中国科学技术大学
University of Science and Technology of China

Gender Differences in Corruption Cases in China: An Analysis Based on Deep Learning

Jingxin Yang,Runpeng Fu,Heshu Wang,Xinchen Wang

scaulife@stu.scau.edu.cn

1. Abstract

- Background: Despite the increasing research on corruption in academia, there is still a lack of consensus on the causes and analysis paths of corruption.
- Content: This article adopts a deep learning research approach to analyze the role of gender differences in corruption cases.
- Method: Research the establishment of the latest dataset based on China Judgment Document Network judgment documents, and apply a series of experiments to analyze whether deep learning can classify corruption gender, predict gender corruption, and identify differences in corruption characteristics between different genders.
- Result: Machine learning can perform gender classification and corruption prediction, and there are significant differences in corruption characteristics between different genders.

2. Background

Gender Differences in Corruption

- Criminology and behavioral science generally believe that there is gender difference in crime, and it is generally believed that men have a higher criminal tendency than women.
- The earliest research shows a essentialism biased view that women's innate moral sense is the decisive factor for their lower corruption.
- Scholars have also questioned this simple association. Hung EN Sung (2003) believes that the connection between gender and corruption is false and emphasizes the construction of a fair political system.
- In addition, there are also studies in the academic community specifically targeting gender differences in beliefs about corruption.

Hedging perspective and nepotism

- The hedging perspective is a new perspective proposed by Justin Esarey and Gina Chirillo (2013), combining previous gender studies on corruption.
- Making Bureaucracy Work: Patronage Networks, Performance Incentives, and Economic Development in China. For China's existing democratic centralized system, the recruitment and promotion of officials mainly rely on top-down appointments and are premised on entering the existing bureaucratic system.

Discussion on Gender Differences in Corruption in the Chinese Context

- Chinese female officials are often excluded from the bureaucratic network in China.

3. Objectives

The gender issue in corruption has gradually gained widespread attention in the academic community, and scholars have begun to attempt to construct models through a large amount of deep learning data to seek more accurate and innovative research results. Therefore, this study aims to integrate two major research trends by analyzing data hidden in CJO regarding corruption cases through deep learning. Specifically, we hope to address the following issues through this study:

(1) Is there a significant difference between men and women in corruption cases in China under deep learning?

(2) Can deep learning accurately predict gender in corruption cases?

(3) Is there a difference in the characteristics of corruption cases between men and women, and if so, what are the characteristics?

4. Method

Data Sources

This study obtained criminal first instance judgments on corruption and bribery crimes from China's Judgment Document Network, with a judgment date from 2020 to September 2022. This is a huge dataset, with 27946 bribery convictions and 23942 corruption convictions.

Data Preprocessing

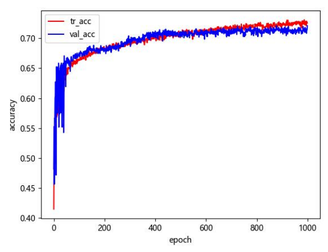
Use the third-party library NGender to predict gender. NGender is a natural language processing (NLP) tool designed to help users classify gender in text. Using NGender can easily identify male, female, or unknown gender in text and provide a reliable inference result. NGender's accuracy in inferring gender based on Chinese names is 70%

Deep Neural Network model

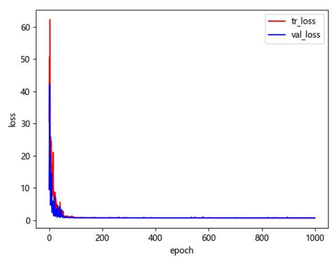
This article chooses to use a supervised machine learning model of Deep Neural Networks (DNN) to build input layers, hidden layers, and output layers. In this model, sofamax activation function is used to deal with the input of each layer nonlinearly. At the same time, the data is divided into 80% of the training set and 20% of the testing set, and the DNN model is trained based on this division. It is generally believed that DNN can overcome the bias of single elements.

5. Results

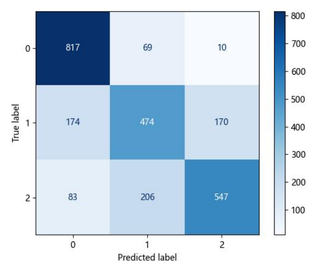
1. Gender Classification



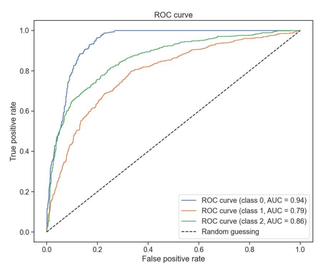
ACC curve based on DNN model



LOSS curve based on DNN model

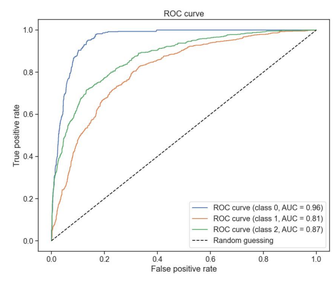


Confusion matrix based on DNN model

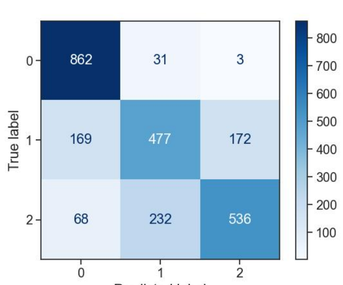


ROC curve based on DNN model

➢ LightGBM control group

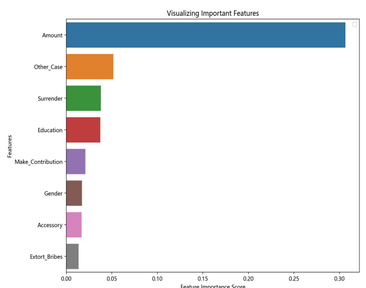


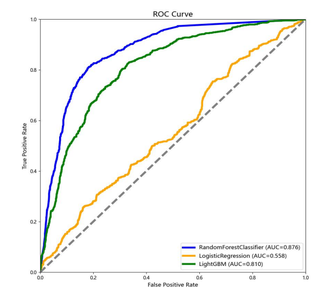
ROC curve based on LightGBM model



Confusion matrix based on LightGBM model

➢ Conclusion of Multi Models

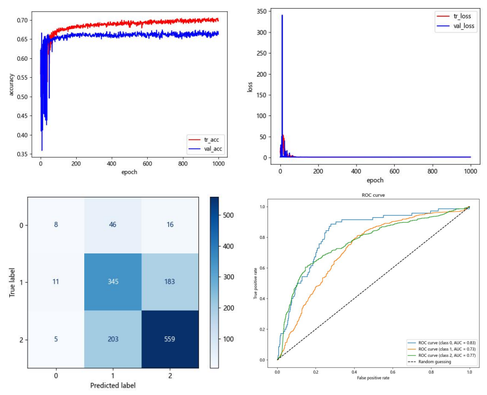




✓ As shown in the above figure, the DNN model has the best effect and performance. Meanwhile, DNN can accurately classify the gender of both men and women.The results indicate that there is indeed a significant difference between men and women, so further analysis can be conducted

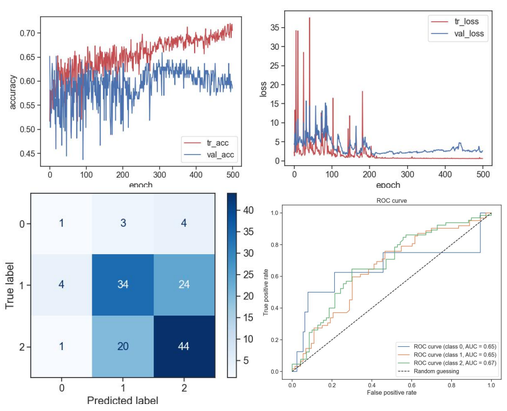
2. Prediction of male&female corruption

➢ Prediction of male corruption



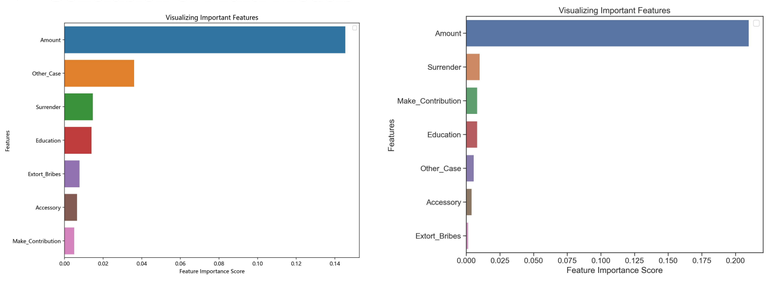
✓ Based on the experimental results on the left, we can conclude that the variable characteristics obtained from the study can effectively predict the likelihood of male corruption crimes.

➢ Prediction of female corruption



✓ Based on the experimental results on the left, we can conclude that the variable characteristics obtained from the study can effectively predict the likelihood of female corruption crimes.

➢ Conclusion of Multi Models

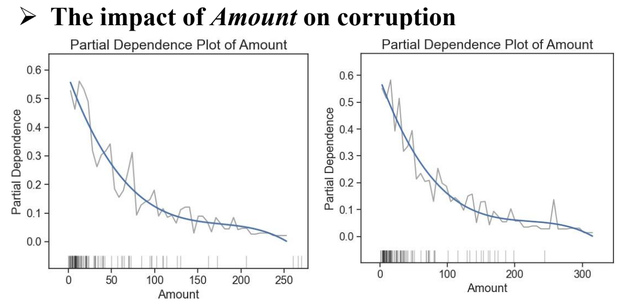


✓ Among the characteristics of *surrender , make consideration , other case* and so on, there is a significant difference in the characteristics of corruption between men and women

6. Conclusions

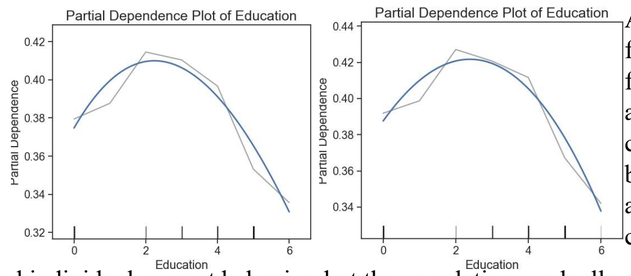
3. Comparing the Characteristics of Gender Corrupt Behavior Based on PDP

➢ The impact of *Amount* on corruption



As can be seen from the above figure, there is a high correlation between the amount of cases involved and individual corrupt behavior, but the correlation gradually decreases as the number increases

➢ The Impact of *Education* on Corruption



As can be seen from the above figure, there is a high correlation between the amount of cases involved and individual corrupt behavior, but the correlation gradually decreases as the number increases; This section shows that the impact of education level on corruption presents a "inverted U-shaped" characteristic, among which it is worth noting that corrupt individuals with a junior high school education have a higher likelihood of corruption.

96 Jinzhai Road, Hefei, Anhui, 230026, P.R.China
Fax: 0551-63632579 Email: oic@ustc.edu.cn https://oic.ustc.edu.cn

A Review of the Relationship between Gender and Corruption: A Machine Learning Analysis based on China Judgement Online Data)

Jingxin Yang [a,1], Runshu Fu[2]

ABSTRACT

Despite the growing number of studies on corruption in the academic community, there is still a lack of co-factual opinions on the causes and analytical paths of corruption . Meanwhile, gender differences are widely used in the analysis of causes of corruption , population attitudes and behaviors. Therefore, this paper establishes an updated dataset based on the judical documents in the China Judgements Online (CJO) dataset and applies a series of experiments to analyze whether deep learning is able to conduct three problems: classify and predict corruption by gender and identify the differences in corruption characteristics among different genders. By building a model with better classification performance under the comparison of various machine models, and then predicting gender on the basis of the model, we provide a new perspective for the behavioral analysis and research method of the corruption problem, and argue that machine learning is able to conduct gender classification and corruption prediction, and at the same time there is no significant difference in the corruption characteristics of different genders, and the results of the study do not support the gender-specific theory, which provides a prudent approach for the development of anti-corruption strategy development.

