Copyright Expectancy Right: Paradigm Reconstruction of AI Training Data Governance

Wenzhou Shu

French School, Sichuan International Studies University, Chongqing, China

School of International Law, Southwest University of Political Science and Law, Chongqing, China

Email: 20202402110033@stu.sisu.edu.cn

Abstract

This paper focuses on the intricate challenges in AI training data governance and innovatively introduces the theory of copyright expectancy right as a potential solution. It first dissects the existing "trilemma dilemma" in AI copyright governance, encompassing the ambiguous rights of data sources, the paradox in determining copyright for AI-generated content, and the lag in regulatory frameworks. Subsequently, the study conducts in-depth jurisprudential validation of the expectancy right theory, exploring its legal philosophical foundations—drawing on Locke's labor property theory (framing data contributions as "digital labor") and Rawls' principle of justice (ensuring fair data distribution under the "curtain of ignorance")—and highlighting its institutional comparative advantages, such as breaking the "all-or-nothing" logic of traditional copyright, compensating for the passivity of the unjust enrichment system, and compatibility with the EU Text Mining Exception. A three-stage governance model (technology, institution, and ethics) is constructed to propose a gradient implementation path, including blockchain traceability, Shapley value-based dynamic distribution, obligation configurations for different scenarios, and ethical guidelines like Habermasian interaction rationality and the "glass box" principle of algorithm transparency. Additionally, a formula (ER=(Q×0.6+V×0.4)×C) is developed for quantifying the realization of copyright expectancy right. Finally, the paper returns to the humanistic value of intellectual property law, reinterpreting the incentive theory and constructing a ternary balance paradigm of technological innovation, institutional protection, and humanistic care. The research aims to provide a new paradigm for AI training data governance, balance the relationships between technological progress, institutional fairness, and humanistic care, and contribute to improving the global AI governance system.

Keywords

Artificial Intelligence, Copyright Expectation Right, Data Governance, Humanistic Value

1. Introduction: the "Trilemma Dilemma" of AI Copyright Governance

1.1 Origin of the Problem

At a time when Artificial Intelligence (AI) is developing rapidly, the copyright issue arising from the governance of its training data has come into focus. Zhang Jiaxin (2024) [1] points out that the rights and interests of data sources are in a state of suspense.AI training relies on massive amounts of data from a wide range of sources, including numerous copyright-protected works, but data providers find it difficult to clarify their own rights and interests under the current system, which has led to an imbalance between their contributions and rewards. For example, a large number of images, texts, and other contents shared by Internet users on social media are accessed and used by AI training data collectors, but users fail to get corresponding rights recognition and financial compensation.

Wang Qian (2024) [10] suggests that there is a paradox in the determination of the copyright of generated content. AI-generated content is not traditionally created by human beings, but also exhibits work-like characteristics in some aspects, which makes it difficult to clearly define the attribution of its copyright according to the existing copyright law. Taking AI painting as an example, the paintings generated after inputting simple commands, the autonomy of AI and the influence of user commands are intertwined in the creation process, and it is difficult to judge whether it meets the requirement of originality under the copyright law and to whom the rights should be attributed.

Sun Yang (2025) [3] reveals the crisis of regulatory lag. AI technology is iterating rapidly, while legal regulation is difficult to keep pace. In the process of acquiring, using, and disseminating AI training data, new forms of infringement continue to emerge, such as data crawling and unauthorised data integration, etc. Existing laws are unable to regulate these behaviours in a timely and effective manner, resulting in frequent infringement but difficult to hold people accountable. It is difficult to pursue responsibility.

