Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.22528

A Review of Ethical Issues in the Field of Cybersecurity

Zijian Cui

Beihang University, Beijing, China
ddcuizijian@126.com

Abstract: The ethical dilemma of cybersecurity in the era of big data presents a multi-dimensional outbreak trend. As the speed of technological change far exceeds the update of laws and regulations, the crisis of technological trust and social ethical conflicts are intertwined, forcing the reconstruction of the global governance system. This study comprehensively deconstructs the ethical disputes in the current field of cybersecurity and summarizes five core contradictions, including the zero-sum game between monitoring rights and privacy rights, the systematic spread of algorithmic discrimination, the new hegemony of data colonialism, the gap between rights and responsibilities of vulnerability disclosure, and the failure of the AI application attribution mechanism. Based on the comprehensive analysis and systematic integration of multi-dimensional fragmented cases, this study proposes a hierarchical and progressive governance paradigm: the technical governance layer pragmatically improves the existing system, the institutional coordination layer links multiple mechanisms, and the cultural identity layer improves the digital citizen literacy. By balancing the dual logic of technological innovation and value constraints, this framework provides an operational governance path for cybersecurity regulatory departments to optimize ethical risk assessment tools, Internet companies to establish algorithm audit committees, and technology research and development institutions to improve ethical embedded design. It has important theoretical reference value for building a humanistic-oriented digital civilization order.

Keywords: Cybersecurity, ethics, artificial intelligence

1. Introduction

With the rapid development of information technology, cybersecurity has become a major issue that permeates all aspects of society and concerns the national economy and people's livelihood. It has evolved from the relatively simple virus attack and defense in the early days to network confrontation and even the formation of APT organizations and ransomware industry chains. In recent years, the popularization of artificial intelligence has brought more opportunities and challenges, and the ethical issues faced by this field have also increased.

Cybersecurity ethics is a discipline that studies the moral issues and responsibility conflicts arising from the application of digital technology and risk prevention and control [1]. The key lies in building a balance mechanism between technological needs and the protection of human values, covering the value trade-offs of multiple rights and interests such as personal privacy, corporate responsibility, and national security. Its core principles include controllability, situational integrity, and defense balance. In the past five years, the application of artificial intelligence has further expanded the scope of

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.22528

cybersecurity ethics, and new issues such as the intensification of attack and defense asymmetry and AI algorithm discrimination have gradually become the focus.

Through the summary and analysis of ethical issues in the field of cybersecurity in the past two decades, it can be seen that the research on cybersecurity ethical issues always lags behind technological iteration. On the one hand, because corporate R&D is guided by commercial interests, technological breakthroughs often precede ethical considerations, and ethical issues always appear after technology is put into use. On the other hand, due to the complexity of ethical issues themselves, the construction of ethical consensus takes a long time and may last for several years. The two lawsuits of Austrian privacy activist Max Schrems I and II spanned 7 years (2013-2020) before pushing Europe and the United States to abolish the Privacy Shield Agreement and reach a new Transatlantic Data Privacy Framework in 2022 [2]. In the period from 2018 to 2022 alone, 5G technology has been iterated three generations. In addition, many technical characteristics themselves will also lead to inherent ethical contradictions. For example, the tamper-proof nature of blockchain can both ensure data security and make it difficult for ransomware to track payments, which puts ethical assessments in a dilemma.

Therefore, this article systematically and comprehensively reviews the current ethical issues in the field of cybersecurity, which is conducive to integrating fragmented cases, establishing a systematic theoretical framework, and then revealing the development trend of current cybersecurity ethical issues, providing theoretical guidance for technical improvements and the construction of relevant policies and regulations, and promoting the construction of the field of cybersecurity ethics.