1. Introduction

At present, the fight against corruption is an important part of the work of governments to foster a wholesome political atmosphere and improve administrative efficiency. Although scholars in various countries have long explored how to reduce the problem of corruption, for the time being, the situation of corruption is still not optimistic. As the subjects of corruption cases often have different characteristics (e.g., gender, education background, etc.), and the current cases in mainland China are not studied using a relatively comprehensive and multi-year dataset, such as the difference in the bribery amount age at the time of bribe-taking, the education level of the involved officials, the sentence of the court between men and women. With the disclosure of the corruption cases written judgment information disclosure in various countries, a large amount of crime data is applied to the study and prediction of corruption.

Criminology or behavioral science generally assumes that there are gender differences in crime, and it is generally accepted that men have a higher propensity to commit crimes than women. In the process of research on corruption mitigation, two schools of thought have emerged, the ‘gender-specific’ and the ‘institutional-specific’, with regard to the key

1 Corresponding Author, Jingxin Yang, Nr. 5, Xihe East Road, West District, Guangzhou City, Provincie Guangdong; E-mail:2656229868@qq.com.

characteristic of gender differences. The ‘gender-specific’ posits that the female gender is inherently more integrity-prone, and that its traits include higher moral standards and more pronounced risk aversion (Dollar et al. 2001).[3] However, ‘institutional-specific’ suggests that different institutional cultures are determinants of gender differences in corruption (Esarey & Schwindt-Bayer, 2018).[4]

The earliest research demonstrated a bias in favor of the 'gender-specific' view, which argues that women's innate sense of morality is a determining factor in their lower levels of corruption. This argument is exemplified by the fact that many previous studies have shown that women's motivations and behaviors towards corruption are less positive than men's on a number of dimensions. Dollar (2001)[5] and Swamy (2001)[6] pioneered the finding that the more women representatives in parliament, the less corrupt they are. Chandan Kumar Jhaa and Sudipta Sarangi (2018) used the IV method to similarly demonstrate the negative correlation between women's parliamentary share and corruption rates, expanding to the fact that this relationship also exists when women are given the same social status as men.[7].Sasiwimon W. Paweenawat (2018) examined data from Asian countries based on the IV method while applying the generalized method of moments (GMM) to observe the persistence of corruption, yielding a similar result to that of Dollar (2001) and Swamy (2001), which is generally consistent with the conclusions.[8]

Some scholars have also questioned this simple correlation, leading to the development of the ‘institutional-specific’. Hung-En Sung (2003) argues that the link between gender and corruption is spurious, and emphasizes more on the construction of an equitable political system.[9] Namawu Alhassan-Alolo (2007) stresses that Relationship networks are more important in eradicating corruption, otherwise measures to increase the quota of women in the public sector as an anti-corruption strategy are likely to fail.[10]

As mentioned above, in past research on corruption, theoretical paradigms or empirical model has often been used to argue the issue. Although there is no lack of excellent research results, the accuracy of data and experiments in research is still lacking. Currently, in the latest research on corruption, people have begun to focus on individual influences and analyze them in depth and detail, including the impact of partisan alliances[11] and cultural stigmatization (Agoff, 2022) on corruption, and[12] also include the study of corruption by

---

[3] Dollar, D., Fisman, R., & Gatti, R.V. (2001). Are women really the 'fairer' sex? Corruption and women in government. Journal of Economic Behavior and Organization, 46, 423-429.
[4] Esarey J, Schwindt-Bayer L A. 2018.Women's representation, accountability and corruption in democracies [J].British Journal of Political Science, 48(3) : 659-690.
[5] Dollar, D., Fisman, R., & Gatti, R.V. (2001). Are women really the 'fairer' sex? Corruption and women in government. Journal of Economic Behavior and Organization, 46, 423-429.
[6] Swamy, A.V., Lee, Y., Azfar, O., & Knack, S.F. (1999). Gender and Corruption. development economics.
[7] Jha, C. K., & Sarangi, S. (2018). Women and corruption: what positions must they hold to make a difference? Journal of Economic Behavior and Organization, 151, 219-233. https://doi.org/10.1016/j.jebo.2018.03.021
[8] Paweenawat, S.W. (2018), The gender-corruption nexus in Asia. asia Pac Econ Lit., 32: 18-28.
[9] Sung, H. (2003). Fairer Sex or Fairer System? Gender and Corruption Revisited. Social Forces, 82, 703 - 723.
[10] Alhassan-Alolo, N. (2007), Gender and corruption: testing the new consensus. Public Admin. Dev., 27: 227-237.
[11] Alexander Stoecker, Partisan alignment and political corruption: Evidence from a new democracy, World Development, Volume 152, 2022, 105805, ISSN 0305-750X. https://doi.org/10.1016/j.worlddev.2021.105805.
[12] Carolina Agoff, Gustavo Fondevila & Sveinung Sandberg (2022) Cultural stigmatization and police corr uption: cannabis, gender, and legalization in Mexico, Drugs: Education, Prevention and Policy, 29:4, 373-3 81,

machine learning (Blasio, 2022). [13]It can be seen that acquiring, analyzing massive data set and ultimately studying the factors that characterize the problem of corruption and its causes by means of techniques such as machine learning is a new trend in the current research in this field.

As a result, we extract 8,768 first-instance criminal judgments of embezzlement and bribery from CJO (China Judgements Online), and employs deep learning to categorize different genders and analyze differences in corruption behaviors between male and female officials, which does not support the gender-specific or institutional-specific theories, but rather argues that the highly relational and male-political dominated bureaucratic corruption network, which is unique to the Chinese context, obliterates possible gender-specificity.

2. Nepotism and Risk Aversion

The risk aversion perspective is a new perspective proposed by Justin Esarey and Gina Chirillo (2013) in conjunction with previous research on the gender of corruption. [14]In the risk aversion theory, the link between the 'gender gap in tolerance of corruption' and 'institutionalized democracies/centralized regimes' was first tested. Empirical results show that there is little gender difference in attitudes towards corruption in authoritarian regimes, while men are more tolerant of corruption than women in more democratic countries (Justin & Leslie, 2017). [15]Second, the theory analyzes the relationship between women's political participation, institutionalized democracy, and corruption, and concludes that there is no correlation between women's political participation and corruption in authoritarian countries, but a negative correlation in democracies (Justin&Leslie,2019). [16]In response to this phenomenon, it has been argued that the different responses to corruption as an external stimulus can explain this phenomenon. Under the current premise that people generally have systematic discrimination against women, women face a lot of difficulties in entering the officialdom compared to men, and at the same time are more vulnerable to criticism from various problems, and it is more difficult for women to maintain their existing political status compared to men. As a result, women are more risk averse and have a stronger desire to follow existing rules.

Nepotism, emphasizing personal relationships, loyalty to groups and client networks (Jiang & Junyan, 2018). [17]For China's existing democratic centralized system, the hiring and promotion of officials mainly rely on top-down appointments and are premised on access to the established bureaucracy. Typically, officials tend to build complex networks of nepotism in the expectation that they will gain political protection and offer bribes and loyalty in return (Jiang & Zhang, 2020). [18]Thus, when women exist in the nepotistic networks of a centralized system, women are often involved in corruption simply because they follow the 'rules' set by

---

[13] de Blasio Guido,D'Ignazio Alessio,Letta Marco. Gotham city. predicting 'corrupted' municipalities with machine learning[J]. Technological Forecasting & Social Change,2022,184.

[14] Justin Esarey,Gina Chirillo. 'Fairer Sex' or Purity Myth? Corruption, Gender, and Institutional Context[J]. Politics & Gender,2013,9(4).

[15] Justin Esarey,Leslie A. Schwindt-Bayer. Women's Representation, Accountability and Corruption in Democracies[J]. British Journal of Political Science,2017,48(3).

[16] Justin Esarey, Leslie A. Schwindt-Bayer. Estimating Causal Relationships Between Women's Representation in Government and Corruption [J]. Comparative Political Studies, 2019, 52(11).

[17] Jiang, J. (2018). Making Bureaucracy Work: Patronage Networks, Performance Incentives, and Economic Development in China. American Journal of Political Science, 62(4), 982-999.

[18] Jiang, J., & Zhang, M. (2020). Friends with benefits: patronage networks and distributive politics in China. Journal of Public Economics, 184.

men who have more political freedom.

3. Governance of Corruption in China

Corruption has always been regarded as a systemic problem that threatens the Party and the country, and China has been taking severe measures to fight corruption since the establishment. Before the reform and opening up, it mainly took the form of campaigns to fight against corruption, and carried out campaigns such as the 'Three Anti-Corruption Campaigns,' the 'Five Anti-Corruption Campaigns' , the 'Rectification Campaign', and the 'Party Rectification Campaign', in order to ensure the Party's progress and purity. Since reform and opening up, China's anti-corruption efforts have gradually become institutionalized and professionalized, with the re-establishment of Party disciplinary committees at all levels beginning in 1977, and the gradual addition of anti-corruption administrative bodies such as the National Bureau of Corruption Prevention and the Ministry of Supervision in the years since, although anti-corruption efforts have mainly been concentrated in the work of the Party's Disciplinary Inspection Committees. Since November 2012, the 18th National Congress of the Communist Party of China (CPC), China has engaged in a new round of anti-corruption campaigns. Unprecedented in scale, scope, and level, party leaders believe that 'anti-corruption is the most thorough self-revolution,'[19] and is an important step in maintaining the party's purity and advancement, and stabilizing its power and credibility. Unlike in the past, this anti-corruption campaign is characterized by a 'multi-pronged, top-down' approach,[20] accompanied by a series of supporting set including political institutions, internal party mobilization, and legal system.

In order to improve the supervisory governance system, China passed the Supervision Law of the People's Republic of China in 2018, establishing the State Supervision Commission as the supreme supervisory authority, merging it with the former Ministry of Supervision and Bureau of Corruption Prevention, and co-locating it with the Central Commission for Discipline Inspection to perform the functions of disciplinary inspection within the Party and state prosecution, respectively, and centralizing the center of anti-corruption efforts to the State Supervision Commission. The data used in this paper comes from the judgment documents generated after corruption cases were examined by the Committee for Discipline Inspection or Supervision Commission and transferred to the judiciary for trial. Since the 1990s, with the establishment and rapid growth of China's socialist market economy, China's approach to large-scale corruption has shifted to a 'Grand Theft and Access Money' approach, in which political elites and social enterprises collude with each other.[21] This is a recurring theme in the Party's disciplinary or supervisory circulars.

4. Discussion on Gender Differences in Corruption in China

In China, a typical collectivist society, the network of bureaucratic relationships plays an

---

[19] Xi Jinping, 'Holding High the Great Banner of Socialism with Chinese Characteristics and Striving in Unity for the Comprehensive Construction of a Modernized Socialist Country-Report at the Twentieth National Congress of the Communist Party of China' (October 16, 2022), People's Publishing House, 2022 edition, p. 21

[20] Gong, Ting, and Wenyan Tu. 'Fighting Corruption in China: Trajectory, Dynamics, and Impact.' China Review, vol. 22, no. 2, 2022, pp. 1-19. 2022, pp. 1-19. JSTOR, https://www.jstor.org/stable/48671497. Accessed 4 Dec. 2022.

