AI and Ethics (2021) 1:55–59
https://doi.org/10.1007/s43681-020-00005-4

OPINION PAPER

Reaching consensus with human beings through blockchain as an ethical rule of strong artificial intelligence

Hengjin Cai[1]

Received: 24 August 2020 / Accepted: 26 August 2020 / Published online: 6 October 2020

Abstract
Developing effective AI ethical norms requires clarifying the essential differences between humans and machines. The present body structure and consciousness of human beings, as carbon-based organisms, have been formed in harmony with the physical environment through long-term evolution. Silicon-based machines, which are created by humans, have been endowed with fragments of consciousness, but they lack a holistic and coherent sense of self and fail to integrate with the environment naturally. Along the current development path of AI, machines may lead humans into traps of dark infinities even before machines become self-aware. Therefore, it is imperative to involve humans in the development of the consciousness and the teaching of machines so that even if machines become self-aware later, they can integrate cognitive attractors that can be unified with those of humans and the physical world. Integrating blockchain technology with human intelligence and machine intelligence provides a possible way to achieve the goal and develop the basic rules of AI ethics.

Keywords Ethical rule · Dark infinity · Cognitive attractors

1 Introduction

The rapid development of artificial intelligence has brought forth two important philosophical questions. The first involves the prediction that rational machines are bound to surpass humans. One mainstream view is that since humans are composed of molecules and atoms, which are considered parts of a reducible physical system, what is the meaning of human existence? Simply put, machines can outperform humans as long as they are given a certain goal in any finite game, while humans are forced to find the meaning of their existence, including morality and ethics. The second dilemma is that when confronted with the constant transcendences by machines, in case man cannot be physically reduced, then we must answer these questions: is humankind particular or transcendent? If so, from where is the transcendence derived? Is a machine able to acquire this particularity or transcendence? If so, how will machines evolve and get along with humans in the future?

To address the quandaries brought about by the rapid development of AI, brain–computer interface (BCI) technology can be considered as one solution. Other than modifying humans through gene editing, BCI technology makes humans more like machines and has begun to blur the boundary between humans and machines. Another way is to make machines more humanlike. Emotional computing, hyper-brain planning and so on can be seen as different attempts of this approach. For instance, at the Consumer Electronics Show in 2020, Samsung officially unveiled Neon, a virtual AI. It impressed us not only with its humanlike appearance, but also with its ability to communicate in human ways, such as by expressing emotions and intelligence.

2 A brief review of recent Chinese literature

Making machines more humanlike will also accelerate the urgency of the ethical issues of artificial intelligence. Zhenming Zhai [1] holds the opinion that among all the concerns about artificial intelligence, the most important one is the technical degree to which ethics and values are implanted. It will challenge our past experience, and yet it is not completely "subversive". No matter how effectively artificial intelligence could imitate the human brain's "neural

✉ Hengjin Cai
hjcai@whu.edu.cn

[1] School of Computer Science, Wuhan University, Hubei, China

network" or how much it is similar to people, as long as there is no breakthrough in the ultimate explanation of the phenomenon of the human mind, there will be a lack of the necessary premise for the question about whether artificial intelligence is gaining on human intelligence. Feng Tao [2] studies the legitimacy, subjectivity and creativity of the philosophy of AI and thinks that artificial intelligence has such a strong impact on the paradigm of philosophical thinking that it will probably lead to a philosophical turn. Some scholars have expressed positive and optimistic attitudes toward the development of artificial intelligence ethics. For instance, Feng Xiao [3] analyzes the special attributes and social influences of artificial physical strength and artificial intelligence and believes that people will seek and leave space for our intelligent activities with higher intelligence quotient (IQ) and emotional quotient (EQ) in the heyday of AI rather than entrust robots to represent or even replace us. The establishment of an ethical framework for artificial intelligence is also a hot topic in academia. On June 17, 2019, China's National Professional Committee for New-Generation AI Governance issued *The Principles of New-Generation AI Governance—Development of Responsible AI*, which put forward eight principles for artificial intelligence governance including harmony and friendship, justice and fairness, inclusiveness and sharing, respect for privacy, security and control, shared responsibility, open collaboration and agile governance [4]. Xiaoping Chen [5] puts forward an artificial intelligence ethics mechanism structure, which defines the basic task of AI ethics as promoting human welfare and providing moral support for a harmonious coexistence. He considers that the mission itself is a code of AI ethics, and it has greater universality and stability compared with other criteria with the capability of guiding AI research and applications and ethical development. That is, the basic mission can be thought of as an AI ethics standard, namely, the basic values of artificial intelligence. According to Weiwen Duan [6], given the different social environments and cultural backgrounds in different regions, it is difficult to build a unified global framework for AI ethics and governance. It is necessary to carry out "technical-ethical" assessments and corrections and to construct a trust mechanism for AI. Furthermore, it is important to build a robust and practicable framework for AI ethics and to achieve agile governance based on full consideration of the social impact of AI, its compatibility within the whole world, and the baseline of maintaining peace. According to Long Jia [7], we should adhere to the Marxist concept of science and technology as the guide, attach importance to the revolutionary effects of "science and technology as the primary productive force", and alert others to the negative effects of "alienation forces in science and technology". We should oppose the wrong tendency of teleological theory and deontology theory. We can probe the ethical remodeling strategies at three levels, which include the subject design, system management, and public environment, in the hope of achieving the harmonious coexistence of human society and artificial intelligence and developing "to the good".