2. Core ethical conflict domain

2.1. Zero-sum game between surveillance and privacy

In the context of the digital age, surveillance is an important means of network security protection, and data collection and analysis activities for detecting and preventing threats are indispensable. The corresponding right to privacy, as a basic human right, emphasizes the protection of personal privacy and opposes redundant data collection. There is a natural contradiction between the two. The dispute over the balance point between surveillance and privacy has continued. The International Association of Privacy Professionals (IAPP) pointed out in the "2022 Data Governance White Paper" [3] that the proportion of redundant data in enterprise data lakes reached 34%, which increased the intrusion attack surface by 23%, becoming a major hidden danger of privacy leakage.

A typical infringement case is the FTC's lawsuit against Amazon's smart doorbell company Ring [4]. Between 2016 and 2020, Ring employees and their contractors accessed user video data without business needs and failed to implement basic privacy and security protections, allowing hackers to control consumers' accounts, cameras and videos. The case ended in 2023 with Ring refunding users a total of more than US$5.6 million and establishing a systematic data access approval process.

Similar privacy infringement incidents are widely distributed in various fields. The medical and financial sectors have long been the hardest hit by information leakage due to the high value of data. For example, in 2020, the well-known serious safety accident of the Brazilian Ministry of Health exposed the medical information of more than 243 million Brazilians on the Internet. Retail, social media and other fields are also important sources of privacy leakage due to their deep connection with users' daily lives.

In terms of geographical distribution, North America and Europe have a large number of cases and high fines due to the concentration of technology companies and the maturity of relevant laws and regulations. In recent years, the number of privacy leakage cases in the Asia-Pacific region, such as China and India, has also been increasing year by year, which may be related to the continuous advancement of digitalization and the imperfection of relevant laws and regulations.

DOI: 10.54254/2755-2721/2025.22528

2.2. The systematic spread of AI algorithm discrimination

In the field of cybersecurity, algorithmic discrimination in AI systems is mainly manifested in the systematic bias of security decisions against specific groups. This discrimination can be reflected in many aspects, such as race, gender, age, consumption, and employment. For example, in a 2018 study, Joy Buolamwini and Timnit Gebru tested the facial recognition systems of IBM, Microsoft and other companies and found that the error rate among dark-skinned women (especially African-American women) was as high as 34.7%, while the error rate among light-skinned men was as low as 0.8% [5]. In AI defense systems such as DDoS protection systems, the probability of misblocking normal traffic in some regions may be significantly higher than that in other regions.

The causes of AI algorithmic discrimination are relatively complex. It may be due to the algorithm designer's own bias projected into the algorithm, or it may be caused by the lack of common samples and training data, or it may be caused by the poor self-explanation and extreme complexity of the algorithm black box itself [6]. This is the key reason why AI algorithmic discrimination is difficult to audit and prevent.

2.3. Data colonialism expansion

Data colonialism is a general description of the phenomenon of network colonization in the era of big data. It is mainly manifested in the fact that technological powers use digital platforms and network equipment to collect, analyze and control data, and then control and exploit other countries or regions in multiple fields such as politics, economy and culture. It is an emerging order that uses data relationships to occupy and exploit resources for profit [7].

The gap in the field of cybersecurity is essentially a technological extension of global political and economic inequality. American cloud service giants such as AWS, Azure, and Google Cloud have taken advantage of their price and quality advantages to occupy a monopoly position in the global market, causing a large amount of data from all over the world to gather in their data centers, posing a risk of data being accessed by foreign governments. On the other hand, the international censorship and bans encountered by emerging technology companies such as Huawei and TikTok during their expansion may also be regarded by Western countries as a reverse case of data colonialism. For example, part of the reason why the United States banned TikTok was because it was worried that data would flow to China, which to a certain extent reflects the struggle for data control in the era of data colonialism.

In essence, the birth of data colonialism does not deviate from the critical framework of Marxist political economy [8]. With the development of society and the iteration of technology, compared with the traditional violent monopoly in the fields of politics, economy, culture, etc., the bourgeoisie is more inclined to adopt soft violent exploitation. Using digital platforms is undoubtedly the best choice in the era of big data. From this analysis, data colonialism can be seen as a specific manifestation of neo-colonialism at present.