[21] Bakken B, Wang J. The changing forms of corruption in China. Crime Law Soc Change. 2021;75(3):247-265.

important role in the bureaucracy.[22] Tu,W and Guo,X. (2021) argue that in China's bureaucratic system, people tend to emphasize nepotism, personal loyalty, and are willing to link this network of relationships more closely through the common characteristics of different people, such as teacher-student, secretarial or alumni relationships. [23]Under this democratic centralized system, whether or not to get hired and promoted is inevitably linked to whether or not the official can be closely connected to the relationship network.Gong,T. (2015) argues that the corruption problem in China is mainly manifested in the problem of agency loss caused by individual mistakes and organizational abuse of power. [24]And the corruption characteristics of Chinese bureaucrats are likewise closely linked to system.Toke S. (2020) empirically analyzed that the bribes received by officials increase with the rank of the official, and that officials with both economic power to decide on expenditures and regulation have more bribes than those who hold purely administrative positions. Meanwhile, in Børge Bakken's (2021) argument, in China, which experienced high economic growth in the 1990s, the state-owned enterprises (SOEs), which were mainly engaged in the exploitation of land, mines and other resources, developed a high level of corruption that went hand in hand with the high level of economic growth, which made the anti-corruption system in China ineffective.[25]

Female officials in China are often excluded from the network of relationships in the bureaucracy in China. In a research of the entire bureaucracy of a county in east-central China, Feng Junqi (2010) found that there was a serious imbalance in the gender distribution at the deputy section level and above, with women accounting for only 6% of the total, and that serving female officials enjoy policy factors based on the preferential protection of female quotas.[26] Tu, W&Guo, X. (2021) argues that despite China's efforts to ensure gender parity and to provide female cadres with a unique promotion space, the gender ratio of female officials is significantly lower than that of developed countries, both at the local and central levels. Moreover, female officials are generally employed in positions that are not at the center of power.

Gender inequalities in political participation in China have also prompted scholars to examine the relationship between gender and corruption. Toke S. Aidt, Arye L. Hillman, and Liu Qijun (2020) find that gender is unrelated to the magnitude of bribe taking based on an empirical study of individual-level data on convicted corrupt officials in China.[27] Wenyan Tu and Xiajuan Guo (2021) use the Corruption Tolerance Survey of In-service Civil Servants in China to find that female survey participants are less tolerant of corruption than males. [28]They argue that due to the mediating role of nepotism, women are easily excluded from

---

[22] Ling Li (2011) Performing Bribery in China: guanxi-practice, corruption with a human face, Journal of Contemporary China, 20:68, 1-20,

[23] Tu, W., & Guo, X. (2021). Gendered clientelism and corruption: Are women less corrupt than men in China? International Feminist Journal of Politics, 23(4), 517-534.

[24] Gong, T. (2015). Managing Government Integrity under Hierarchy: anti-corruption efforts in local China. Journal of Contemporary China, 24(94), 684-700.

[25] Børge Bakken,Jasmine Wang. The changing forms of corruption in China[J]. Crime, Law and Social Change,2021,75(prepublished).

[26] Feng Junqi, Zhongxian Cadre, Ph.D. Dissertation, Department of Sociology, Peking University, 2010, pp. 27-30.

[27] Toke S. Aidt, Arye L. Hillman, LIU Qijun, Who takes bribes and how much? Evidence from the China Corruption Conviction Databank, World Development, Volume 13, 2020, 10498, ISSN 030-750X. Volume 133, 2020, 104985, ISSN 0305-750X

[28] Tu, W., & Guo, X. (2023). Gendered clientelism and corruption: are women less corrupt than men in China? 243.

networks of deals and shelter regarding corruption, thus leading to women's low acceptance of nepotism and low tolerance for corruption. Ni Xing and Xiong Jing (2020), relying on data from the 2016 China Integrity Survey, confirm that women have higher perceptions of integrity than men.[29]And judicial data from Taiwan, China shows that the main reason for women's lower corruption is due to women's marginalization in the bureaucracy rather than risk aversion, and the more marginalized women are, the less corrupt they are (Yi-Ming Yu et al., 2022).[30]

5. Methodology

In this paper, we use a data source that excludes the subjective factors of traditional questionnaire-based research and features a comprehensive volume of data, combined with machine learning , which has a higher prediction accuracy compared to data mining (Saeed, R. & Abdulmohsin, H., 2023),[31] to analyze the data hidden in the CJO about corruption cases and to achieve the double optimization of the data and the research methodology.

Deep learning analysis can process a large amount of text data to extract patterns and trends in it, so as to provide a more comprehensive and objective understanding of the perception of gender-unintuitive differences in it.The accuracy of deep learning analysis is usually affected by many factors, such as the quality and completeness of the dataset, the design of the model and the adjustment of the parameters. Therefore, before performing machine learning analysis, we cleaned and preprocessed the data to ensure the accuracy and credibility of the analysis results as much as possible. In addition, it is generally necessary to uphold an agnostic stance when using machine learning for social science research, and it is also not assumed that a particular machine learning method is the most applicable (Justin Grimmer, Margaret E. Roberts & Brandon M. Stewart, 2021).[32] Therefore, this paper also tries as much as possible and at the same time by plotting ACC curves, LOSS curves and confusion matrices to test the accuracy of the model, the fit of the model in the training samples and the recognition ability of the model, and tries to find out the model with the best performance.

Specifically, in accordance with the research problem of this paper, we need to first classify males and females to verify whether there is a significant difference between males and females in the problem of corruption discrepancy, and then predict the characteristics of male and female cases respectively, trying to find the differences between males and females in case characteristics. In this paper, we choose to use the supervised machine learning model of Deep Neural Networks (DNN) to build the input, hidden and output layers. The sofamax activation function is specifically chosen in this model to nonlinearize the inputs of each layer. At the same time, the data is divided into 80% training set and 20% testing set, and the DNN

---

[29] Ni Xing, Xiong Jing. Gender Differences in Individual Characteristics, Social Structure, and Perception of Cleanliness--An Analysis of Data Based on the 2016 National Integrity Survey[J]. Theory and Reform, 2020, No.234(04):161-175.

[30] Yu, Y.-M., Wu, Y.-M., Chiu, P.-L., & Chen, K.-H. (2022). Marginalization or risk aversion? Using big data to examine why women are found to be less corrupt in court judgments. Governance, 1- 20. Governance, 1- 20. https://doi.org/10.1111/gove.12752

[31] Saeed, R. & Abdulmohsin, H. (2023). A study on predicting crime rates through machine learning and data mining using text. Journal of Intelligent Systems, 32(1), 20220223. https://doi. org/10.1515/jisys-2022-0223

[32] Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: an Agnostic Approach. Annual Review of Political Science, 24, 395-419. https://doi.org/10.1146/annurev-polisci-053119-015921

model is trained based on this division. It is commonly believed that DNN can overcome the single-element bias. In order to validate the performance of the model, we first predicted the outcome of the trial and subsequently the gender.

In order to conduct comparative experiments with DNN, we test Random Forest, Logistic Regression and LightGBM with the same data and data structure.To observe the overall performance of the models, we focus on the AUC as well as the differences in AUC between the models. Also to test the accuracy of the models, the fit and recognition ability of the models in the training samples, we plotted the ACC curves, LOSS curves and confusion matrices. In order to observe the overall differences between males and females in the cases, we used difference analysis, divided into two groups, males and females, to see if these differences were statistically significant.

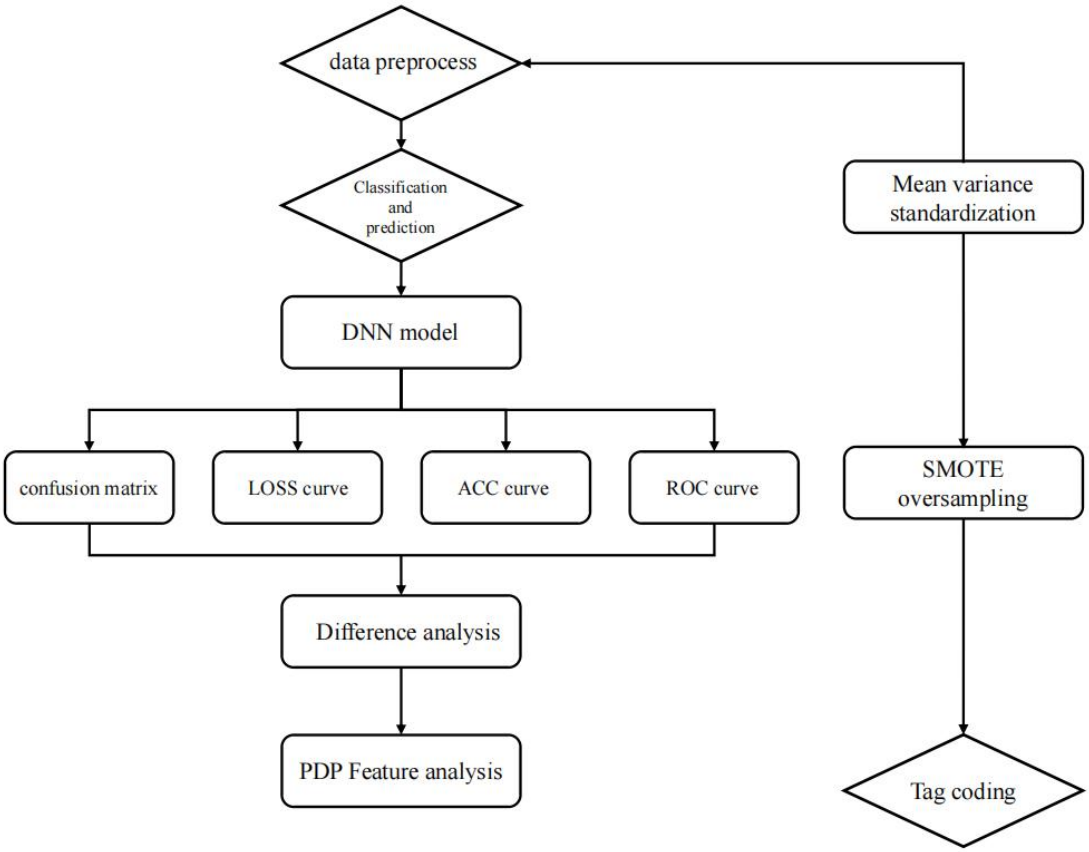


Figure 1. Model flow chart

Ultimately, we hope to address the following questions through this study:

（1） Are there significant differences between men and women in corruption cases in China under deep learning?

（2） Can deep learning accurately predict gender in corruption cases?

（3） Are there differences in the characteristics of corruption cases between men and women and, if so, in which characteristics?