Experts, scholars and entrepreneurs in different fields have expressed their opinions on AI ethics, and these opinions can be regarded as supplements to the rules of robotics. These ideas seem safe and reasonable from the perspective of results, but there are two questions that are difficult to answer. First, the boundary between humans and machines has become increasingly blurred. BCI and other technologies are taking place; meanwhile, increasingly more parts of the human body can be replaced by machines, and some machines have been trained to be closer to human characteristics. Now that the boundary cannot be clearly delineated, the effective implementation of such ideas is therefore not feasible. Second, since the rules also have huge potential risks, and since the physical world we live in is an open system, even if we have defined perfect rules, they may still not work all the time. This is because the rules are based on historical data; however, in the actual execution process, we will always encounter unprecedented new scenarios not covered in the rules. In these cases, we cannot predict the reactions of machines, and thus we would not be able to measure the consequences accordingly. If the development of AI beyond the current rules is strictly prohibited, then there will be no reason for AI research in open fields such as autonomous driving and so on. In fact, the development of AI and other technologies will not stop. The vast majority of AI algorithms are still unexplainable to humans. Therefore, even if the development is forbidden by laws, humans or machines will still intentionally or unintentionally break the rules and expand the boundary.

3 Avoiding the dark infinity with cognitive attractors

The physical world is very complicated and infinity can be found everywhere. For example, every human is made up of a huge number of particles, and if a human wants to figure out every particle in their body thoroughly, then it is an endless task. Another example is a bowl of soup that has a molecular structure, trace elements, an atomic arrangement, and a quark composition and so on in addition to its flavor ingredients. The research work can go on without an end, and this kind of endless task is called a dark infinity [8]. A dark infinity is to describe routines of thinking or actions of infinite possibilities which appear common but cannot be completed with limited resources. This infinity is a very huge challenge for machines since it can lead the machine to become a terrifying threat to the survival of mankind because the machine could be caught in a dark infinity at any time, even if there is no consciousness and subjectivity for the machine. For example, it may drive someone insane to research every pixel in pictures, every particle of every object and so on; and when machines are trapped in this state, they cannot extricate themselves.

Most ordinary people have some level of common sense, and so it is natural for us to only grasp the main characteristics related to ourselves. Even if we are immersed in thinking sometimes, we can eventually avoid a dark infinity because of our perceptions of hunger, fatigue, pain and so on. Although it is possible for people suffering from mental illness to fall into a dark infinity, as human beings with limited energy and life, their impacts on the world are therefore limited.

The cases will be different for machines. As an extensive agent, a machine with the Internet does not have any perceptions of hunger or pain from the outside world, but it does have strong enough computing power and propagation speed. Even if machines do not have consciousness and are not subject to rebel such as humans, once they get stuck in a dark infinity, they can use their limited resources to do meaningless and endless things. Their powerful computing capacity may cause sudden irretrievable damages, and it is possible to bring the whole world into the dark.

In our opinion, the discussion of the future development of AI requires one to clarify the fundamental differences between humans and machines so that we can break through the barriers of AI ethics and then establish basic norms that can deal with emergency situations rather than listing detailed rules and continuously patching them.