2.4. Ethical dilemma of vulnerability disclosure

Unlike "black hat" hackers who destroy information systems and seek illegal profits, "white hat" hackers refer to technical experts who have a strong interest in computers and the Internet, and hope to explore the vulnerabilities and causes of software and hardware through their own hacking skills, and even take the discovery and reporting of vulnerabilities as their responsibility. If guided reasonably and correctly, this type of "white hat" hacker can become an important force in protecting network security. For example, in 2018, the Singapore government invited about 300 "white hat" hackers from all over the country to infiltrate the computer network system of the Ministry of Defense

DOI: 10.54254/2755-2721/2025.22528

to detect and eliminate system security vulnerabilities. However, "white hat" hackers often face various disputes when disclosing vulnerabilities on the Internet.

The first is the boundary dispute of testing authority. Security vulnerability testing activities that are not socially harmful and are conducted in good faith are generally referred to as “good faith intrusions,” but their clear definition and boundaries are still under discussion [9]: whether it is allowed to invade sensitive data areas under the pretext of testing, whether it is permissible to overstep authority when discovering vulnerabilities in affiliated companies, and other issues need to be resolved urgently. Some good faith intrusions may violate the law, such as the widely watched Yuan Wei case in China in 2016, in which white hat hacker Yuan Wei was arrested for suspected criminal violations while testing security vulnerabilities for a dating website. At the same time, there are also some hackers who secretly leak data under the name of “white hat” hackers, making it more difficult to determine the boundaries of testing.

Secondly, after the vulnerability is disclosed, the manufacturer’s delay in repairing the vulnerability may give attackers an opportunity to take advantage of it, and the responsibility in this case is unclear. The manufacturer may delay the progress of vulnerability repair for reasons such as saving costs and maintaining market reputation, resulting in a window period of several months between “disclosing the vulnerability” and “fixing the vulnerability.” During this period, hackers can easily exploit the vulnerability to attack, and there is significant controversy over the responsibility in such cases.

In addition, whether authorized vulnerability testing will lead to service interruption or data corruption is also a question worth considering. For example, during a penetration test, a cybersecurity company accidentally triggered a fault in a hospital system, causing some medical services to be interrupted. Such accidents that may have a huge negative impact are likely to be fully or partially the responsibility of the tester. In addition, the "dilemma of proof" caused by the company's possible attribution of existing system defects to the testing behavior has also become an important reason for some white hat hackers to hesitate.

2.5. The failure of the attribution mechanism of AI applications

The liability issues caused by AI applications in the field of network security present a complex situation of multiple subjects and cross-levels.

On the one hand, the complexity and black box nature of deep learning algorithms themselves [10] make it difficult to explain the mechanism of vulnerability occurrence once problems occur in AI applications. The technical untraceability of the "algorithm black box" makes it extremely difficult to determine responsibility.

On the other hand, with the further deepening of globalization, the responsible entities involved in artificial intelligence in the field of network security, such as developers, operators, and regulators, may span multiple countries. Even if the problem of the vulnerability generation mechanism is solved, considering the differences in laws and regulations between different countries and regions, it is bound to be difficult to divide the responsibilities between legal entities.

3. Governance framework construction

3.1. Technical governance layer: pragmatically improve the existing system

3.1.1. Establish a dynamic balance between privacy and monitoring

In the foreseeable future, monitoring and privacy in the cybersecurity field will remain in a zero-sum game for a long time, but a dynamic balance can be achieved through technical improvements. Consider improving existing methods such as federated learning + differential privacy technology

DOI: 10.54254/2755-2721/2025.22528

[11,12] to inject controllable noise into the monitoring system. For example, companies such as Google and Apple have applied local differential privacy technology to their products, making individual data unidentifiable while maintaining macro threat analysis capabilities and monitoring levels. It is also possible to adopt secure multi-party computing + data desensitization [13,14] to reduce security risks and enhance privacy protection while improving collaboration efficiency. At its core, the core is to minimize privacy exposure while ensuring necessary monitoring functions through the coordinated application of encryption algorithms and de-identification technologies, ensuring the confidentiality and controllability of the entire data processing chain.