6. The China Judgement Online (CJO) Dataset

China Judgment Online (CJO), the official platform designated by China's Supreme People's Court (SPC), has a relatively complete set of official records and judicial data of

anti-corruption agencies, which provides possibilities for further empirical research (Weimin Zuo & Chanyuan Wang, 2020).[33] A number of scholars have already conducted research on criminal justice based on Chinese adjudication documents (Yiwei Xia,Tianji Cai&Hua Zhong, 2019; James T. Graves&Alessandro Acquisti, 2023; Peng,Yali&Jinhua Cheng , 2022; Moulin Xiong,Siyu Liu&Bin Liang, 2022; Ma,Chao,Chao-Yo Cheng&Haibo He, 2022; Yin Qi&Ruobing Wang, 2022).[34][35][36][37][38][39]

Therefore, this study obtained criminal first-instance judgments for embezzlement and bribery with decision dates from 2020 to September 2022 in China from the China Judgment Online[40] . This is a large dataset with 27,946 acceptance of bribe judgments and 23,942 embezzlement judgments. First, judgments containing multiple indicted persons are excluded for one case has only one defendent. As China is a typical statutory law country with more rigid sentencing standards for the crimes stipulated in the criminal law, China adopts uniform sentencing standards and sentencing circumstances for acceptance of bribe and embezzlement, both of which are based on the criteria of 30,000 yuan to 200,000 yuan, 200,000 yuan to 3 million yuan and more than 3 million yuan as three different sentencing grades, respectively, with aggravating or mitigating factors according to the stipulated circumstances.[41][42]Therefore, the data on acceptance of bribe and embezzlement can be combined . As a result, we screened out the valid cases and pre-processed the data for the remaining data of 8,768 valid ones.

Before moving on to the next step of model building and analysis of variance, the Table 1. below provides the results of the descriptive analysis of the main statistical variables.

Table 1. Full sample descriptive statistics (n=8768)

Demographics	
Officer gender	Male:91.1%

[33] ZUO, W., & WANG, C. (2020). Judicial Big Data and Big-Data-Based Legal Research in China. Asian Journal of Law and Society, 7(3), 495-514.

[34] Xia, Y., Cai, T., & Zhong, H. (2019). Effect of Judges' Gender on Rape Sentencing: a Data Mining Approach to Analyze Judgment Documents. china Review, 19, 125 - 149.

[35] Tianji Cai, Li Du, Yanyu Xin & Lennon Y.C. Chang (2018) Characteristics of cybercrimes: evidence from Chinese judgment documents, Police Practice and Research, 19:6, 582-595,

[36] Peng, Y., & Cheng, J. (2022). Ethnic Disparity in Chinese Theft Sentencing: a Modified Focal Concern s Perspective. China Review 22(3), 47-71. https://www.muse.jhu.edu/article/ 864166.

[37] Xiong, M., Liu, S., & Liang, B. (2022). Death Sentence Review by the Supreme People's Court in China: Decision Patterns and Variations*. China Review, 22(3), 137-166.

[38] Ma, C., Cheng, C., & He, H. (2022). From Local to Upper Capture: The Chinese Experiment of Administrative Courts. China Review 22(3), 9-46. https://www.muse.jhu.edu/article/864165.

[39] Qi, Y., Wang, R., Cao, N., & Xi, C. (2022). When a Judicial Mistake Went Viral: The Diffusion of Law in China. China Review, 22, 106 - 73.

[40] See the China Judgment and Order Website at https://wenshu.court.gov.cn/website/wenshu/181029CR4M5A62CH/index.html?cid=012001007, accessed May 31, 2023, which allows case searches by keywords and various case conditions The CJO website provides access to all types of judicial cases in mainland China, but access to large-scale judicial documents and datasets based on them is likely to become increasingly difficult as the CJO website has begun to impose technical restrictions on bulk access to documents.

[41] Supreme People's Court and Supreme People's Procuratorate, Interpretation of the Supreme People's Court of the Supreme People's Procuratorate on Several Issues Concerning the Application of Laws in Handling Criminal Cases of Embezzlement and Bribery, (Adopted by the Trial Committee of the Supreme People's Court at its 1680th meeting on March 28, 2016, and by the 12th Procuratorial Committee of the Supreme People's Procuratorate at its 50th meeting on March 25, 2016, with effect from April 18, 2016) effective from April 18, 2016), Legal Interpretation [2016] No. 9

[42] Wang, G.(2020).Reflections and Prospection on the Legislation of the Criterion forJudicial Sentencing in Connection with Corruption and Bribery Crimesin the Past 70 Years since the Founding of the People's Republic of China,Journal of Yunnan Normal University(Humanities and Social Sciences Edition),52(03). Journal of Yunnan Normal University(Humanities and Social Sciences Edition),52(03),92-101.

	Female:8.8%
Charges committed	embezzlement: 39.8006%.
(Name of crime committed)	bribe: 60.1994%
	Other charges (other than embezzlement and bribery): 15.0%
	None (no academic information): 20.5%
	Illiteracy: 0.2 %
	Primary school: 2.9%
Education level	Junior high school (middle school): 9.8%
	High school/Technical secondary school (high school/secondary school): 13.4%
	Undergraduate/Associate college: 49.3%
	Postgraduate: 3.6 %
Average amount involved	¥369452.06340825
(average amount involved)	
	Make contributions: 8.8%
	Surrender oneself: 43.5%
Sentencing plot	Accessory: 5.1%
	Demand a bribe: 3.0%
	None: 49.2%
	Exemption from Penalty: 13.8%
Verdict	Penal servitude: 4.6%
	Fixed-term imprisonment with suspension: 32.9%
	Fixed-term imprisonment without suspension: 48.4%
	Life imprisonment: 0.0456% (<0.1%)

(1) 91.1% of the accused officials in the dataset are male, accounting for the majority, while only 8.8% are female. There are 6,967 data entries with information on academic qualifications, 0.2% illiterate, 2.9% elementary school, 9.8% middle school, 13.4% high school/secondary school, 49.3% university/tertiary school, and 3.60% postgraduate students, which is related to the improvement of the system of hiring state employees in China, where the public sector has been continuously raising the academic and cultural requirements for the positions of hiring since the reform and opening up of China. The average amount of money involved was 369,452.063,40825 yuan, a rather staggering amount that directly reflects the serious and huge amount of corruption in China's rapid development phase.

(2) In terms of sentencing circumstances, the self-surrendering circumstance accounts for 43.5%, which may be related to China's increasingly sophisticated state supervision system, which creates a strong deterrent effect. Interestingly, in 8.8% of the cases, there will be officials contributing to the detection in order to mitigate the sentence, which indirectly reflects the existence of nepotistic corruption (Weijia Li, Gérard Roland & Yang Xie, 2022).[43]

(3) Accessory roles and active solicitation of bribes(demand a bribe) accounted for a very small proportion of the cases, 5.1% and 3.0 %, respectively. The majority of officials were sentenced to a fixed-term imprisonment, accounting for 81.3%, which included a term

[43] Li, W., Roland, G., & Xie, Y. (2022). Crony capitalism, the party-state, and the political boundaries of corruption. Journal of Comparative Economics, 50(3), 652-667.

of imprisonment without probation and a term of imprisonment with probation. Under China's criminal law, sentences of probation are generally less than three years for lesser offenses. A probationary sentence provides for a trial period during which the offender is required to undergo community corrections and comply with restraining order that may be issued against the offender by the court. Generally, if the offender does not break the law and seriously violate the restraining order issued by the court during the probationary period, the original main sentence does not have to be executed.

Observation of the data reveals that there are 7993 male defendants in the case while there are only 775 females, a serious gender imbalance in the distribution, in order to make the data balanced, we use mean-variance standardization and apply SMOTE oversampling for balancing, and processing, ultimately obtain 12,750 pieces of data.

Finally, in order to better adapt the data to the model, we preprocess the raw data and first numericalize the data with attributes. For example, label coding is used to convert the character variable of male or female in gender to a numerical variable of 0 or 1, which is converted to 1 for males and 0 for females. If the data does not show the gender of the perpetrator, the name usually has a certain gender tendency,[44] Especially in China, there is a longstanding specific cultural phenomenon of commonly used names for males and females, and in general, certain words are specifically used for males or females.[45] Therefore, the gender can be presumed according to the commonly used names of males and females, then eliminate the data that does not show gender. Therefore, it is possible to infer the gender by male and female common names, and then exclude the data that do not show the gender. Gender is predicted using the third-party library NGender, a natural language processing (NLP) tool designed to help users categorize gender in text. Using NGender can easily identify male, female, or unknown gender in text and provide a reliable presumptive result. NGender's accuracy in inferring gender based on Chinese names is generally 82%, and the level of accuracy for our data is shown in Table 2.[46] Similarly, Bribery is labeled as 0, and embezzlement is labeled as 1. Additionally, educational qualifications, other cases, sentencing scenarios, and imposed penalties are also characterized according to the type of Digital transformation.

Table 2. NGender's inferred accuracy

Gender Recognition		
male	Total: 5445	Number of corrections: 5175
female	Total: 505	Number of corrections: 364

7. Experimental Result

In this study, the model is evaluated by four types of indicators : Loss curve, ACC curve, ROC (Receiver Operating Characteristic) curve and Confusion matrix. Among them, Loss curve reflects the change of loss value in model training, which can be categorized into Good fit, Under fit and Over fit according to different performances. ACC curve reflects the change of accuracy rate in model training. Considering the factor of data imbalance, we use the ROC

[44] Mehrabian, A. (1992). Interrelationships among name desirability, name uniqueness, emotion characterist ics connoted by names, and temperament. Journal of Applied Social Psychology, 22(23), 1797-1808. https: //doi.org/10.1111/j.1559-1816.1992.tb00977.x

[45] Xie Yu'e. Personal Names-Gender-Culture-An Examination of the Cultural Phenomena of 'Men's Names' and 'Women's Names'[J]. Chinese Cultural Studies, 2000(01):103-108+3.

[46] NGender Library, https://github.com/observerss/ngender, accessed May 11, 2023 .

curve,which reflects the false positive rate and the true positive rate of a classifier, and the higher the degree of curve convexity, the better the performance of the classifier. The confusion matrix is used to evaluate the performance of the classification model and is used to indicate the gap between the predicted and true values. The confusion matrix has four metrics, which are TP (True Positive), TN (True Negative), FN (False Negative), and FP (False Positive).The larger the values of TP and TN, and the smaller the values of FN and FP, the better the performance of the model is.The horizontal coordinate of the ROC curve is FPR (False Positive Rate), vertical coordinate is TPR (True Positive Rate). AUC (Area Under Curve) is the area under the ROC curve, the larger the value of AUC, the better the performance of the classifier (Fawcett , 2006).[47]

7.1 Gender Classification

First, the study hypothesized that males and females presented significant differences. Further, the study was conducted by training the model for 1000 rounds and recording the results of the loss value for each training, and plotted to obtain the LOSS curves ,as shown in Figure 2, (left), and it was found that as the epoch advances, the training loss and the validation loss are converging, and at the same time the difference between the two curves is getting smaller and smaller, and the model is getting more and more stable.After training 1000 rounds of this model, the accuracy results of each training are recorded, resulting in Figure 2 (right) ACC curve, and it is found that the epoch advances, the accuracy is getting higher and higher until the end of the model's accuracy is close to 71%, and it should be noted that the accuracy of this prediction for the outcome of the criminal penalties is not low, mainly because although China has a fixed amount of punishment according to different amounts of money involved, there is also criminal discretion in each region under the condition of observing the legal sentencing circumstances, and in particular, there are differences in the decisions of the courts between different provinces and cities, and sometimes the judges will give priority to referring to the sentencing results of similar cases in their own city or province before making a decision, but in general, this prediction rate is also in line with the estimation on the regional differentiation of the results of the criminal justice decisions by many other scholars, and the legal sentences guarantee the accuracy of model's prediction. The statutory sentencing ensures the accuracy of the model's prediction. [48]At the same time, we added gender and education background as input features, two factors that judges basically do not consider when sentencing, and it can be seen from below that regardless of male or female, the strongest correlation with the final conviction is still the most important factor when sentencing for the crime of embezzlement and bribe - the amount of money involved in the case, and thus the addition of other factors will affected the

---

[47] Tom Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, Volume 27, Issue 8, 2006, Pages 861-874.