2. Theoretical Deconstruction: Jurisprudential Proof of Expectancy Right Theory

2.1 Legal Philosophical Basis

2.1.1 Locke's New Interpretation of the Labour Property Theory: Data Contributions as "Digital Labour"

Locke's property theory of labour holds that people invest their energy and time in an unowned object through labour, thereby acquiring ownership of the object. In the realm of AI training data, the contributions of data sourcers can be considered a form of "digital labour". They generate data through actions such as creation, sharing, and use, which become the basis for AI training. For example, the original articles and videos released by network creators provide rich materials for AI training, and their labour input should be rewarded accordingly, which is consistent with the viewpoint that work data sources, as the initial generators of data, should receive corresponding property benefit distribution when they actively provide or participate in contributing raw materials of work data, as put forward by Jiaxin Zhang (2025) [1] . This return is not an immediate property right in the traditional sense, but rather an expectancy right based on possible future gains. Because in the complex ecology of AI training data governance, the value of data can only be fully realised in subsequent AI applications and commercial development, as pointed out by Awasthy et al. (2024) [4], the commercial value of AI-generated innovations tends to emerge gradually in the process of their application and dissemination, and data sources should enjoy the right to look forward to the future benefits based on their data contributions. The data source should have the right to expect future benefits based on their data contribution.

2.1.2 Application of the Rawlsian Principle of Justice: Data Distribution under the Curtain of Ignorance

Rawls' principle of justice emphasises the allocation of social resources under the "curtain of ignorance" to ensure fairness and justice. In the distribution of AI training data, if all parties are behind the "curtain of ignorance", i.e., they do not know their roles in the data governance ecosystem (data providers, AI developers, users, etc.), they will be inclined to formulate fair rules for data distribution. From this perspective, copyright expectation rights can provide a fair distribution mechanism. Although data sources cannot immediately obtain full copyright, having anticipatory rights means that they can obtain a fair share of the value generated by the data in the future, avoiding the unfairness caused by the unequal distribution of the initial resources. For example, after the AI commercial application generates revenue, the data source can get a corresponding share based on the anticipation right, which is in line with Feng Xiaoqing's (2025) [5] argument that trade secret protection should balance the protection of innovation and the distribution of benefits to ensure that the data contributors get a reasonable return to ensure their fair rights and interests in the distribution of data.

2.2 Comparative Advantages of the System

2.2.1 Breaking through the "all-or-nothing" Logic of Traditional Copyrights

Traditional copyright follows the "all-or-nothing" logic, i.e., once a work is created, the author either owns the full copyright or does not enjoy any rights. In the AI training data scenario, this logic cannot be adapted. Feng Xiaoqing (2024) [5] points out that the complexity of the sources of AI training data makes it difficult to simply determine that a certain subject owns full copyright. Copyright expectancy rights break this logic by allowing the data source to have the possibility of acquiring rights and interests in the future even though they do not have full copyright at the data contribution stage. For example, in the case of open source data used in AI training, although the contributor does not have full copyright over the code as a whole, he or she may receive a corresponding return when the AI product developed based on the code gains revenue by virtue of the anticipation right, thus motivating more people to participate in data contribution.

2.2.2 Making up for the Passivity of the Unjust Enrichment System

Jiaxin Zhang (2024) [1] argues that the unjust enrichment system is passive in dealing with the issue of rights and interests related to AI training data. When an AI developer obtains benefits from unauthorised use of data, the data source needs to claim the right through litigation and other means, and needs to prove that the other party's profit is not based on lawful grounds and other complex elements. Copyright expectancy rights, on the other hand, are proactive in that they clarify the expectancy of the data source from the very beginning of the data contribution, without having to wait for the fact of unjust enrichment to occur before pursuing remedies. For example, if the expectancy right of the data source is stipulated in the data use agreement, when the AI product enters the market and makes profit, the data source will automatically get the corresponding income according to the agreement, which reduces the cost and uncertainty of defending the right.

