From Virtual Reality to Metaverse : Ethical Risks and the Co-governance of Real and Virtual Worlds

by Yi Zeng, and Aorigele Bao

Note: This paper is originally published in Chinese on Philosophical Trends, vol 9, 2022. Here we provide an English version translated from the original paper.

With the increasing maturity of virtual reality, mixed reality, augmented reality, blockchain, artificial intelligence (AI), computer vision, cloud computing, mobile networks, and other fundamental technologies, “Metaverse” as an emerging concept of convergence technology has received widespread attention. What is “Metaverse”? What are the risks of the “Metaverse”? How to manage the Metaverse responsibly? In this paper, we try to analyze the existing dangers of “Metaverse” technology from the conceptual analysis of “Metaverse” and seek ways to co-govern it.

1. Analysis on the Concept of Metaverse

The “Metaverse” is both “heavy” and “imaginative” in nature. Its heaviness lies in the fact that the existing technologies can partially and preliminarily realize the vision of the “Metaverse” proposed at this discussion stage through technological aggregation. In other words, it is based on virtual reality technology and computer rendering technology to simulate the five senses, blockchain technology, sensor network to realize real-time data feedback, and AI-based information processing and machine learning as the technical basis to learn the virtual experiences. The innovative feature is that the “Metaverse” technology is highly conceptually open and has not yet formed a conceptual consensus. Therefore, based on technology development, it is a first priority to provide a relatively precise conceptual description of the “Metaverse.”

First, the “Metaverse” promises the public an unknown contextual architecture but faces the dilemma of conceptual ambiguity. On the one hand, in the public context, the “Metaverse” is considered as an emerging technological product by default. However, the promised architecture of unknown scenarios does not show uniqueness and does not explain the differences compared to very related concepts such as virtual reality, mixed reality, and augmented reality. The vague description of the unknown scenario architecture leads to conceptual ambiguity, partly because the current social innovation cannot support the development of “technological myths” and thus faces a conceptual interpretation dilemma in the sense of technological incompleteness. On the other hand, the translation of “Metaverse” as “Meta Universe” is inherently ambiguous. In the context of technological possibilities, the “Metaverse” does not share the characteristics of the universe at the abstract level, and

defining it as a “Meta Universe” leads to a giant gap between “high expectations” and “unfulfilled expectations. “Meta” has the connotation of “up”, “beyond”, etc. The technical description based on “Metaverse” is more appropriate than these, which is derived from “back” and “virtual frontier of copies”. As for the part of “universe,” most of the academic discussions about “meta-universe” are that “meta-universe” is a virtual world composed of a large number of virtual spaces, which indicates that “meta-universe” cannot cover the real universe, but is an extension of the real universe and the real world.

Second, the “Metaverse” attempts to achieve an open architecture through incompleteness but fails to address the need for contextual interactions. The “Metaverse” describes uncharted territory as an explorable incompleteness, giving users the promise of open architecture. Open architecture implies abandoning more external constraints and the generation of multiple and diverse virtual worlds based on technological convergence. Thus, the rise of the “Metaverse” is an alternative reflection of the individual’s need for freedom under the promise of technological development. Individual immersion is achieved through solid interactions in the promise of creative “freedom”. Immersion is the nature of human and technological creation through deep coexistence and interaction, expressing the desire to transcend the limits of the body. However, achieving immersion requires computer rendering technology and the user’s “active deception” to shape the “real” experience within the self through immersion.

Third, the “Metaverse” faces the “digital twin” dilemma. In the application of “Metaverse,” “digital twin” is mainly used to support the real-time modeling of the virtual world to physical world entities and enhance the immersion of “Metaverse” through data fusion modeling. However, today’s “digital twin” technology has limited scope and capability for real-time modeling, and the traditional human-environment interaction includes not only the exchange of scenery but also the interaction of humans and the environment. The requirements of the experience are not only limited to sensory interaction but also include the inherent need for interdependence between humans and the environment. Therefore, the secondary modeling in the virtual environment cannot simulate the whole interaction between humans and the environment.