[48] In order to compare the level of accuracy in our sentencing studies, the accuracy rates of some previous researchers in conducting sentencing studies are listed here. Xia Wei (2022) had an accuracy rate of 83. 7% and 72.5% for the Conviction Model of the Crime of Buying Abducted Women and the Conviction Model of the Crime of Buying Abducted Children, respectively; Xia Wei, Research on the Conviction and Sentencing Rules of the Crime of Buying Abducted Women and Children, Journal of Southwest University of Political Science & Law, Apr. ,2022.Vol. 24, No. 2. p139-152.; Wang Weijiu (2022) et al. use XGBoost algorithm to predict the Crime of Illegal Business Operations with Accuracy of 80.02% and 78.46%, Wang Weijiu , Xu Minya, Xu Boxi, Meng Siyu, Wei Zhao, Crime of Illegal Business Operations Sentencing Prediction Based on Text Mining and XGBoost Model, Information Research, No. 9 (Serial No. 299). p20-28.; Nikolaos Aletras (2016) et al. predicted judicial decisions of the European Court of Human Rights with 79% accuracy using natural language processing, Aletras, Nikolaos & Tsarapatsanis, Dimitrios & Preotiuc-Pietro, Daniel & Lampos, Vasileios. (2016). Predicting Judicial Decisions of the European Court of Human Rights: a Natural Language Processing Perspective. peerJ in Computer Science. 2016. 10.7717/cs.93.

accuracy of the model,but basically accord with the actual situation of China's criminal justice.

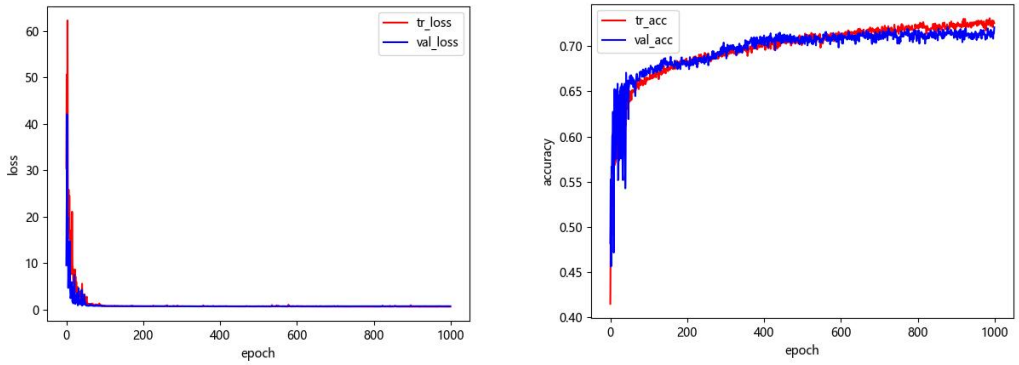


Figure 2. LOSS curve vs. ACC curve

Meanwhile, we construct the confusion matrix to obtain the results as shown in Figure 3(left), as shown in the figure, it can be seen that the TP (true class) is 817, the TN (true negative class) is 1406, the FN (false negative class) is 79, and the FP (false positive class) is 257, and it can be seen that the model's accuracy of the classification performance index is 72.078% by the calculation.

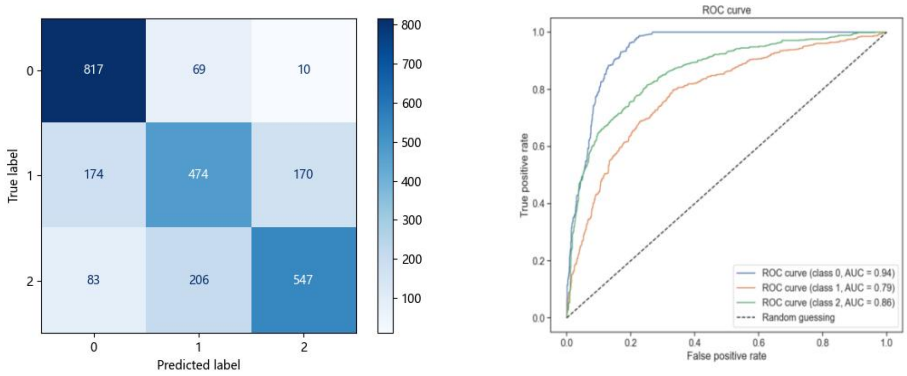


Figure 3. Confusion matrix and ROC curve

In the ROC curve, the AUC represents the area under the ROC curve, which can evaluate the performance of the model. TPR represents the probability of being able to classify a positive case correctly, and FPR represents the probability of misclassifying a negative case as a positive case. As can be seen from the Figure 3(right), the AUC values of the DNN model are 0.94, 0.79 and 0.86 for the different trial outcome classifications, 0.876 for Random Forest , 0.558 for Logistic Regression, and 0.810 for LightGBM, so the DNN model is the most effective as well as the best performer.The DNN is able to classify male and female more accurately.

7.2 Difference analysis

Dividing the females and males into two groups, assigning the same variables and analyzing the differences, the results can be obtained as in Table 3. From the table, it can be seen that the p-value of education is less than 0.05, which presents significance and the

difference of these variables between the two samples is significant. While the p-value of amount, other cases are greater than 0.05 and cannot reject the original hypothesis, the difference of these variables between the two samples is not significant. At the same time, it can be obtained that the magnitude of importance of the impact of the characteristics of different variables on the results in the study. Accordingly, this paper concludes that despite the large difference in sample size between males and females, the fact that males and females can be accurately categorized implies that the characteristics of corruption of different genders show some differences.

Table 3. Table of differences between groups

	Label	education attainment	sum of money	Other cases
t-statistic	-1.430	-3.252	-0.136	-1.877
p-value	0.1528	0.0012	0.8915	0.0605
(number of) degrees of freedom (physics)	8766.0	8766.0	8766.0	8766.0

The study used a male sample (n=91.1%) to make predictions to determine if the model could predict what type of characteristics a male sample possessing would have a higher rate of corruption.

First, the deep neural network model is trained using training samples. The study plots the Loss curve, by training the model for 1000 rounds and recording the loss values, as the epoch advances, the training loss and validation loss are converging, and the model performs as Good fit. Meanwhile, the model is trained for 1000 rounds and the accuracy is recorded to get the ACC curve as shown in Figure 4., as epoch of advancing , the accuracy rate is getting higher and higher.

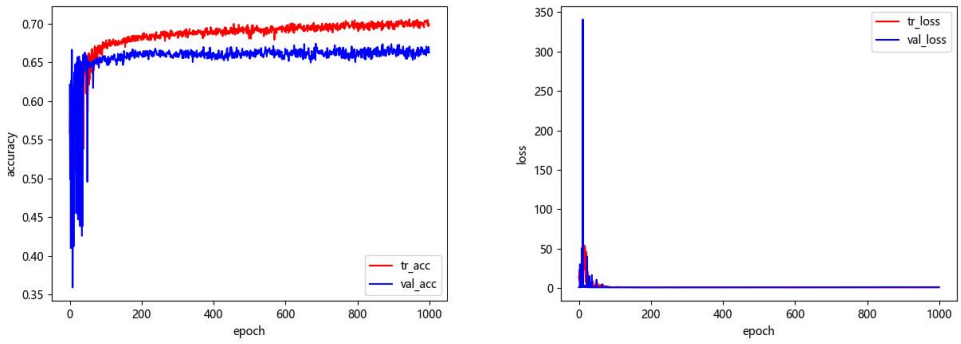


Figure 4 LOSS Curve vs. ACC Curve for Predicting Male Prison Sentences

Meanwhile, we construct the confusion matrix to evaluate the model performance. From Figure 5., it can be seen that TP (true) is 8, TN (true-negative) is 1290, FN (false-negative) is 62, and FP (false-positive) is 16, and it can be known that the accuracy rate of the model is

66.279% by calculation. And in order to reflect the false positive rate and true positive rate of the classifier, we use the ROC curve. Through the experiment, it can be seen that Testing AUC is 0.4771828620951143, while the curve is convex at a high level, and the model performance is good. Based on the above, it can be concluded that some of the influential features proposed in the study constitute a significant influence on the corruption of male samples and have high accuracy in predicting male corruption.

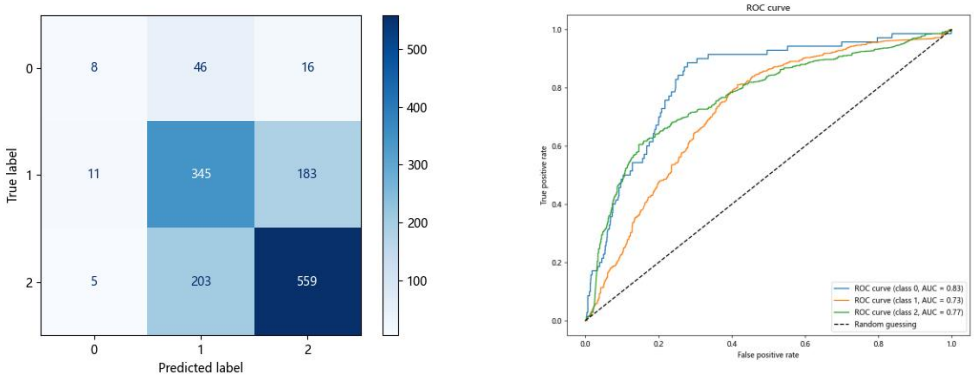


Figure 5. Confusion matrix vs. ROC curve for predicting male prison sentences

As in the study above, predictions were made using a female sample (n=8.8%) to determine if the model could predict what type of characteristics a female sample possessing would have a higher rate of corruption.

Firstly, training the deep neural network model is carried out 1000 times and the loss value is recorded and the Loss curve is plotted. It can be seen through the curve that as the epoch advances, the loss value is decreasing as a whole. At the same time, the ACC curve is plotted by training the deep neural network model 1000 times and recording the accuracy, and it can be seen through the ACC curve that as the epoch advances, the overall accuracy is increasing. Thus the model prediction has low loss values and high accuracy with ideal results.

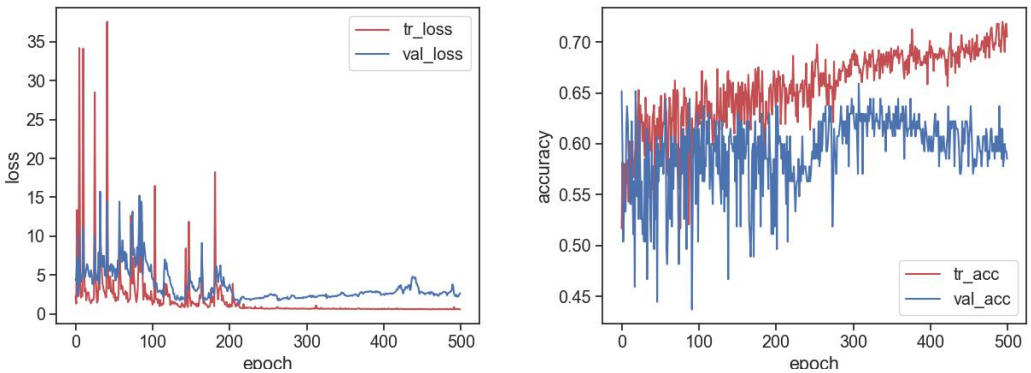


Figure 6. LOSS Curve vs. ACC Curve for Predicting Female Prison Sentences

Meanwhile, the confusion matrix of the model for predicting women's imprisonment was constructed, which can be seen in Figure 7, TP (true)is 1, TN (true negative) is 122, FN (false negative) is 7, and FP (false positive) is 5. The accuracy of the model can be calculated to be 59.848%, which is a desirable result. In this group of study the Testing AUC in ROC curve is

0.766 and again the performance is good. Based on the above results, we can conclude that the characteristics of the variables derived from the study are good predictors of the likelihood of female corruption offenses.