2.2.3 Compatible with the EU Text Mining Exception

Ivana Kunda (2024) [6] found that the EU's Single Digital Market Copyright Directive provides a text mining exception, which allows text mining of copyrighted works under certain conditions, such as scientific research. The copyright expectation right is compatible with this exception. Where the text mining exception is met, the data source cannot prevent the data from being used for AI training, but can be compensated for any subsequent revenues arising from the use of the data based on the expectancy right. For example, if a scientific research organisation uses AI to conduct medical text mining research, and the medical data of the data source is used, when the medical AI products developed based on the research results are profitable in the future, the data source can obtain a corresponding share based on the

right of expectation, which not only protects the progress of scientific research, but also safeguards the rights and interests of the data source, which is consistent with the viewpoint of balancing the interests of the AI service provider and the data source as put forward by Yang Li-Hua (2025) [7] . source's interests, which is in line with Yang's (2025) view of balancing the interests of AI service providers and data sources.

2.2.4 Expectation Right Realisation Formula (Innovation Model)

In order to more accurately measure the degree of realisation of copyright expectation rights, the following formula is constructed:

$$ER = (Q \times 0.6 + V \times 0.4) \times C \quad (1)$$

Where, Q (data_quality) represents data quality, covering dimensions such as data accuracy, completeness, and uniqueness, with a value range of 0-1 (0 indicates extremely poor quality and 1 indicates extremely high quality); V (commercial_value) refers to the potential value of data in commercial applications, such as the role of data in enhancing the market competitiveness and profitability of an AI product after being applied to it, with a value range also of 0-1; C (contribution_index) is the contribution index of data providers, determined according to their participation in data generation, collection, and sorting, with a value range of 0-1. This formula shows that the realization degree of copyright expectancy right （ ER ） depends on data quality, commercial value, and the contribution degree of data providers, providing a preliminary framework for the quantitative evaluation of expectancy right.

From the perspective of parameter weight design, the priority of data quality (60%) is higher than that of commercial value (40%). This setting conforms to the core logic of information protection—that is, the secrecy and uniqueness of information are the premise of its value. For example, a high-precision medical case dataset (Q = 0.9) with professional annotations in the medical field, its scarcity and accuracy play a decisive role in the training effect of AI diagnostic models, which is far beyond that of ordinary public data. The introduction of the contribution index (C) reflects the principle that data providers should receive reasonable returns for their creative contributions to value increment. For instance, there should be a significant difference in the distribution of rights and interests between structured work data directly provided by original authors (C = 0.8) and non-original data reprocessed by third-party institutions (C = 0.3).

A specific calculation is made with academic paper data as an example: Suppose an AI company uses paper data from an academic database to train a generative AI model. The dataset has undergone professional proofreading (Q = 0.85), and after application, it increases the model's market share in the field of academic writing by 30% (V = 0.7). The original contribution of paper authors as original data providers corresponds to (C = 0.9\). According to the formula calculation: ER =(0.85×0.6+0.7×0.4)×0.9=(0.51+0.28)×0.9=0.79×0.9=0.711. This result indicates that the realization degree of copyright expectancy right in this scenario is relatively high, which not only reflects the core value of high-quality academic data but also reflects the reasonable rights and interests returns that original authors should obtain, consistent with the governance goal of balancing data utilization and the interests of providers.

3. Gradient Implementation Path: Three-Stage Governance Model

3.1 Technology Layer

3.1.1 Blockchain Traceability

Blockchain technology has the characteristics of decentralisation, tampering and traceability, which can provide an effective authentication and traceability solution for AI training data. Based on Wang Qian's technical solution, the blockchain is upgraded to generate a unique identification for each piece of data, and record the information of the whole life cycle of the data from generation to use. For example, at the data generation stage, information about the data source, data generation time, and data content summary are uploaded onto the chain; during the data flow process, information about the user, purpose of use, and time of use of the data are recorded for each data. In this way, it ensures that the source of data is clearly traceable and provides technical support for copyright expectation rights. When the data generates revenue in the future, the blockchain records can accurately determine the data contributors and the degree of their contribution, guaranteeing the realisation of their expectancy right.