The digital twin for individual avatars is also without completeness. On the one hand, the “digital twin” emphasizes the multidimensional generation of similarities, i.e., it illustrates the distinction between the physical body and the virtual body, but the rationality of the difference is not defined. In the redundant promise of an alternative “universe” the physical body still exists as the basis of individual life, and the virtual body is far from the level of physical body and is only an extended experience of the physical body. **Today’s AI technologies only support “symbolic” identity control on virtual platforms and cannot “clone” humans into digital systems. AI also lacks human understanding, awareness, and autonomy, and current intelligent information processing tools are far from being able to support a true digital twin, or a “virtual human” or “digital life” that is generated purely digitally without a human user as the target. The “digital twin” dilemma prevents the “Metaverse” from achieving true completeness.**

It is important to note that the "digital human" and the "digital twin" are not unattainable in the future, but the "digital human" generated by AI and the "digital twin" through "digitally cloning" an actual human facing related but not identical dilemmas. Whatever form and image the current "digital person" takes to the public, it is still essentially an information processing tool that seems intelligent. The modern AI-generated "digital human" is without the sense of "self", no "conscious mind", no ability to really understand, hence no basis for intelligence and wisdom. If we look at the "digital person" in terms of the reality of "information processing tools", its definition loses the possibility of being valid. If we want to realize the true meaning of "digital human", we need to reshape the foundation of modern AI that is with the purpose of constructing "tools", and start from the sense of "self", and build a "conscious mind", and realize the true meaning of "understanding". The vision of the "digital twin" is to "clone" real people, to create "digital bodies" and "digital doppelgangers" so that real people can be distributed into different scenarios with "doppelgangers". However, if the "digital twin" stays at the stage of "information processing tool," it will not be able to meet the needs from social applications and human society. If the "digital twin" has a "self" and "conscious mind," it will have or evolve "reflective ability", then the "digital twin" will not be satisfied with the existence of "avatar" and "doppelganger", and will even challenge the "cloned" original natural person in various forms and eventually leads to existential risks. Moreover, if the "digital twin" has a "self" and a "conscious mind", it will be difficult for the "cloned" orgininal natural person to accept that the "digital twin" is not only its "avatar" and "doppelganger". Thus, the "meta-universe" cannot negate its virtual nature, i.e., it cannot become "another universe".

The technology of the "Metaverse" should be highly related to the real world and cannot be completelly separated from it. The "meta" in "Metaverse" creates a sense of a more virtual world than the natural world and does not reflect the aggregated nature of the technology. The name "universe" is also exaggerated and does not indicate that today's "Metaverse" is more like a virtual reality frontier. Thus, the "Metaverse" is an immersive, interactive, and aggregated virtual experience domain. The "Metaverse" is not independent of the natural world of individual existence but is a virtual experience frontier based on the real world.

2. Stereoscopic real-existing risks

When AI and the "Metaverse" are developed as two separate technologies, their potentials can create two possible worlds and correspond to two possible world risks. When AI is a crucial supporting technology of the "Metaverse" or even a "responsible entity" in the "Metaverse", the risks of the two possible worlds are not simply a linear superposition but are presented as real three-dimensional risks, manifesting as three kinds of challenges: security and safety dilemma, cognition dilemma, and morality dilemma.

For security and safety concerns, first and foremost, is about privacy security. The threat to privacy security from "Metaverse" technology is mainly related to three aspects. Firstly,