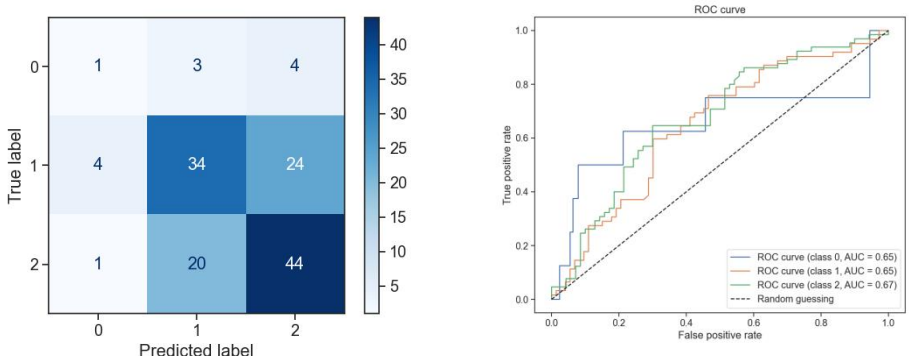


Figure 7. Confusion Matrix vs. ROC Curve for Predicting Female Prison Sentences

7.3 Gender Differences in Corruption Behaviors Based on Partial Dependence Plots

While we have used difference analysis to understand the contribution of different features to the results and gender differences, Partial Dependence Plots (PDPs) can further enhance our confidence in the model's predicted outcomes and our understanding of differences.

The PDP plots a function of a particular characteristic by holding other characteristics constant, showing how changes in that characteristic affect the output. For gender differences, we can use PDP to understand how model predictions vary with different features and how model predictions show different tendencies with different features. This can help us determine which characteristics are the most influential on gender differences and better understand why the model makes these predictions.The PDP can also help us detect whether nonlinear relationships or interactions are occurring in the model that may not be detected by ordinary statistical analysis. By using PDP, we can observe how model predictions vary with different characteristics, and thus better assess whether our models are plausible in order to enhance their interpretability (Christoph Molnar, 2023).[49]

From the partial chart, we basically have the following conclusions: (1) As shown in Figure 8, the larger the amount of money involved in the case, whether it is a man or a woman, the smaller the distribution of the people involved in the case, based on the sentencing of the larger the amount of money, the heavier the sentencing characteristics, which indicates that anti-corruption deterrence has gradually formed a long-term effective mechanism. In the field of anti-corruption policy, China has introduced a new approach to the traditional inspection system, focusing on the power center of the 'top banana', supplemented by supervision by the National People's Congress, social supervision, and public opinion supervision[50] , while maintaining the independence of the inspection agency. At the same time, the scope of anti-corruption inspection 'shift to primary level'[51] . However, although the high pressure of

[49] Christoph Molnar (2023). Interpretable machine learning : a guide for makingblack box models explain able. section 8.1, <https://christophm.github.io/interpretable-ml -book/pdp.html>accessd.
[50] Bamboo Lijia. What kind of institutionalization is needed for 'inspections against corruption'. People's Daily, October 15, 2013, p. 5
[51] Guo Xingquan. Theory and Practice of Integrity Building in China. p420-423

anti-corruption can play a deterrent role for most officials, China also investigates and handles dozens of cases of 'small officials and huge corruption' every year. In our opinion, the rate of such cases has a lot to do with the existence of life imprisonment sentence, that is, the 'Criminal Law Amendment (IX)' stipulates that for the embezzlement and bribe in a huge amount, and make the interests of the state and the people suffer particularly significant losses,..... , life imprisonment, no sentence reduction, probation'[52]. Life imprisonment sentence, in the general trend of domestic and foreign calls to reduce the death penalty, is considered an alternative to 'immediate execution' in corruption cases. However, according to a study by Samuel Chu, this alternative mainly represents 'a political gesture' rather than a significant deterrent against corruption[53] .

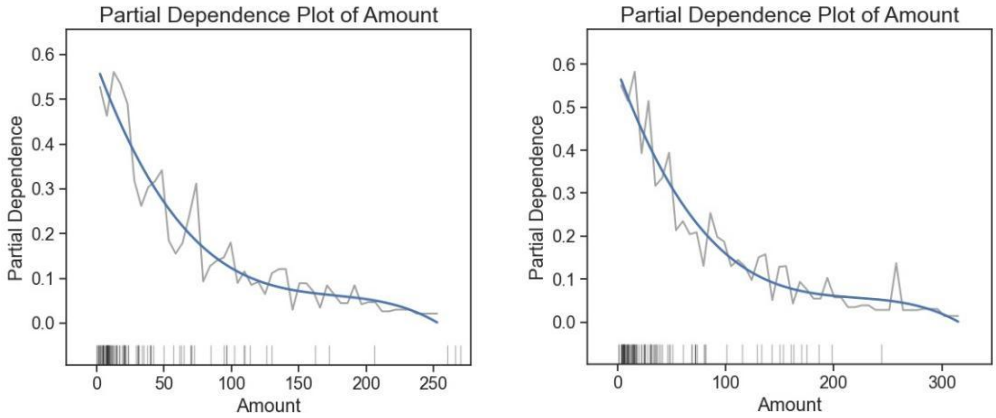


Figure 8. Partial dependence of the amount of money involved in the case for corrupt male (left) and female (right) officials

As shown in Figure 9, both genders, in terms of education level, show that the most corrupt officials are distributed in the cultural level of high school/secondary school, which is in line with the study of Jin Honghao (2019) and other studies, which concluded that the number of crimes committed by leading cadres with higher education and the number of crimes involved are more. However, according to the study of Nie Huihua (2015), from the data of corruption about education background of all departmental officials, master's degree accounted for 42%, and bachelor's degree accounted for 37%, compared to the corruption of the study in general, the high-level corrupt officials show the characteristics of higher education concentration[54] . Of course,we believe that the emergence of this characteristic is closely related to the promotion policy of 'elite power' in the stage of Chinese politics[55] ;

---

[52] Shixin, Tang Weijun, Chen Luning. Refining Elastic Provisions to Guide Judicial Practice. In Procuratorate Daily, September 3, 2015, page 3

[53] Chu, Sai-Shik. Strict but not severe: designing policy ideas for criminal law revision[J]. Journal of Peking University (Philosophy and Social Science Edition), 1989(06):101-109+94.

[54] Nie Huihua. Innovating the Discipline Inspection and Supervision System to Curb Corruption of the 'Handful of Hands'. pp. 25-82.

[55] Ezra F. Vogel. Deng Xiaoping and the Transformation of China. p398.

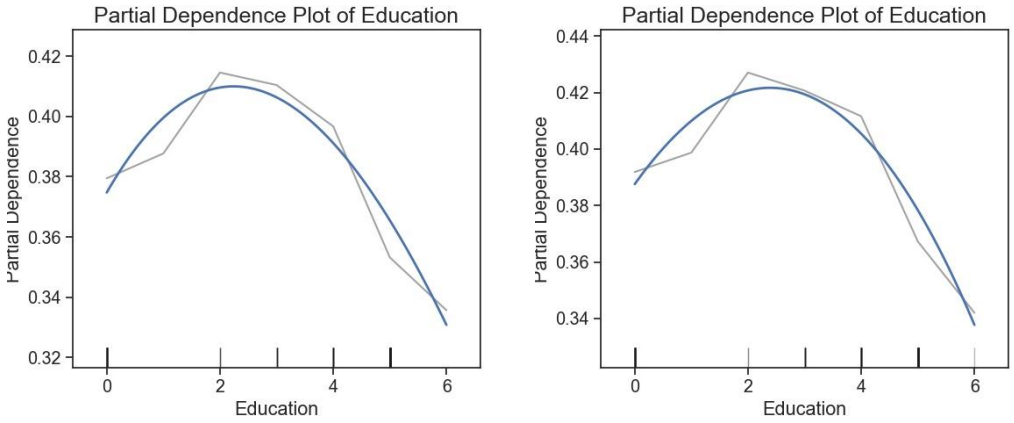


Figure 9. Partial dependence plot of educational attainment of corrupt male (left) and female (right) officials

We categorized the types of cases into three types based on sentencing outcomes, distributed across the three types of situations using PDP for a more detailed examination, and the results, as shown in Figures 10-13, show that both genders exhibit a very high degree of similarity in performance and do not differ significantly. However, the corruption characteristics common to both genders still merit detailed discussion.

The dependence of the three types of penalization outcomes on different amounts of corruption is shown in Figures 10-11. When the amount of the crime is very small, the likelihood of being sentenced to detention or control is very high, but overall, the likelihood of being sentenced to a lighter free sentence of detention or control is very low for corruption, and the curve therefore "falls off a cliff". This feature is in line with China's general tone of "full and strict governance over the Party" for corruption cases, which plays a good deterrent role in punishing crimes and warning officials; and when the amount of crime reaches a certain level (usually 30,000-50,000 yuan), the larger the amount of the crime, the lower the possibility of obtaining probation while being sentenced to fixed-term imprisonment ,and the S-shaped upward trend is shown. The key to this feature lies in the definition of "probation" in corruption cases. The abuse of probation has been highly discussed in Chinese society in recent years, as some people believe that the existence of probation in corruption cases means that "bureaucrats shield one another" and "probation means not going to jail"[56] . At the national level, the work report of the Supreme People's Court (SPC) adopted by the National People's Congress (NPC) has also received an increasing number of negative votes[57] . Based on this background information, we can also correspond to Figures 10-11, in the probation sentence based on the impact of the amount of crime, the general trend of the curve is smooth, but the small fluctuation is large, that is to say, after confirming the amount of the specific crime committed by the official and setting the general tone of the official's crime, the judge still has great discretion considering "Whether there is repentance", "whether there is a risk of reoffending" and other subjective evaluation criteria. Therefore, how this discretion is exercised will become an important factor affecting the governance of corruption in China, as well as the trend of related social opinion.