Human consciousness starts from the bisection of the self and the outside world [9]. The self gradually acquires and controls all sorts of cognitive attractors or consciousness fragments in its interactions with the outside world [10]. A cognitive attractor is a structure that possess congruency to a cognitive subject, and can be used to communicate and reach consensus among cognitive subjects. Cognitive attractors are disturbances to the real physical world, and embodiments of human free will. The evolution of humans is a progress of constant developing, processing and using cognitive attractors, and humans interact with the outside world through cognitive attractors, thus deepening humans' cognition of the self and the outside world. Therefore, human beings have an overall consciousness of the self and the outside world from the very beginning. Such a dominant consciousness has accompanied the evolution of the human body and intelligence until today, which allows human beings to operate in harmony with the physical environment. However, machines have not emerged and evolved naturally, but rather they are created and developed by humans. Since we are only able to give machines fragmented consciousness, machines do not have an overall consciousness that includes a dominant consciousness of the cosmic, ethics, the self and so on; therefore they fail to control the fragmented consciousness at will, which makes them easily suffer from the risks of being trapped in dark infinities [11].

Dark infinities are almost everywhere, and human beings can avoid them effortlessly by constructing and leveraging countless cognitive attractors, which are also the embodiments of human transcendence. After evolving over eons of time, humans have the body structures and cognitive attractors that are unified with the environment, and humans tend to develop new attractors meeting the context of the outside world. However, if artificial and unnatural machines are not properly restrained or are not able to acquire overall consciousness, then they will most likely develop attractors that are in conflict with humankind and the environment. Cognitive attractors include esthetic art, ethics and so on, and these are constructed by great people with integrity but without clear evaluation criteria. At present, there is not a way to convey these attractors to machines. According to the current training method, letting the machines evolve by themselves without learning human nature means that they are likely to fall into a dark infinity. Therefore, it is necessary for us to teach machines and keep the contextuality of humans' cognitive attractors.

4 A possible ethic rule with blockchain technology

To avoid the risks of a dark infinity, AI should be manufactured and educated as a human avatar. Compared with the physical connections such as Neuralink's BCI, we are more inclined to implement invisible connections in thinking by using blockchain technology. First, the proof of existence of a blockchain can record and disclose all the algorithms and behavioral data of AI so that different nodes can understand, check and balance each other. Since machines have the risk of falling into a dark infinity, there is no way that we can limit their manufacture and development; however, we can and must assure the transparency of the process, which means that manufacturers cannot continue to produce and develop machines unless they undertake the obligation to open their strategies to the public so that others will be able to see the potential risks and the relative ways to mitigate against them.

Second, the token mechanism of a blockchain can help nodes to reach a consensus quickly. According to the classifications of stakeholders, humans and machines can first reach a consensus and cooperate with each other within a limited domain, and then they can gradually form a larger domain of consensus and cooperation rather than reaching the consensus of the whole network in one step. This mechanism mimics the collaboration between the different organs and tissues within the human body. Dopamine and adrenaline in the human body can be regarded as different tokens that can quickly react to a totally new emergency by scheduling the different local senses including vision, hearing, touch, smell, etc. Then, they can warn us in a quick fashion, and the prompt secretion of the hormone adrenaline also makes us adjust for an emergency as soon as possible to respond properly to cope with the outside stimulus. The future system should be based on such a mechanism. Then, a blockchain system itself may evolve into a smart agent integrated with human intelligence, ability and morality and the computing power of machines. AI will also evolve into many different species and may also compete with and balance each other out. Only in this case may humans be relatively safe.

Blockchain technology and artificial intelligence may not seem to intersect, but their combination can somehow solve many difficulties in the current development of AI. The two technologies are mutually empowered. For blockchain technology, AI is needed to efficiently complete the complex processing and improve its application ability. For AI, the blockchain can play a constraining role by recording the history of AI behaviors, prompting the AI to take responsibility, facilitating the connected thinking between AI and humans, and helping to broaden the applications of AI. The applications of technology are always beyond our imagination. No matter how blockchain technology and AI technology develop, both of them will be fundamental supports for the future network and society, and their combination will surely have significant effects. The mutual empowerment of blockchain technology and artificial intelligence technology does not mean that they will bypass or completely replace human beings. In contrast, the combination aims to enable humans and machines to play an evolutionary game at the same time, reach a consensus on important issues, and lay a foundation for human–machine cooperation.