3.1.2 Shapley Value Method Dynamic Distribution

Zhang Jiaxin's model proposes to use the Shapley value method to allocate data rights and interests, and optimise on this basis. The Shapley value method can fairly distribute benefits according to the degree of contribution of each participant in the cooperation. In AI training data governance, data sources, AI developers, data users, etc. are regarded as cooperating parties. According to the inputs and outputs of each party in data generation, algorithm development, data application and other aspects, the Shapley value of each party is calculated dynamically. For example, as the market performance of AI products changes and the commercial value of data changes, the Shapley value of each party will be recalculated in time to adjust the proportion of revenue distribution. Through this dynamic distribution mechanism, the copyright expectation rights of data sources can be reasonably realised at different stages, and all parties are incentivised to continuously participate in the data governance ecosystem.

3.2 Institutional Layer

Scenario	Obligation Configuration	Realisation Mode
AI Commercial	Mandatory disclosure + revenue sharing	Collective management organisations on behalf of
Research Use	Statutory Exceptions	Compensation system

3.2.1 AI Commercial Scenarios

In the AI commercial scenario, in order to safeguard the right of copyright expectation, a mandatory disclosure obligation should be configured, and AI developers should disclose key information such as the source of their training data and the way the data is used, so as to enable the data sources to understand the use of their own data. At the same time, a revenue-sharing system should be implemented, whereby AI developers are required to allocate a certain percentage of the commercial revenues from their AI products to the data sources. This process is carried out by collective management organisations, which, as a professional organisation, negotiate with AI developers on the proportion of revenue sharing, collect and manage data sources' information, and reasonably distribute the revenue sharing to data sources, so as to reduce transaction costs and improve the efficiency of realising the right of expectation.

3.2.2 Research Use Scenarios

For research use, set up statutory exceptions. On the premise of complying with scientific research ethics and legal regulations, researchers can use data for research without individual authorisation from data sources. However, a compensation system is established to protect the copyright expectation right of the data source. Compensation will be paid to the data source when the research project is funded or when economic benefits are generated from the research results. The amount of compensation is determined according to the scope of data use, the impact of the research results, and other factors, and in this way the reasonable rights and interests of the data source are safeguarded while promoting scientific research.

3.3 Ethical Layer

3.3.1 Habermasian Interaction Rationality Guides Multi-Party Consultation

Habermasian rationality emphasizes reaching consensus through equal and open dialogue and negotiation. In the governance of AI training data, it involves data sources, AI developers, users, regulators and other parties. Based on the rationality of interaction, all parties negotiate and jointly formulate data governance rules. For example, data governance forums are regularly held with the participation of multiple parties, so that all parties can fully exchange opinions on data use boundaries, distribution of rights and interests, etc., reach a consensus based on mutual understanding and respect, and form a governance scheme that meets the interests of all parties, so as to create a favourable atmosphere for the realisation of the right of anticipation of copyright.

3.3.2 The "Glass Box" Principle of Algorithm Transparency (Beyond the Black Box / White Box Controversy)

Traditional algorithms are subject to the controversy between black box (the internal mechanism of the algorithm is not transparent) and white box (the algorithm is completely open). Black-box algorithms make it difficult for the data source to know how the data is being used, affecting their copyright expectation rights; white-box algorithms are open and transparent, but may disclose trade secrets. The "glass box" principle transcends this debate by requiring algorithms to be transparent at key points. For example, the algorithm should disclose to the data source the core information such as the algorithm's processing logic, data filtering and integration methods, so that the data source can understand the flow of data in the algorithm, and at the same time, provide reasonable protection for the part involving commercial secrets, so as to guarantee the data source's right to know about the use of their own data, and thus better realise their right to expect their copyrights.