private information can be maliciously stolen and exploited in a specific scenario to limit the subject's privacy. The development of the "Metaverse" will expand the need for privacy information from relatively homogeneous information to a broader level of life information. Secondly, privacy protection initiatives are prone to "consent fatigue". Excessive privacy permissions are more demanding on the public, and a specific understanding of data permissions is required to use privacy-protected virtual services. In order to protect their privacy, the public is confronted with issues of informed consent and revocation of authorization for privacy provisions. The lack of clarity between the responsibilities of the service providers and the users results in too much cognitive responsibility being placed on the user. Users' expectations are not respected by the service providers but create additional pressure of responsibility in the end. Thirdly, using non-essential personal information relies on the self-censorship of the technology service provider. Although technology providers are obliged to self-censor "how information is collected", there are many challenges at the implementation level. Second, security and safety concerns are manifested in the threat to emotional safety posed by "Metaverse" technologies. Some studies have shown that if one is exposed to external violence or aggression in a virtual reality environment, the harm is more severe than the emotional harm in reality. (see. Madary & Metzinger, p. 5.) Considering the wide range of possible audiences for "Metaverse" services, the possible emotional risks need to be considered in the design of the technology. Finally, security and safety concerns are manifested in the threat of technology to individual autonomy. As a possible technology that has yet to take shape, the "Metaverse" is imagined by the public's expectations and the designer's thinking, and the "Metaverse" will be different for different individuals' perceptions. As a comprehensive description of future technology, the "Metaverse" experience is also a limited imagination after technologization. According to M. Hilty's "techno-parental" measure, autonomy in the "Metaverse" is described as "the tendency to act and think in the user's favor", controlled by technological rationality in the opposite direction – in the name of "protection", admonishing the user. The "technological parent" is present in the Internet of Things (IoT) and AI technologies. (see Hilty, p. 14.) These technologies are also the basis of the aggregated technologies on which the "Metaverse" is based. When considering the security and safety concerns of the "Metaverse", it is necessary to ensure that the technology does not turn into a admonishing subject and that the individual's autonomy is guaranteed.

The extensive involvement of the "metaverse" can cause individuals to face cognitive dilemma. First of all, according to the "rubber hand illusion", when a person is under visual and tactile illusions, he or she corrects his or her proprioception to match the external illusion. Thus, the mere illusion of perception is sufficient to influence an individual's self-identification of himself. (see Botvinick & Cohen, p. 756). Thus, the immersive nature of virtual environments must significantly affect an individual's perception of the natural world and mental health. Once the psychological symptoms are extended to a broader range of user experiences, a more comprehensive range of cognitive risks is created. Second, individuals cannot complete their entire life cycle in a discontinuous virtual environment. People face the crisis of depersonalization in alternating between the virtual and the real. The cognitive dilemma of the virtual environment shaped by the "Metaverse" stems from the

discontinuity of the virtual experience itself. Individuals immersed in virtual experiences for a long time have difficulty adapting to the real world and confuse the boundaries between reality and the virtual. Finally, the potential failure of the expectations of "Metaverse" technology can easily lead to "addiction" to the virtual world. This is because, although one considers the "Metaverse" as a tool for achieving intentions when entering it, it is challenging to distinguish effectively between the realm of intentions and the realm of immersion. The discontinuous addiction to virtual worlds creates a psychological expectation failure of the individual to the virtualization technology. Thus, cognitive difficulties include the risk of mental illness, depersonalization, and addiction in the "Metaverse". The further going of cognitive difficulties will impact the individual's moral judgment.

Moral dilemmas are first manifested in the absence of moral risks. The emergence of moral ethics is related to the vulnerability of individuals. Because of individuals' vulnerability, we need to establish ethical contracts between different individuals based on mutual non-harm, respect, individual responsibility, and obligations. However, in the world of avatars, where autonomy is promised, personal experience is the first requirement. The lack of existential risk weakens the need for an ethical contract between individuals. Secondly, moral doubts are expressed in the difficulty of moral evaluation of virtual behavior. In the "freedom" experience, the individual has a similar experience in the "metaverse" that is free from physical limitations. The harm to others is also unrestricted by the transparency of the body. The social individuality of the individual is considered as a restriction of freedom in the description of the "Metaverse". The freedom of the virtual user is interpreted as "the ability to exercise autonomous choice within a virtual scenario". The moral evaluation of virtual behavior is directly related to the question of "how to perceive the intentions of individuals acting in virtual scenarios". Although the virtual act does not necessarily result in physical harm, the intent is created. The difference between unethical virtual behavior and real-world behavior is primarily only about degree and manner. The distinction of degree leads to the moral choice of virtual behavior as "trying to save others from harm". Thus, the justification for virtual behavior lies in the fact that virtual behavior affects the moral intuition of individuals, and there is no guarantee that individuals who are not morally constrained in the virtual environment will not blur the boundary between the virtual and the real in reality. Finally, in reality, behaviors that are wrong are also wrong in the virtual environment. The authenticity of the individual's experience ensures that the actions in the Metaverse involve actual causality, and causality ensures that the individual can gain real experience by trusting the Metaverse technology. In reality, the motivation for virtual behavior is no different from the motivation for moral behavior. Immersion also ensures that wrongdoing is not perceived as a symbol but only as natural, reprehensible behavior.