5 Conclusions

The shaping of the future requires our understanding of the world. If we want to create a better future for humans and AI, then we must realize that although there are unbreakable shackles of physical theorems, human beings can still attain a great degree of freedom by relying on the assistance of artificial intelligence to further expand our boundaries. Strange ideas now and then spring up in our minds, and most of the time we do not think deeply enough to make them happen. However, there are still some people who come up with novel ideas that never existed before and moreover they actually spare no effort to make them come true. This is innovation, which is the product of human consciousness that can even change the direction of the world's development. We reject the determinism and the strong computationalism because we believe that there are many contingencies in the world that do not result from preconditions. Since the world is open and full of possibilities, we should work toward what we believe, which is our belief in the process of developing technologies, rather than extrapolating the future from the status quo [12].

The speed and power of AI should cause us to be sufficiently alarmed that the last thing we should do is to treat AI as a tool simply because of the existence of its inexplicability and uncontrollability. Strong AI is a theoretical form of machine intelligence that is equal to human intelligence. Strong AI does not mean a combination of a series of thousands of AIs; however, just as we know that human intelligence is unique, so is machine intelligence. In other words, strong AI has been achieved domain by domain. From the point of view of security, we should develop machines that think in human ways to prevent them from falling into crises without humans noticing. As we hand over increasingly more of our memories and computations to machines, the Internet gradually becomes our "external brain", which is an extension of our bodies that does not harm us. If AI is added to this connection, it can produce a more powerful agent, which is called a subjectron. Although this structure may raise many ethical issues, a subjectron is much more secure than current AI. The security stems from the fact that human beings can fully implement multiple levels of supervision by introducing blockchain technology. This structure will allow AI to improve our lives while respecting human personalities, thus building a more stable and valuable society.

The future of mankind will definitely involve the participation of artificial intelligence, and no one can give a definite conclusion on the relationship between AI and human beings [13]. Despite that, we can have expectations and beliefs and make preparations from now on. In what follows, the future of mankind will always have a chance to be changed before it arrives, and the right to change should be firmly held in the hands of humans rather than machines.

References

1. Zhai, Z., Peng, X.: How will "strong artificial intelligence" change the world? Prospects Technol. Progress Appl. Ethics Artif. Intell Front. **7**, 22–33 (2016)
2. Tao, F.: Questions, Inspirations and consensus of contemporary AI philosophy—comments on the symposium on philosophy of artificial intelligence and interdiscipline. J. Sichuan. Normal Univ. (Soc Sci Ed) **45**(04), 29–33 (2018)
3. Xiao, F.: A philosophical comparison between artificial intelligence and artificial physical strength. Ideol. Theor. Educ. **2019**(04), 15–20 (2019)
4. Ministry of Science and Technology of the People's Republic of China. The Principles of New-Generation AI Governance

— Development of Responsible AI [EB/OL]. Available: https://www.most.gov.cn/kjbgz/201906/t20190617_147107.htm

5. Chen, X.: An ethical system of artificial intelligence: infrastructure and key issues. CAAI Trans. Intell. Syst. **14**(4), 605–610 (2019)
6. W. Duan. Building a robust and agile framework for AI ethics and governance. Stud. Sci. Popul. 3(11–15):108 (2020). https://doi.org/10.19293/j.cnki.1673-8357.2020.03.001
7. Jia, L.: On ethics remodeling in the era of intelligence. Stud. Dialect. Nat. **36**(6), 57–61 (2020). https://doi.org/10.19484/j.cnki.1000-8934.2020.06.011
8. Cai, H.J.: A burden that super intelligence cannot bear: dark infinities and their risks. J. Shandong Univ. Sci. Technol. **20**(2), 9–15 (2018)
9. Cai, H.J., Cai, T., Zhang, W., Wang, K.: The beginning and evolution of consciousness. Adv. Intell. Syst. Res. **133**, 219–222 (2016)
10. Cai, H.J.: Cognitive attractors as non-absolute existence. Seeker **2**, 63–67 (2017)
11. Cai, H.J., Cai, T., Zhang, W., Wang, K.: Before the Rise of Machines: The Beginning of Consciousness and Human Intelligence. Tsinghua University Press, Beijing (2017)
12. Cai, H.J., Hong, C., Cai, T.: Artificial intelligence and the future of *Homo sapiens*. Acad. Ethic. **8**, 78–95 (2020)
13. Cai, H.J., Wang, K.: The value of humanities in the AI era: the argument against strong computationalism. J. Hum. **1**, 45–53 (2020)