3.1.2. Algorithmic discrimination mitigation

The mitigation methods for algorithmic discrimination mainly focus on the institutional level, but by deploying real-time correction modules in relevant programs, the generation and aggravation of algorithmic discrimination can also be effectively curbed at the technical level. For example, we can consider adding a "fairness feedback link" to the intrusion detection system, referring to the existing Adversarial Debiasing tool [15], and automatically adjust the threshold when the false alarm rate of certain specific groups is abnormal, so as to alleviate possible algorithmic discrimination.

3.1.3. Using cost war to curb data colonialism

The core problem faced by "colonial countries" under data colonialism is that they lack the technical level of cloud service giants in developed countries. Such countries can consider increasing investment in scientific research, promoting lightweight operating systems developed locally, and compressing AI security model parameters as much as possible so that they can be deployed under low computing power conditions. At the same time, the RISC-V open-source architecture is used to replace the traditional chip architecture, while improving transparency and security. Reduce economic costs and further enhance technological autonomy.

3.2. Institutional coordination layer: linkage multi-party mechanism

3.2.1. Standardization of vulnerability disclosure

In order to deal with situations such as manufacturers delaying the repair of vulnerabilities, network security regulatory authorities can refer to the Google Project Zero policy and require manufacturers to complete the repair within three months of receiving the vulnerability report. Otherwise, they will be punished accordingly, and "white hat" hackers will not be responsible for possible vulnerability attacks.

At the same time, the "white hat" hacker certification system and code of conduct can be further standardized. The United States and Europe are at the forefront of the times in this field. For example, Europe affirms the vulnerability detection rights of "white hat" hackers and provides legal protection through the "authorization + guarantee" model. The U.S. "Cybersecurity Act" stipulates the disclosure rules for "white hat" hackers when they legally obtain security vulnerability information without obtaining manufacturer authorization, that is, to ensure that the information is kept secret and eliminate user privacy information, etc. This kind of practice is worth learning from other countries.

3.2.2. Improve the responsibility identification system

When AI applications cause losses, a responsibility ladder system should be implemented, and a multi-dimensional scoring rule should be established to determine the weight of each party's responsibility by comprehensively considering factors such as AI autonomy, technical transparency, and data quality index. For example, for an AI defense system at level L3, with a technical

DOI: 10.54254/2755-2721/2025.22528

explanation coverage rate of <60% and a data deviation of <15%, the developer bears 80% of the responsibility, and the operation and maintenance party bears 20%. Different industries can flexibly adjust the influencing factors and weights according to actual conditions. An ideal promotion plan is to give priority to the implementation of the responsibility ladder system in high-risk industries such as medical care and electricity, and gradually promote it to the entire industry after optimizing it based on experience and lessons learned.

At the same time, it is possible to consider establishing and regulating a cybersecurity insurance system, and to establish "AI liability insurance" in the same way as medical and life insurance to encourage critical infrastructure to purchase AI liability insurance, so as to maintain a good AI market ecosystem.

3.2.3. Promote cross-border data governance cooperation

At present, international cybersecurity norms are still in the process of being gradually formulated and generated. Governments are committed to reaching a certain normative consensus, but are often affected by political games between countries. Therefore, establishing confidence building measures (CBMs) is the top priority for promoting consensus building [16]. The United Nations has always played an important role in the construction of international cybersecurity rules. Its subordinate Group of Governmental Experts on Information Security (UNGGE) and Open-Ended Working Group on Information Security (OEWG) are both important institutions for practicing confidence measures. The measures emphasized by the UNGGE, such as strengthening joint law enforcement, enhancing information sharing, and improving dialogue transparency, are all effective means to promote international cooperation in cybersecurity.

