buddhism also makes the Chinese take a more tolerant attitude toward AI.

Finally, the introspection, self-cultivation and consciousness advocated in Chinese philosophy also drive the Chinese to look back on their past when faced with AI and realize that perhaps the crux of the problem lies in ourselves. The common feature of the three ideological schools, Confucianism, Taoism and Buddhism, is that people should be self-disciplined, constantly reflect on themselves, and make endless efforts to achieve inner holiness. Moreover, these schools of thought assume that everyone should start with his own self-cultivation to build a good society, and a good society is hard to realize without good self-cultivation.

Therefore, in the eyes of Chinese philosophers, we should examine ourselves and obtain inspiration from the evolutionary history of human society when discussing the future direction of human technological development and debating the survival threat brought by technological progress. In other words, we need to review our past and realize that the key to the problem may lie in ourselves. We can never produce

morally satisfactory AI products unless we can reflect on our morality and assume our responsibilities.

As humans are facing an increasing number of global challenges, we may broaden our thinking and obtain inspiration from ancient Chinese philosophical traditions. It's time to get rid of the zero-sum game mentality, the preference for maximizing personal wealth, and fully unrestricted individualism. To create "human friendly" artificial intelligence or other cutting-edge technology products, we must work to build a global system that is inclusive, harmonious and compassionate to mankind. We can see that human self-cognition is constantly enriched through practical activities, and technological progress is expanding people's abilities and the "cognitive scale" of themselves. Therefore, it is necessary to conduct a more profound study on the philosophical values and connotations of the east in this process to obtain a new understanding of the ethical relationship between man and machine in an intelligent society, which is also the core of the discussion in this section.

2.3 Ethics and Moral Responsibilities of AI

As discussed above, we have benefited greatly from the development of AI, but we should also attach importance to the many potential ethical, moral and legal risks. For example, there are still no clear answers to a series of questions, such as the definition of medical accident responsibility of "Leonardo Da Vinci" surgical robot, the ownership of copyright of "XiaoIce" poetry collection, and the ownership and responsibility of human death caused by Uber unmanned motor vehicles. To avoid the moral hazards of AI, the idea of a moral machine has begun to emerge and received much support, but the view is based on emotion rather than rationality. The proposal of machine morality stems from human concern and panic about the development of AI technology. We hope that machines can be so moral that they will not harm us while serving us.

Machine morality is the expansion of human morality in the era of AI, while artificial intelligence is artificial. Whether AI has human sociality is still controversial in the academic community. Additionally, the study of AI at the legal level cannot be separated from the basic category of how AI bears responsibility. The premise to solve this problem is to clarify the legal status of AI and its relationships with humans, especially the impacts and challenges of future strong AI with self-awareness and independent decision-making ability on the existing judicial system, the right of personality, the subject status and legitimate rights of humans.

Compared with the practice of traditional ethics—the moral practice of "man", the moral practice of intelligent machines such as autonomous driving vehicles has significant particularities. First, machine behaviors are determined by the computational process of artificial design instead of a natural causal process. This distinguishes machine behaviors from ordinary accidents or natural disasters. Second, machine behaviors are not determined by its users to a great extent, which makes the machine different from the tools in the traditional sense. Finally, for machines with

wide applicability and the ability to deal with complex situations, especially those with learning ability, it is difficult or impossible for designers and manufacturers to accurately predict or control their behaviors.

Thus, traditional responsibility partition attribution is not suitable for intelligent machines due to these particularities. Take a self-driving automobile as an example. If a highly reliable autonomous vehicle accidentally violates traffic rules or causes traffic accidents, then it is not convincing to regard the incident as an accident, requiring passengers or designers to bear the corresponding responsibilities. To endow machines with moral subject status is to make them responsible for their own behaviors and decisions. In the long run, machines are able to be responsible for their behaviors and decisions only if they have moral subjectivity, which means machines have free will and can behave freely. It is reasonable to attribute moral responsibility to machines only when machines can act freely.

AI has the following characteristics that are highly similar to those of human beings: 1. Rational conditions—AI can independently complete the set tasks based on the algorithm model obtained from training and learning. At the same time, it can also realize unsupervised learning and evolution according to the real-time feedback data, which is similar to human learning behavior. 2. Moral internalization—human moral code of conduct is not something we are born with but something acquires in the environment. If the code of ethics can be quantified into data that can be recognized by the computer and input into AI, it can also be acquired based on its own learning characteristics. 3. Artificial intelligence has the possibility of enjoying corresponding rights and bearing related obligations and responsibilities. 4. Strong AI may have “free will” similar to human beings and generate human-like production behaviors through the mode of “cognition-learning-creation”.