4. The Return of Humanistic Value: The Essence of Intellectual Property Law

4.1 Reinterpretation of Incentive Theory

Feng Xiaoqing (2024) provides an in-depth discussion of intellectual property incentive theory, which is traditionally "author-centred" and emphasises direct incentives to creators to promote the creation of works. In the AI era, this model has become limited. Shifting from "author-centred" to "ecological incentives" and integrating Kunda's view [8] implies not only incentivising creators in the traditional sense, but also focusing on all parties in the AI training data ecosystem . The copyright expectation right of data sources who participate in the ecosystem by contributing data is an incentive mechanism. When data sources know that their data contribution may bring benefits in the future, they will more actively participate in data generation and sharing, enriching AI training data resources and promoting the development of AI technology. At the same time, AI developers and users will have more opportunities for innovation and development in a good data governance ecosystem, forming a virtuous circle and promoting the prosperity of the entire AI industry and ecosystem.

4.2 Ternary Balance Paradigm (Innovation Framework)

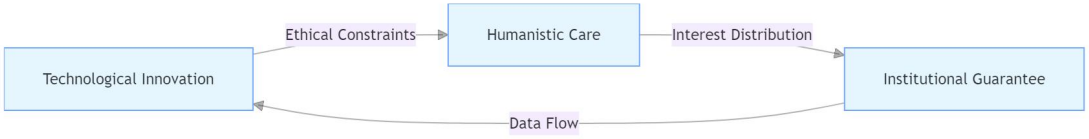

Figure 1. Triadic Balance Paradigm of AI Training Data Governance

In the governance of AI training data, a three-dimensional balanced paradigm of technological innovation, institutional guarantee and humanistic care is constructed [9]. Technological innovation is the core driving force of AI development, providing resources for AI training through data flow and promoting algorithm optimisation and application expansion. Institutional safeguards regulate data governance behaviours and protect the rights and interests of all parties through reasonable legal systems and policies, of which the copyright expectation right system is a key part. The distribution of benefits is an important element of institutional protection, and the benefits brought by AI development are reasonably distributed to ensure that data sources and other parties receive corresponding returns [10]. Humanistic care embodies respect for human dignity and rights, and ethical constraints are used to regulate the direction of technological innovation and avoid technological abuse. For example, in the process of collecting AI training data, ethical principles are followed to protect the privacy of data sources. This triadic balanced paradigm promotes and constrains each other, ensuring that AI training data governance takes into account institutional fairness and humanistic values while realising technological progress. The interaction mechanism of the three dimensions is visually illustrated in Figure 1: Triadic Balance Paradigm of AI Training Data Governance.

5. Conclusion

5.1 Anticipatory Rights Theory's Practical Echo of the "Twenty Articles on Data"

The "Twenty Articles on Data" emphasise the release of value and standardised governance of data elements, highlighting the dual goals of activating data circulation and safeguarding the legitimate rights of data subjects. The copyright expectancy right theory resonates strongly with these objectives in practice. By establishing a dynamic rights framework that links data contribution to future value distribution, the theory provides a concrete operational path for protecting data sources' interests—addressing the "value-reward mismatch" that often arises in large-scale data utilisation. For instance, the three-stage governance model (technical, institutional, and ethical layers) directly responds to the "Twenty Articles on Data"'s call for "improving data property rights systems" and "establishing benefit-sharing mechanisms." At the technical layer, blockchain-based traceability ensures transparent data flows, aligning with the requirement for "standardised data circulation"; at the institutional layer, revenue-sharing mechanisms and statutory exceptions balance efficiency and fairness, echoing the emphasis on "orderly data utilisation"; at the ethical layer, the "glass box" principle of algorithm transparency safeguards data sources' right to know, fulfilling the mandate to "protect data subjects' rights." Such a multi-dimensional design not only promotes the efficient allocation of data factors but also prevents the alienation of data value from its original contributors, thereby reinforcing the implementation of the "Twenty Articles on Data" and fostering a sustainable data factor market.