International organizations such as the European Union have played a demonstration role in promoting cross-border governance cooperation. By requiring member states and the European Defense Agency to develop a cyber defense framework, promoting communication and collaboration among civil and military actors throughout the EU, and ensuring dialogue with international partners such as other international organizations [17], the EU has achieved remarkable results. These practices are worth emulating and promoting by other international organizations.

Internet companies, non-governmental organizations and other private entities should also give full play to their subjective initiative, actively participate in the formulation of norms, put forward initiatives to make up for the current deficiencies in norms, and participate in cross-border data governance cooperation "from the bottom up" [18].

3.3. Cultural identity layer: improving digital citizen literacy

3.3.1. Establishing a global cybersecurity ethics education system

In order to popularize ethics education in the field of cybersecurity, enhance public awareness of prevention, and bridge the gap between cultures and countries, a feasible solution is to establish a cybersecurity ethics course framework led by the United Nations, and popularize relevant knowledge to civil servants, business leaders and even college students in various countries through international MOOC platforms. At the same time, full consideration should be given to the differences between different countries in the field of cybersecurity ethics, and the greatest consensus should be sought in the course, seeking common ground while reserving differences. This approach is conducive to breaking down cognitive barriers and shaping the common denominator of values.

DOI: 10.54254/2755-2721/2025.22528

3.3.2. Cultivation of narrative community

In the field of daily life, people can consider supporting transnational media to jointly produce cybersecurity ethics propaganda films or documentaries, using emotional narrative methods such as "privacy leakage/algorithm discrimination victim stories" to break cultural barriers, strengthen value recognition between countries and cultures, and build a narrative community across national boundaries.

4. Conclusion

The ethical dilemma faced by the field of cybersecurity has expanded from a single dimension to the level of social civilization reconstruction. Its core contradictions are concentrated in five intertwined dimensions: the game between monitoring rights and privacy rights has led to a value paradox between public security needs and the maintenance of individual dignity; the spread of algorithmic discrimination has led to the systematic exclusion of specific groups in AI security decisions; data colonialism relies on the monopoly of digital platforms to build a new hegemonic system, and developing countries encounter structural oppression in situations such as the transfer of data sovereignty; the ethical dilemma of vulnerability disclosure highlights the dual dilemma of the vague legal identity of "white hat" hackers and the evasion of corporate repair responsibilities; the AI application attribution mechanism often faces the dilemma of failure in scenarios such as autonomous driving accidents and smart medical misdiagnosis, and the diversity of technical black boxes and responsible subjects has disintegrated the traditional legal attribution framework.

In response to the above multi-dimensional ethical challenges, the hierarchical governance model constructed in this paper presents a three-dimensional response logic in practice: the technical governance layer pragmatically improves the existing system, the institutional coordination layer encourages the linkage of multiple mechanisms, and the cultural identity layer improves the digital citizen literacy. The practical significance of this governance framework is to provide methodological support for cybersecurity regulatory authorities to optimize ethical risk assessment tools, shifting the focus of supervision from end-point governance to full life cycle ethical review; guiding Internet companies to build algorithm ethics committees and data trust architectures, balancing technological innovation and rights protection in business decisions; and inspiring technology research and development institutions to transform ethical embedded design principles into specific development processes. More importantly, this model reveals the transformation path of the governance paradigm in the era of digital civilization - from technology control to value co-construction, from local repair to system reconstruction, and provides a theoretical anchor for building a global digital governance order from the perspective of a community with a shared future for mankind. Future research can be based on this model to further explore the ethical paradigm breakthroughs brought about by disruptive technologies such as quantum computing and brain-computer interfaces, as well as the rights and responsibilities reconstruction mechanism of virtual-real interaction in digital twin environments.