The reason people have personality is that they have rationality. AI has a human-like rational essence. In the eyes of scholars who oppose “Anthropocentrism”, morally, we should treat all rational beings like human beings. As long as this rational relationship exists, no matter how different it is from *Homo sapiens* in biology, it is wrong to disrespect its human nature or treat it in an inhumane way. If we fail to follow this principle, on what grounds can we treat all intelligent beings equally? The premise of establishing such a rational relationship is that the object can obey morality like human beings and cause no harm to society, or the harm is generally within control. Only by doing so can we judge that the AI product has moral ability and meets the standards established based on ethical personalism. However, those unmoral AI products that may threaten the interests of the human race should not be regarded as moral subjects.

It should be noted that whether AI can be considered a moral subject is not necessarily related to the correct values that AI should follow. Even if AI does not have the status of moral subject, it has to observe correct ethics and values. Second, with the continuous expansion of the ability of AI, it bears greater responsibility. Meanwhile, restrictions on moral subjects will also limit responsibilities that humans should not bear. Finally, before judging that AI has moral ability, reliable standards should be set up to assist us in judgment.

The necessary and insufficient conditions for AI with moral subject status are cognitive ability, intent capacity and moral ability. Cognitive ability mainly refers to the human-like ability of perceiving the objective material world and spiritual world. It covers a wide range of elements, including memory, curiosity, association, perception, introspection, imagination, intention, and self-consciousness. Cognitive ability can be embodied as a process by which AI receives information from the environment and other objects through sensors and then creatively applies them in the process of self-calculation. Cognitive ability is also the building block of rationality.

Of course, only having cognitive ability is far from sufficient. Passively obeying the goals set by the programmer makes AI one of the ordinary tools without moral subject status. AI needs to set action goals for itself; that is, it should have “intent capacity”. The so-called intent capacity refers to behavioral competence or capacity for responsibility, which means “man has the ability to independently and responsibly determine his existence and relationships within a given range of possibilities, set goals for himself and limit his behaviors according to his essential attributes”. This ability includes communication, prudence, choice, decision and free will.

Although the levels of intelligence and cognition of AI on this occasion are no different from those of human beings, due to the lack of ethical regulations, the behaviors of AI can hardly center around the rights and interests of human beings, which may cause serious harm to human social order, core interests and even survival. On this proposition, we can cite a principle put forward by Finnis: “if the potential risks of damaging public good arise from the behaviors of our compatriots, we must accept them; but if they are caused by life beings other than our compatriots, then we don’t need to accept them”.

This requires AI to have a human-like moral ability and to try to reduce or eliminate the risk of harm to human racial interests as much as possible. Moral ability refers to the ability to judge and control one’s behavior in strict accordance with moral practice under the framework of moral norms and ethical regulations. The premise for someone to have rights is that he is essentially a person with ethical significance. If anyone wants to obtain moral subject status relative to others, that is to say, he has the right to ask others to respect his own personality and has the obligation to respect others’ personalities. Subjects with moral ability have sufficient moral reasons to ask others to respect their legal personality, while those without moral ability have no capacity to respect others’ personality or the need to be respected by others.

Since the proposal of AI in the 1950s, AI still remains in a weak stage after half a century of development. The technology of AI is mainly based on the promotion of deep learning, big data applied in specific application scenarios and computing power is in the primary stage. It takes a large number of data elements as the basis for thinking, decision-making and action. Static AI based on big data belongs to weak AI. Its algorithms may achieve exponential performance improvement in virtual environments, but it can only realize linear performance improvement in real data environments, which has certain limitations.

From the technical level, the next generation of AI may move forward in the following three directions:

First, from the perspective of the underlying technology paradigm, AI technology with complex computing ability based on dynamic and real-time environmental changes could be a development direction, including big data intelligence, swarm intelligence, cross media intelligence, hybrid augmented intelligence and intelligent unmanned systems. These are all directions that are being promoted in China as the new generation of AI technology. In particular, the last two, which are implemented through a new bottom intelligent system, are particularly noteworthy.

The second direction is AI technology based on a new application paradigm formed under complex goals from the perspective of application target scenarios. The complex goal here is to build an artificial intelligence technology for an intelligent society with urban groups as the goal, which involves new technical directions such as the landing of urban full-dimensional intelligent perceptual reasoning engines and autonomous learning oriented to media perception. The typical examples are the “Woven City” initiated by Toyota in Japan and the future city “Neom” being promoted by Saudi Arabia.