5.2 China's Programme for Global Governance (compared to the EU AI Bill)

Compared with the EU AI Act, the AI training data governance paradigm based on copyright expectancy rights proposed in this paper offers distinct advantages rooted in proactive source governance and dynamic balance. The EU AI Act adopts a risk-oriented regulatory approach, focusing on post-hoc oversight of high-risk AI applications (e.g., mandatory impact assessments and usage restrictions) to mitigate potential harms. While effective in risk prevention, this model may inadvertently stifle innovation by imposing rigid compliance burdens on AI developers. In contrast, China's proposal addresses governance challenges at the data source stage: by institutionalising copyright expectancy rights, it creates a incentive-compatible mechanism that aligns the interests of data sources, AI enterprises, and society at large. For example, the quantitative assessment formula for expectancy rights ($ER = (Q \times 0.6 + V \times 0.4) \times C$) provides a replicable tool for equitable benefit distribution, avoiding the ambiguity in "fair compensation" clauses often seen in international frameworks.

Furthermore, the Chinese paradigm transcends the binary tension between "protection" and "innovation" by embedding humanistic care into technical and institutional design. Unlike the EU's emphasis on procedural compliance, China's three-stage governance model integrates ethical constraints (e.g., Habermasian communicative rationality in multi-stakeholder negotiations) with technological solutions (e.g., Shapley value-based dynamic distribution), ensuring that data governance serves not only economic efficiency but also broader social values such as equity and respect for creators. This holistic approach is particularly relevant for developing countries grappling with balancing AI development and rights protection, offering a middle path that avoids overly restrictive regulation while preventing exploitative data practices.

In the context of global AI governance, such features make China's proposal a valuable complement to existing frameworks. By prioritising source-level rights clarity and flexible benefit-sharing, it addresses gaps in current global norms—such as the lack of actionable standards for data contribution recognition—and provides a blueprint for building an inclusive governance system. As AI technology becomes increasingly transnational, the copyright expectancy right model, with its emphasis on proportionality, transparency, and multi-stakeholder participation, can serve as a bridge between divergent regulatory philosophies, facilitating international consensus on data governance and advancing the goal of a fair, innovative, and human-centric global AI ecosystem.

References

[1] Zhang Jiaxin. (2025). Research on the Benefit Sharing Mechanism of Work Data Sources in Artificial Intelligence Training. Intellectual Property, (5).

[2] Wang Q. (2023). Re-examining the characterisation of AI-generated content in copyright law. Politics and Law Forum, 41(4), 17-22.

[3] Sun, Yang. (2025). The copyright exception system of generative artificial intelligence and its construction. Journal of Shenzhen University (Humanities and Social Sciences Edition), 42(3), 87-97.

[4] Awasthy, D., Bishnoi, A., & Meena, R. (2024). AI and Intellectual Property Law: Challenges and Opportunities in Digital Age. In 2024 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM). https://doi.org/10.1109/IIPEM62726.2024.10925706

[5] Feng, Xiaoqing, Li Ke. (2025). Reshaping the Rules of Trade Secret Protection in the Age of Artificial Intelligence. Intellectual Property Rights, (5).

[6] Ivana Kunda. Artificial Intelligence as a Challenge for European Patent Law [C]//MIPRO 2024, Opatija, Croatia, May 20 - 24, 2024.

[7] Yang, L.. (2025). Study on the Duty of Care of Generative Artificial Intelligence Service Providers. Comparative Law Studies, (3), 54-68.

[8] Kunda, I. (2024). Artificial Intelligence as a Challenge for European Patent Law. In 2024 47th MIPRO ICT and Electronics Convention (MIPRO). https://doi.org/10.1109/MIPRO60963.2024.10569722

[9] Khalifa, M., & Sabry, M. (2024). The Challenges of The Artificial Intelligence of Law in The Context of Technological Development. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS). https://doi.org/10.1109/ICETSIS61505.2024.10459547

[10] Al Nagrash, A., Alareed, N., Aldulaimi, S., Abdeldayem, M., & Aswad, A. R. (2024). Unveiling the Legal Implications of Regulating Information Technology Crimes in Violations of the Social Insurance Law. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS). https://doi.org/10.1109/ICETSIS61505.2024.10459364