References

[1] *Yanyan Qiu. Cybersecurity and Ethics Construction [J]. Intelligence Magazine, 2000, 19(1):17-18, 21. DOI:10.3969/j.issn.1002-1965.2000.01.006.*

[2] *Yongjiang Xie, Lin Zhu, Jie Shang. The European and American Privacy Shield Agreement and its Implications for my country [J]. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 2016, 18(6):39-44. DOI:10.3969/j.issn.1008-7729.2016.06.007.*

[3] *International Association of Privacy Professionals. (2022). Data governance in the age of hyperconnectivity: 2022 whitepaper. IAPP Resource Center. https://iapp.org/resource-center/white-papers.*

DOI: 10.54254/2755-2721/2025.22528

[4] Federal Trade Commission. (2023) FTC Sends Refunds to Ring Customers Stemming from 2023 Settlement over Charges the Company Failed to Block Employees and Hackers from Accessing Consumer Videos.
[5] Buolamwini, J., Gebru, T. (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency, 81:1–15, 2018.
[6] Huaijun Wang, Xuhua Ru. Discrimination in artificial intelligence algorithms and its governance [J]. Research in Philosophy of Science and Technology, 2020, 37(2):101-106.
[7] Couldry N., Mejias U. A. The costs of connection: how data is colonizing human life and appropriating it for capitalism [M]. California: Stanford University Press, 2019.
[8] Shuai Li, Zhenbang Li. Research on data colonialism from the perspective of information politics [J]. Journ al of Changsha University of Science and Technology (Social Science Edition), 2024, 39(2):33-41. DOI:10.1 6573/j.cnki.1672-934x.2024.02.005.
[9] Ye Lou, Wenying Gao. Suggestions on the legalization of "white hat" hacker vulnerability detection behavior [J]. Journal of Hunan Institute of Engineering (Social Science Edition), 2019, 29(1):74-80. DOI:10.3969/j.is sn.1671-1181.2019.01.016.
[10] Yanhong Liu, Shanyao Gong. Research on the criminal liability of network service providers for AI decision s [J]. Journal of Guangxi University (Philosophy and Social Sciences Edition), 2022, 44(3):164-173. DOI:10. 13624/j.cnki.jgupss.2022.03.005.
[11] Shufen Zhang, Hongyang Zhang, Zhiqiang Ren, et al. A review of fairness in federated learning [J]. Journal of Computer Applications, 2025, 45(1):1-14. DOI:10.11772/j.issn.1001-9081.2023121881.
[12] Zhiqiang Gao, Yutao Wang. Research progress of differential privacy technology [J]. Journal of Communications, 2017, 38(z1):151-155. DOI:10.11959/j.issn.1000-436x.2017241.
[13] Wenke Zhang, Yong Yang, Yu Yang. Research on secure multi-party computation [J]. Information Security and Communication Confidentiality, 2014(1):97-99. DOI:10.3969/j.issn.1009-8054.2014.01.027.
[14] Huihui Jia, Chao Wang, Xiangmin Ji. Research on key technologies of data desensitization [J]. Microcomputer Applications, 2024, 40(6):164-167. DOI:10.3969/j.issn.1007-757X.2024.06.041.
[15] Reimers, Christian, et al. Conditional adversarial debiasing: Towards learning unbiased classifiers from biased data. DAGM German Conference on Pattern Recognition. Cham: Springer International Publishing, 2021.
[16] Geng Zhao. Research on the application of confidence-building measures in the formulation of international cybersecurity rules [J]. Intelligence Magazine, 2022, 41(11):89-96, 47. DOI:10.3969/j.issn.1002-1965.2022.11. 014.
[17] Qiang Li, Zeng Wei. International collaboration in cybersecurity governance [J]. China Science and Technology Forum, 2016(11):26-31. DOI:10.3969/j.issn.1002-6711.2016.11.005.
[18] Lei Wang. Bottom-up norm formulation and the generation of international cybersecurity norms [J]. International Security Studies, 2022, 40(5):130-156. DOI:10.14093/j.cnki.cn10-1132/d.2022.05.006.