---

[56] Yao, Jianlong. Criminal Law Issues in Social Change. p377-379.2019.6.
[57] Huang Yong and Jin Sheng, "Township Cadres Set Up Cells to Detain Two Hundred People," in China Youth Daily, Nov. 24, 2000

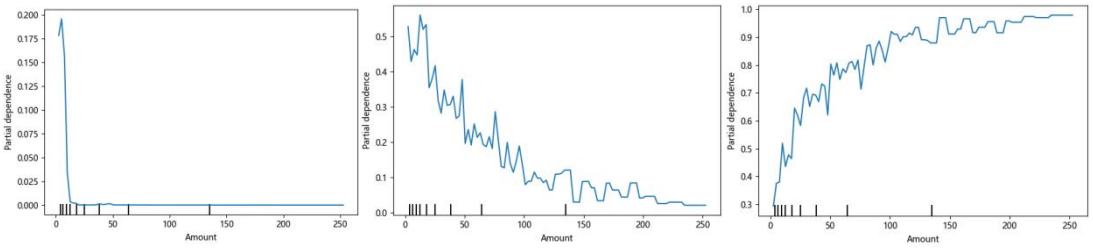


Figure 10. Partial dependence of amounts for male officers based on three types of punishment outcomes: detention or control (left), suspended sentence with imprisonment (center) and suspended sentence with imprisonment without probation (right)

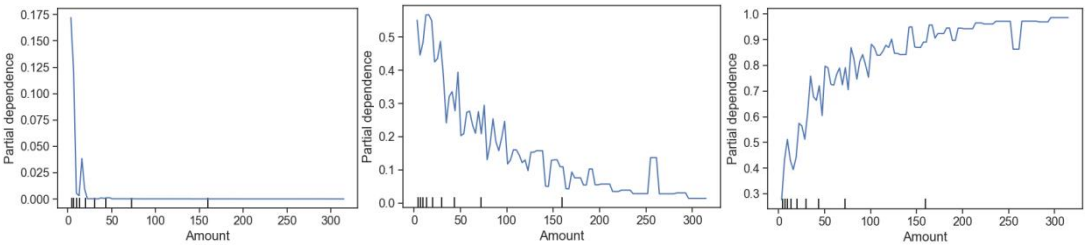


Figure 11. Partial dependence of amounts for female officials based on three types of punishment outcomes: detention or control (left), suspended prison sentence (center) and suspended prison sentence without probation (right)

The level of education of an official has a more diverse impact on the sentence than the amount of the crime. As shown in Figures 12-13, for criminals sentenced to detention or control, "primary school" and "undergraduate/associate college" education have higher probability of being sentenced to detention or control, i.e., they are at the "peak"; "illiterate", "junior high school" and "high school/secondary school" are relatively less likely to be sentenced to detention or control and at the "trough". Besides, "postgraduate" and "PhDs" degrees, the likelihood of being sentenced to a lighter liberty is significantly lower, showing a "straight-line decline". In the case of fixed-term imprisonment, it can be clearly seen that, similar to the case of detention or control, the likelihood of "postgraduates" and "PhDs" being sentenced to probation is significantly lower, and the likelihood of having no probation is higher; while in the case of "illiterate" to "undergraduate/associate college", the likelihood of being sentenced to probation increases and then decreases with a maximum at "junior high school".

The existence of the above characteristics plays a significant role in analyzing and studying the impact of academic qualifications/educational backgrounds on corrupt officials. First of all, we believe that the common feature of relatively heavier penalties for highly educated offenders, with few suspended sentences, has much to do with China's periodic political policy of "elite power". During Deng Xiaoping's administration, China's official promotion policy was boldly shifted from the "class composition theory" to the "elite rule theory". As a result, since the 1980s, a large number of highly educated graduates have entered important central and local organizations to hold state power and have unique advantages in promotion.Consequently, it is reasonable to assume that highly educated officials usually occupy high positions of power and hold the core of departments, and therefore are more likely to commit corruption in larger amounts and receive heavier sentences. Secondly, in response to the question of the higher likelihood of being

sentenced to detention or control for "primary school" and "undergraduate/associate college", we believe that offenders at these two stages are often in the "middle ground", i.e.,unlike the "illiterate" and "junior high school" groups, who lack entry and promotion possibilities, or the "postgraduates" and above groups, who have access to more specialized promotion policies. As a result, these two groups in the middle ground are relatively less likely to commit serious corruption and more likely to be sentenced to control or detention. As for the problem of the extreme value of "junior high school" education, we suspect that it is closely related to the special characteristics of the "junior high school" group in Chinese society.

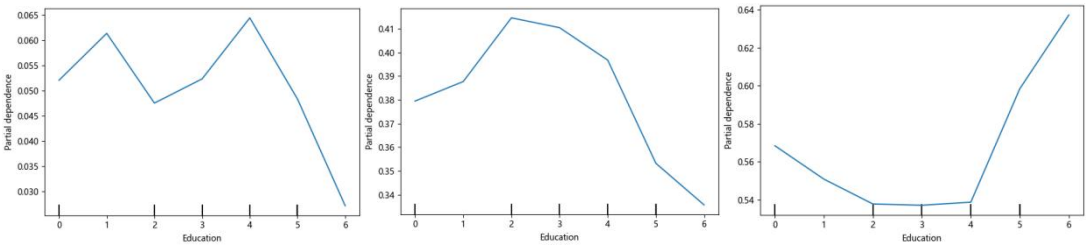


Figure 12. Partial dependence of educational attainment of male officials based on three types of punishment outcomes: detention or control (left), suspended prison sentence (center) and suspended prison sentence without probation (right)

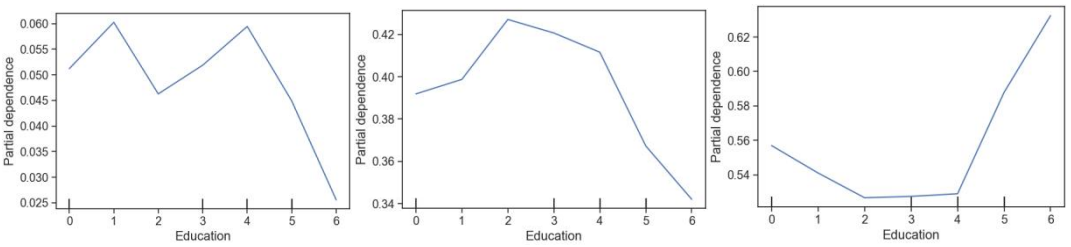


Figure 13. Partial Dependence of Educational Attainment of Female Officials Based on Three Types of Penalty Outcomes: Detention or Control (left), Probation with Imprisonment (center) and Imprisonment without Probation (right)

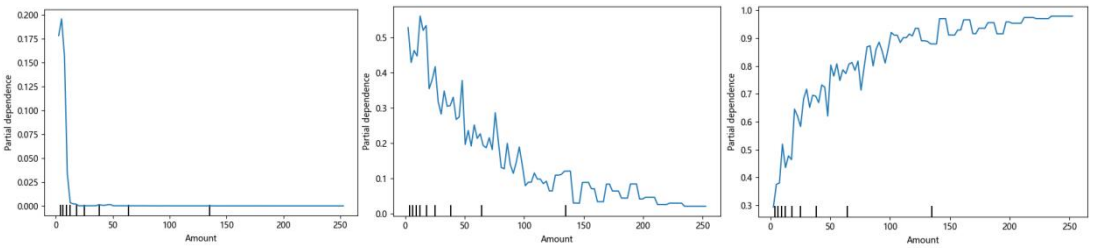


Figure 10. Partial dependence of amounts for male officers based on three types of punishment outcomes: detention or control (left), suspended sentence with imprisonment (center) and suspended sentence with imprisonment without probation (right)

8. Conclusion

This paper utilizes deep learning to analyze gender differences in corruption in China, which fills the gap in the existing literature in terms of methodology and research content. At the same time, with the gradual deepening of China's judicial data disclosure, CJO and other adjudication document platforms and case filing platforms will provide a large number of rich data sources for the study of corruption and crime, so as to facilitate in-depth innovation in

data for other topics. The findings of this paper on gender differences in corruption in China are of great significance for the behavioral analysis of corruption and the innovation of research methods. We have created a new dataset based on the referee documents, which makes the data source for studying this issue more objective research material, and the data come from the facts and conclusions of the investigation of the judicial institutions, which has a high credibility and authority. Deep learning has shown notable performance in categorizing both male and female corruption cases, which provides an effective method for strategic governmental decision-making in the fight against corruption.

The experimental results in this paper fulfill and answer the three research ideas presented above. By dividing the experimental dataset using DNN models and performing nonlinear processing, deep learning was able to successfully classify males and females in corruption cases. Simultaneously, this study predicts male corruption cases with higher accuracy and smaller loss values, which gives more excellent results compared to females. Of course, the deep learning analysis in general maintains high accuracy in both sets of experiments, and the results are more excellent compared to previous random forest, logistic regression and LightGBM models.

Our study draws a cautious recommendation for anti-corruption strategies: in public institutions where power is highly centralized and networks of relationships are dense, gender policy adjustments may not be a good choice for anti-corruption considerations. The reason is that, at least according to the conclusions of the analysis based on the limited data, there is no significant difference in the level of corruption between male and female officials in China, which is similar to the results of Gao, Bo and Miao, Wenlong's (2013) empirical analysis of 897 corruption-related cases, and to some extent to the results of Zhao, Bin's (2014)[58] social survey of 2,643 state-owned enterprise (SOE) employees, which showed no significant difference in perceptions of corruption between different genders on the types of corruption. The lack of significant differences in perceptions is echoed to some extent. The gender-specific theory that women are more upright is not supported in this study because even though women make up a very small percentage of the official workforce, their performance in embezzling and bribes is not very different from that of men. Also, contrary to the institutional-specific insistence that the more marginal women are, the less corrupt they are, a small number of female officials reciprocally can still produce as much corruption as men.

One possible explanation is that highly relational bureaucratic networks of corruption can virtually erase the behavioral factors brought about by gender-specific traits, or that thick bureaucratic networks form male-dominated political cultures and behavioral habits, with political interests and bureaucratic collective ends replacing gender motivations. For example, Feng Junqi (2010) found in his investigation of the officialdom in Zhongxian County that 'in the jungle of male leading cadres in Zhongxian County, it can be clearly felt that some female leading cadres, especially those above the level of the head of the main section, are shrewd, bold, and capable, but also splashy, and fiery, and sparring with the male leaders over drinking and being tactful, with varying degrees of masculine temperament and such image and temperament are often more recognized and accepted by the male leadership cadres.

---

[58]Zhao Bin. A Sociological Study on the Cognition of State-owned Enterprises Employees on Anti-Corruption and Integrity Promotion [D]. Doctoral dissertation, Wuhan University, 2014

Forgetting gender and becoming part of the male jungle is both a growing experience and a growing concern for some female cadres.' The process of marginalization plays a screening role in which the process of rent-seeking requires the collaboration of hidden relationships to be accomplished, and therefore greater care is taken in selecting women for entry into a web of corruption. This also creates a situation where women are not less tolerant of corruption or more incorruptible, but rather lack criminal opportunities due to the inaccessibility of political resources.

It is important to emphasize that this phenomenon is still the result of a system of inequality, and the low number of female officials due to artificial marginalization is a global phenomenon (Mona Lena Krook, 2010)[59] , which is related to the gender bias in the selection and appointment of officials, whereby male corruption is a way of getting into a network of corruption, first by obtaining a posting (and possibly also by obtaining a posting precisely through the network of relationships), and then by integrating into a network of relationships (which is not necessarily networks of corruption), which can morph into networks of communities of corruption. While women first have to more or less overcome their marginalization, this effort begins when they seek a public position, and the acquisition of a position does not necessarily mean that they are integrated into a network of relationships, much less a network of corruption. Thus it can be seen that women go through more screening than male officials to enter a hidden corrupt network, and it can also be argued that the screening of women by the corrupt network ensures that women who are granted a license to enter the network can be just as corrupt as men, and that once in the privileged class, men and women are equal, or there is only one gender identity permitted to exist in this class - the privileged males.[60] This also suggests that the marginalization and risk aversion perspective theory is explanatorily incomplete and needs to be extended and refined by theoretical and empirical research methods (Russil Durrant, 2018)[61] .

In addition, according to the prediction results, the two experimental groups of different genders are equally significantly different in terms of the characteristics that influence corruption cases. Specifically, male corruption cases are greatly influenced by the outcome of probation, while less influenced by variables such as age, education, and other cases; while in female corruption cases, probation likewise plays the role of the greatest influence, but at the same time the existence and number of corruption cases also constitute a greater impact on female corruption. At the same time, the three variables of education, age, and sentencing circumstances have a significantly larger share in the degree of influence of female corruption. Therefore, this study provides experimental results for analyzing the causes of gender differences in corruption behavior and the extent to which the variables influence corruption behavior differently by gender.