The third direction is a revolutionary AI technology based on the new chip architecture. In the Von Neumann architecture, the mainstream architecture of AI chips, there is a ceiling of memory performance due to separated units of computing and memory. The result also leads to limited growth space for processor performance and affects the effectiveness of the computing power of the current neural network model for complex scenario processing. However, the adoption of in-memory computing in the new chip architecture is of great value. Therefore, international academic conferences on semiconductors such as ISSCC or semiconductor manufacturers such as IBM and TSMC are committed to promoting and implementing the innovation of in-memory computing architecture.

We can see that although the so-called “strong AI” with “free will” has become a hot topic among the supporters of “subjectivity”, the following three characteristics of the current development of AI have forced us to doubt that scholars holding this view have fallen into a state of “utopianism”.

In terms of research direction, the research on AI has not explicitly moved toward the field of strong AI, and strong AI is still an illusory direction at present and will not significantly improve the value and benefits of AI to mankind. In terms of technical architecture, neither the computing architecture based on semiconductor materials nor the algorithm development based on binary logic can realize the “independent evolution” of AI. To realize strong AI, human beings need to break through the shackles of existing logic circuits and data structures, which is almost impossible in the short term. The lack of a sufficient theoretical basis is another barrier. In terms of research principles, both scientific research and industrial development need to strictly abide by ethical and moral principles. Even if strong AI can be realized, its risks are far greater than the benefits. The Pandora box should not be opened.

Artificial intelligence is a factitious nonliving entity based on intelligent technology rather than life metabolism. First, AI lacks the physiological basis of the human “carbon-based” brain. The human brain is the core of human life. Brain death marks the loss of legal personality. The “silicon-based” AI program or robot does not meet the physiological requirement of subject personality based on the existing

moral framework. Second, AI does not have rationality and will. Rationality covers people's perception of the characteristics and laws of the things around, the understanding of and compliance with ethics, and the empathy of emotion. Kant proposed that "human beings are rational and themselves are ends". Hegel agreed that "man has the purpose because of rationality", man has the ability of introspection and self-examination because of rationality. Different from human rationality, the "rationality" of AI is essentially a logical relationship based on an algorithm model, let alone human "natural" rationality. The value of its "rationality" is also to realize human will and purpose and much less introspection and self-examination, so it is impossible to break away from the restrictions of human regulation from the algorithm level for critical reflection. Although AI has deep learning ability, algorithm-based correction is essentially a kind of pattern recognition, rather than truly having natural rationality. The "intelligence" is a representation of algorithm simulation given by human beings.

Will is not simple logical calculus, and it contains two key factors: desire and action. Driven by desires, human beings make judgments on the ethical value of realistic propositions they face. Humans' desire and pursuit for survival, quality of life, identity and personality drive them to constantly transform themselves and nature to create a better living environment or make the most advantageous judgment. Neither AI has desire of its own and public nor emotional fetters. In the face of similar propositions such as the "trolley problem", it is difficult to evaluate the rationality of the decisions made by AI. In most cases, AI only performs probability calculus through human algorithms, let alone the possibility of establishing free will without the help of human beings.

Some critics believe that this idea of regarding AI as a person with nonfull capacity is full of discrimination against special populations who lack rationality and will as well as protection in human society, which is essentially an irrational embodiment. Moreover, children will also have rationality and will in adulthood. Vegetables with brain damage may also come to themselves and regain rationality and will, which obviously contradicts the characteristics of AI.

It should be emphasized that AI cannot bear responsibility under the existing moral framework, and giving it subject status is only a formal regulation. However, there are many gaps and blanks in propositions such as whether the control of AI on property and the ownership of property should go to AI itself or the human subject who makes it. Theories such as "finite personality" have essentially proved that it is unnecessary to endow AI with the qualification of moral subject, and the practice cannot pass the test of "Occam's razor principle".

The above is our preliminary thinking on the moral responsibility of AI. Later, we will discuss AI moral and philosophical propositions in various fields. The basic starting point for us to pay attention to these propositions in this book is that they are closely related to the fate of human civilization. As Will Durant and Ariel Durant defined in the last volume of their masterpiece, *The Story of Civilization*: civilization is social order promoting cultural creation. Four elements constitute it: "economic provision, political organization, moral traditions, and the pursuit of knowledge and the arts. It begins where chaos and insecurity end. When fear is overcome,

curiosity and constructiveness are free, and man passes by natural impulse toward the understanding and embellishment of life”.

In brief, only when we overcome the moral fear of AI in future human civilization can we create a brighter civilization. Human civilization pushes us to find solutions, and what we do is to find a way for humans and AI to achieve long-term coexistence while promoting civilization.