3. Co-governance of Virtual and Real Worlds

At present, humanity cannot find a perfect solution to the possible risks created by AI and the "Metaverse", let alone the risks arising from their combined effects. The "virtual world" that the "Metaverse" claims to build cannot be separated from the real world. It is necessary to consider further how to co-regulate the real and the virtual and

to ensure that this new scientific and technological product can be used reasonably and rationally.

One is the consideration of collective responsibility. To solve the above problem, it is necessary to consider how to attribute technical responsibility first. According to the "control theory of responsibility", the agent of the act should have a certain degree of control over the outcome of the act and thus be responsible for the act. (see Fischer & Mark, p.15). Suppose the AI technology on which the "Metaverse" is based has produced sufficient machine decision-making capabilities. In that case, the cybernetic account faces an explanatory dilemma when exploring the responsibility of the engineer or the relevant stakeholders. The conflict between the promise of automation and the cybernetic account arises, and the question of responsibility becomes more complex. Second, the "Metaverse" technology challenges assigning responsibility to the technology stakeholders. One is the "many hands problem". The "Metaverse" technology developer is not a single entity. However, it consists of many stakeholders, including AI technology providers, real-time communication technology providers, virtual reality technology providers, users, risk management agencies, regulatory agencies, etc. Therefore, the attribution of responsibility needs to consider the allocation of stakeholder responsibility. (See Van de Poel, Ibo, et al., p.49) The second is the retrospective responsibility dilemma. Due to the aggregated nature of the "metaverse" technology, it is tough to trace back to the previous technological stage, i.e., accurately, it is more difficult to effectively trace back the responsibility in the face of risk consequences. Thirdly is the dilemma of assigning responsibility by stage. In the "metaverse", technology innovation, technology product construction, marketization, technology maintenance, and other stages involve how to allocate responsibility. However, because too many stages and the parties involved are intertwined, it is challenging to allocate responsibilities.

If the metaverse becomes part of life, it must first be empowered by society. The needs of society determine the socialization of technology. The process of demand identification is the process of technological empowerment of society. Society defines its technology empowerment norms. Based on this norm, each emerging technology is identified as a need in the sense of society as a whole and allows the advancement of technology. In this sense, the reason for collective responsibility lies in that society empowers technological development and collectively bears technological responsibility. In addition, the emerging commitment to convergent technologies explains more deeply the universal vulnerability of human beings. In the virtual environment, technology becomes more vulnerable regarding intentions, patience of human, perception of the world, and the means of validating the value of life. Individual vulnerability also makes collective responsibility an option.

Thus, based on the cybernetic account, a good governance framework for the "Metaverse" should first consider the collective assumption of technical responsibility and understand the universal participatory nature of responsibility attribution in the sense of social acceptance of the "Metaverse" technology. Collective responsibility requires an understanding of the individual's responsibility for technology. The obligation to order the "metaverse" of

technology is assumed by each organic individual and is maintained through the commitment to the “metaverse” technology.

To protect the collective virtual order, to be responsible for their actions in participating in technical services. Thus, collective responsibility requires each user participating in the “Metaverse” to be involved in the risk perception process of the “Metaverse” and actively participate in the dialogue. The provider of technical services has an interpretable responsibility towards the user, through effective interaction between the user and the service provider, to ensure that the user does not waive his moral obligations on the grounds of autonomy. Secondly, it is essential to consider that not every user can afford social responsibility. Children and adolescents, who have not yet developed the capacity for self-responsibility, should not be pushed into this unknown space when the basic concepts, visions, and applications of the “Metaverse” are still very premature and with significant risks. Finally, it is essential to ensure service providers do not avoid responsibility based on “multi-handedness”. One is that the responsibility for the consequences of the behavior does not change for the technology provider because the consequences are not controllable. When it is not controllable or does not meet expectations, it is the responsibility of the technology provider to change the original design intent to make the technology consistent with human values. Second, the service provider must also consider the user’s “pre-determined informed consent”. Meeting public expectations should be an essential requirement for responsible service providers. Third, the risk governance of collective responsibility involves how people are perceived—considering people as entirely rational individuals or as beings based on a reward system faces significant problems. In an uncertain world, people can only seek satisfactory solutions through limited rationality but cannot reach optimal solutions.

Similarly, in the technological portrayal of human beings, the human group, shaped in an “optimal solution” manner, also faces a misinterpretation of human beings. Autonomy means not determining decisions through static images but returning autonomy to people based on incompleteness. The protection of autonomy should be based on the transparency of the technology. Technological transparency requires that producers promote the interpretability of the “metaverse” technologies to have a broad discussion inside and outside the technology community. The best result of public participation in technology discussions is a high sense of responsibility on the part of technology providers to construct good technology outcomes for people.

The second is the consideration of the nature of reality. The cognitive dilemma suggests that fundamental reality is more precious than artificial reality because of its not completely artificial properties. The depth of interaction on virtual platforms reduces the frequency of fundamental interactions. It limits the emotional demands of the social body dimension to interactions between virtual identities. It is vulnerable not only to the body but also to the vulnerability of interaction as a social identity. Individuals affected by technological immersion consider privacy, security and safety, and ethics risks as external constraints, giving rise to a new sense of “technological powerlessness”.

“Technological powerlessness” arises from the fact that the new technological participation is more attractive than the old individual life. The repressive society and the declining attractiveness of reality lead to an escapist mentality, and under the influence of consumerism, the quest for immersion becomes mainstream. High levels of immersion have traditionally been associated with emotional arousal, an alternative manifestation of dissatisfaction with reality. The problem with virtual environments is that although their immersion and interactivity can highly influence individual perceptions, the discontinuity and incompleteness of virtual experiences expose people to possible risks in the alternation between the two experiences. Only by allowing individuals to return to reality and participate in real human-human and human-nature interactions can the risks associated with virtual reality immersion be genuinely resolved and the priority of reality be restored.

Thus the creative presence of social activity and culture can manage possible risks in a broader sense instead of pursuing the digital twin of second nature, engaging in the interaction of cultural environments, focusing on the natural characteristics of the body, and replacing virtual immersion with real immersion. The virtual commitment is considered an exceptional state, and the concept of “Metaverse” is relegated to the original meaning of “extension of reality” to build a co-governance between the real and the virtual. The people create the value of the natural world in it, and the co-governance of reality should be the cultural co-governance of reality.

The promise of a “metaverse” exposes individuals to a generalization of identity forms in a postmodern context. The widespread dissemination of information and differences in cultural backgrounds have led to a more significant impact on normative social identities.

On the one hand, a series of social components such as individuals, social groups, states, and intergovernmental organizations and their relationships have been formed through a long history of evolution and political games. **Suppose the “Metaverse” tries to set aside the existing social relations. In that case, the vision of constructing new social relations in the virtual world should be bound by natural relations. Eventually, it can only become a subordination to or an extension of the actual social relations. Whether subsidiary or extended, the governance of the virtual world will form a “virtual-real co-governance” with the governance of the real world. The rational real-world regards the “Metaverse” as a new stage of the Internet application form, so the governance of the “Metaverse” is bound to inherit the governance norms and models of the modern Internet, AI, and other related technologies make necessary extensions. The governance of the natural world will consider the virtual world as an extension of reality, based on the governance model of the natural world and giving the extended virtual world a specific space for adaptation within a controlled range.**

On the other hand, “Metaverse” technologies offer the promise of sensory immersion, interactive diversity, and more prosperous and diverse information. They require a higher level of information integration from the user, thus supporting the non-fragmentation of virtual participation. The cultural self is at the heart of both the real and the virtual worlds. As

the real world is difficult to escape from in its entirety, the virtual world is still brutal to separate from the cultural self, as cultural influences remain with the user over time. The cultural self requires an appropriate cultural framework to end the information chaos promised by virtual technologies.

Mutual cultural understanding and trust require that the development of new technologies begins with responsible upstream research in the form of an exchange of ideas between researchers from different cultural backgrounds. Moreover, in a territorial sense, it ensures that all relevant stakeholders participate in a "governance community". The risk of "Metaverse" is not monocultural but a risk in the sense of aggregation of complexity. Cross-cultural understanding and mutual trust enable technical participants from different professional and cultural backgrounds to share technical knowledge and risk management initiatives. In a broader sense, the virtual environment reinforces the original individual cultural identity. Conflicting values are magnified in the virtual environment, leading to a broader range of conflicts. Therefore, "Metaverse" governance requires a risk governance framework that builds on the reality of cultural co-governance.

The difficulty of cross-cultural co-governance is dealing with the trust barrier between different cultures. The trust barrier lies in the difference in value norms. (see Higeartaigh，s.，et al. p. 579). A specific analysis of the values of different cultures reveals the phenomenon of value overlap – despite different cultural backgrounds, different cultures value privacy, freedom, and justice important. Thus, the difference in values is a difference in the weighting of values across cultures. The core of the problem of interculturalism and mutual trust lies in how to deal with the difference in value weights.

To reshape the virtual society, history is indelible. Mutual cultural appreciation and trust in virtual-real world co-governance must be based on historical openness. Each participant in the dialogue needs to maintain an open attitude toward other cultures and, based on cognitive openness, make intercultural virtual-real world co-governance not only stay in static participation but also go deeper into the historical dimension, seek the path of value weighting coordination from the universal historical openness, and build a realistic dialogue situation based on shared values.

4. Conclusion

From a technical point of view, when we promise the concept of the "Metaverse" to the global public and the political system, it needs to be based on shared fundamental ideas and vision. Moreover, it needs to be considered technically feasible and socially applicable. New concepts reinvented in complete isolation from the real world can lead humanity to real existential risks. Practical technology design requires an interactive relationship between the technology developer and the public to prevent unintended or intended harm. Informed consent for technology also suggests that dialogue is essential for individuals to participate in technology design. As technology takes shape, people consider it as an actual situation and

adapt to it. The individual's behavior is faced with the problem of providing rapid feedback to the external context. However, this process also faces the problem of technological exhortation, which makes it even more important to consider how to engage the public through dialogue. In the process of symbiotic technological development, the user identity is transformed.

Although Metaverse technologies are continuously evolving and have not yet proven to be fully applicable, some aspects of the future vision of Metaverse based on convergence technologies, such as the demand of some people for virtual reality, are becoming a growing reality in the lives of individuals. In order to address the risks associated with the uncertainty of "Metaverse" technologies, a high degree of risk sensitivity is required, and a framework of responsible co-governance of reality and fiction is needed to promote the sustainable development of "Metaverse" technologies. From the distribution of responsibilities to the mutual appreciation and dialogue among different cultures, and ultimately back to the value of life that each individual cares about. Through reconstructing cultural identity, we can build a realistic path of co-governance through mutual cultural appreciation and trust in a common and open historical dimension.

References

Botvinick， M.&Cohen， J， 1998, "Rubber Hands 'Feel' Touch that Eyes See" in *Nature* 391(6669).

Fischer， J.M.&Ravizza， M.1998, *Responsibility, and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

Hilty， LM.2015, *Ethical Issues in Ubiquitous Computing-Three Technology Assessment Studies Revisited*, Ubiquitous Computing in the Workplace.

Madary ， M.&Metzinger, T. ， 2016, "Real Virtuality: A Code ofEthical Conduct.Recommendations for Good Scientific Practice and the Consumers of VR-Technology" *Robotics and Al* 3(3).

Higeartaigh，s.， et al.，2020, "Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance", *Philosophy & Technology* 4(33).

Van de Poel, I., et al., 2012, "The Problem of Many Hands: Climate Change as An Example", *Science and Engineering Ethics* 1 (18).

About the Authors

Yi Zeng is a Professor at Institute of Automation, Chinese Academy of Sciences, Director of the International Research Center for AI Ethics and Governance, and Director of the Brain-inspired Cognitive Intelligence Lab. He is cross appointed as Professor to School of Future Technologies and School of Humanities at University of Chinese Academy of Sciences. He is the Chair of the Professional Committee on ICT and AI for Comittee of Ethics in Science and Technology at Chinese Academy of Sciences. (http://braincog.ai/~yizeng/)

Aorigele Bao is a Ph.D student at School of Humanities at University of Chinese Academy of Sciences, and Institute of Philosophy, Chinese Academy of Sciences.