Finally, this study realizes the innovation of research method. In the study of gender differences in corruption, most people apply quantitative or logistic regression methods to conduct empirical research and analysis, but lack the research method of constructing

---

[59] Krook, M.L. (2010), Why Are Fewer Women than Men Elected? Gender and the Dynamics of Candidate Selection. Political Studies Review, 8: 155-168. https:// doi.org/10.1111/j.1478-9302.2009.00185.x

[60] Acker, J. (2006). Inequality Regimes: Gender, Class, and Race in Organizations. Gender & Society, 20(4), 441-464. https://doi.org/10.1177/ 0891243206289499

[61] Russil Durrant (2019) Evolutionary approaches to understanding crime: explaining the gender gap in offending, Psychology, Crime & Law, 25:6, 589- 608,

accurate models based on deep learning. Therefore, this study provides a corresponding model example and prediction direction for the gender difference research on corruption, and provides a corresponding reference for the rest of behavioral analysis research and deep learning prediction research.

Deserving Recipients: Defining Copyright and Methodology of AIGC through the Lens of Labour Theory

Runpeng Fu[1]

Jingxin Yang[2]

Mengchi Cai[2]

Heshu Wang[2]

Xinchen Wang[*]

keywords: AI-generated content; Intellectual property rights; Labour theory; Locke; Marx.

Abstract:

The copyright dilemmas surrounding artificial intelligence (AI)-generated works are often approached from the perspective of labour theory of property, which is inadequate in providing a comprehensive solution to the problem. This article proposes a revised labour theory of intellectualintellectual property, drawing on Marxist labour theory, and presents a comprehensive methodology for determining artificial intelligence generated content (AIGC) copyright ownership based on the three elements of labour and Marx's theory of commodity exchange. This new approach not only offers an original theory based on the labour theory of value, but also provides a novel method for determining the copyright ownership of artificial intelligence in complex production chains.

Introduction

The issue of copyright ownership of artificial intelligence generated content (AIGC) has been a subject of prolonged academic discourse. Current solutions to this issue mainly include the "Simulated Authorship Theory", the "AI Rights Subject Theory", the "Non-Copyright Theory", and the "Public Domain Theory." This article employs Marxist labour theory to provide a legal approach to the issue of AIGC's copyrightability, centring around human labour. Specifically, this article offers an interpretation of AIGC's copyrightability based on the theory of human authorship and explores a methodology for determining the legitimacy of AI's simulated authorship. The aim of this article is to provide a comprehensive and original theoretical framework to address the confusion surrounding the copyright ownership of AI-generated works.

The Theoretical Core of Copyright for AIGC: A Recognition of Labour's Expression

Marxist labour theory and Locke's labour mixing argument

Both Marx and Locke emphasised the fundamental importance of labour as it pertains to an individual's survival and development within society.[1] The Marxist Theory of Labour (MTL) is built on a philosophical amalgamation of hunting and gathering, agriculture, handicrafts, manual labour, and large-scale machine industry. Consequently, it offers valuable insights into comprehending labour production in the present-day post-industrial era.[2] While it is not certain whether Locke acknowledged intellectual labour, some scholars of the Lockeian school concur with Marx in recognising the existence of intellectual labour and utilising labour as a theoretical foundation for the argumentation of intellectual property rights.[3] Bryan Cwik's theory of productive capacity, as presented in his 2014 work, aligns closely with Marx's views on intellectual labour. The theory highlights three key ideas: (i) the legitimacy of property is demonstrated through labour, (ii) control over intangible intellectual labour is secured solely through property rights, and (iii) market conditions often dictate the role that intellectual property rights play. However, some scholars argue that Marx's theory cannot substantiate the legitimacy of property.[4] [65] However, this

[*] We would like to deliver our acknowledgment to reviewers and editorial board of EIPR. This article is the research outcome of the Institute for Studies on Artificial Intelligence and Law, Tsinghua University. Runpeng Fu is an undergraduate student at the School of Law, Sun Yat-sen University. He is concurrently conducting research internships at the SYSU Research Institute of Law-Ruled Society Construction and the Institute of Artificial Intelligence and Rule of Law, Tsinghua University. Xinchen Wang is a Research Assistant at Institute of Intellectual Property, University of Science and Technology of China；Mengchi Cai is a Research Assistant at the Institute of International Law, Central South University. Heshu Wang is a Research Assistant at School of Government, University of International Business and Economics. Runpeng Fu is the first author of this article Xincheng Wang is the corresponding author, and Mengchi Cai, Heshu Wang, and Jingxin Yang are the second authors.All authors come from China. If has any question please contact with Xinchen Wang: University of Science and Technology of China, 96 Jinzhai Road, Hefei, OH 230026, China. E-mail addresses: wangxinchen@mail.ustc.edu.cn (X. Wang).

[1] C. Kwok, "The Normativity of Work: Lockean and Marxist Overlapping Consensus on Just Work" (2020) 26(3) *Journal of Human Values* 228–237.

[2] S. Sayers, "The Concept of labour: Marx and His Critics" (2007) 71(4) *Science & Society* 431–454.

[3] A. Mossoff, A. (2012). Saving Locke From Marx: The labour Theory Of Value in Intellectual Property Theory. Social Philosophy and Policy, 29(2), 283-317. doi:10.1017/S0265052511000288.

[4] B. Cwik, "Labor as the Basis for Intellectual Property Rights" (2014) 17 *Ethic Theory Moral Practice* 681–695, available at: *https://doi.org/10.1007/s10677-013-9471-y.*

[5] Drahos, P. (1996). A Philosophy of Intellectual Property (1st ed.). Routledge. https://doi.org/10.4324/9781315263786，111-137.

kind of perspective is considered one-sided, as *Capital* unequivocally states that "[6]the private ownership of production resources by labourers is the foundation of small production, which in turn is essential for the development of social production and the expression of workers' individual personalities". But"[7]the personal and scattered means of production are transformed into the social and accumulated means of production, thereby transforming the small property of the majority into the large property of the few. The vast masses of people are deprived of land, livelihood, and tools of labour. The terrible and cruel deprivation suffered by the masses forms the prehistory of capital." Therefore, "[8]the private ownership acquired through one's own labour, which is based on the combination of independent labourers and their conditions of labour, is excluded by capitalist private ownership, which is based on the exploitation of others but appears to be free." According to textual evidence, it is apparent that Marx acknowledges the correlation between labour and legitimate ownership,, especially in relation to the labour of workers. The key issue, however, lies in the ownership of the means of production, which implies that if the labourer has ownership over the means of production for their labour, the unequivocal outcome of their labour belongs to the labourer. It is undoubted that in capitalist societies, a significant amount of the means of production are typically not owned by the labourers, resulting in a lack of ownership over the labourer's own labour outcomes.

Locke's labour Mixing Argument fails to provide a comprehensive analysis of the labour process, whereas MTL provides a precise description of the tangible and intangible aspects of the labour process and its components, which can help us understand intellectual labour. Lockean labour theories often argue that property originally comes about by the exertion of labour upon natural resources, mainly demonstrated by the labour-mixing theory. However, Locke cannot prove "why one should receive what they have mixed their labour with, rather than losing their labour".[9] Nozick offered a counterexample of pouring tomato juice into the sea to contest Locke's labour Mixing Argument. This example raises the question of "[10]if I own a can of tomato juice and spill it in the sea so that its molecules (made radioactive, so I can check this) mingle evenly throughout the sea, do I thereby come to own the sea, or have I foolishly dissipated my tomato juice?". In response, from the perspective of MTL, it can be argued that Nozick's action does not qualify as labour, as Engels had previously stated that the "[11]Nature, like labour, is also a source of use value.",that is "[12]labour is the father of wealth, and land is the mother". Labour can only be conducted when individuals exercise their sovereignty over the natural resources required for production. Thus, individuals who pour tomato juice into the ocean cannot claim ownership over the ocean. In fact, when Nozick raised this question, he implicitly established a rule regarding ownership, which assumes that when people encounter this inquiry, they would assume that the person who performed the action has ownership over the tomato juice but not over the ocean. We could also provide a straightforward response to this question by saying, "The tomato juice is yours, but the ocean is not." Therefore, you cannot own the ocean.The imprecise nature of this theory has rendered Locke's property theory, which is based on knowledge-based intellectual property, especially fragile in the context of intricate production relations. Whether it pertains to traditional copyright ownership or the current academic debate surrounding the AIGC issue, Locke's property theory is unable to satisfactorily resolve these dilemmas.

In contrast to Locke's viewpoint, MTL argues that labour is the foremost justification for ownership and does not believe that property rights exist independently on their own merits. As a result, MTL provides unwavering support for the labour ownership of workers.[13] We further discuss how people, after engaging in intellectual labour, have their own labour results within the theoretical realm of intellectual property, based on MTL. For someone engaged in physical labour, if they produce some tangible products and someone else takes them without payment, most countries' laws will punish the person who takes them for misappropriation, in order to protect the results of physical labour. However, the use of intellectual results is "non-exclusionary" (sometimes referred to as non-competitive).[14] This is due to the characteristics of the knowledge object itself. As an example, we can consider a salaried employee whose livelihood depends on their writing. In the event that their work is replicated without appropriate remuneration, this would amount to a type of labour expropriation. Despite the fact that such expropriation may not manifest as an overt or heinous act in practice, it nevertheless necessitates extensive safeguarding. When the topic of AI content production is discussed, the advanced intelligence and complexity of

[6] Karl Marx & Friedrich Engels, "Capital: A Critique of Political Economy," in Collected Works of Marx and Engels(Chinese version), vol. 23, People's Publishing House, Beijing, 2006, p. 830.
[7] Ibid.
[8] Ibid.
[9] E.C. Hettinger, "Justifying Intellectual Property" (1989) 18(1) *Philosophy & Public Affairs* 31–52, available at: *http://www.jstor.org/stable/2265190.*
P. Drahos, *A Philosophy of Intellectual Property*, 1st edn (Farnham: Routledge, 1996), available at: *https://doi.org/10.4324/9781315263786.*
[10] R. Nozick, *Anarchy, State, and Utopia* (New York City: Basic Books, 1974), p.174.
Karl Marx and Friedrich Engels, "Capital: A Critique of Political Economy", in *Collected Works of Marx and Engels(Chinese version)*, Vol.23, People's Publishing House, Beijing, 2006, p.830.
[11] Karl Marx and Friedrich Engels, "Criticism of the Gotha Programme", in *Collected Works of Marx and Engels(Chinese version)*, Vol.19 , People's Publishing House, Beijing, 2006, p.15-35.
[12] William Petty, "The Political Anatomy of Ireland"(Chinese version), translated by Dongye Chen et al. (Beijing: Commercial Press, 1978), p.66.
[13] Matthew T. Huber, "Value, Nature, and labour: A Defense of Marx" (2017) 28(1) *Capitalism Nature Socialism* 39–52, available at: *https://doi.org/10.1080/10455752.2016.1271817.*
[14] Ibid.at 4.

machines can lead to overlooking the contribution of human labour. In the *Zur Kritik der PolitischenÖkonomie*, Marx argued that the automation of machines by technicians and scientists through their "general scientific labour" would alter the form of labour that is based on workers' experience.[15] This transformation also applies to intellectual labour, where intelligent machines make it possible for creators to simply operate and supervise them.[16] Nonetheless, it is important to note that "[17]even the most sophisticated machines today are simply loyal agents of their human designers or users". If AI is perceived as a tool for creativity, then it can be classified as a type of labour material in MTL since, regardless of their level of intelligence, machines can only serve as "loyal agents" of their users".[18]

Insights and Innovations in MTL

MTL abstracts the process of labour from a continuous factual state, creatively summarises the three elements of the labour process: (i)purposeful activity or labour itself, (ii)the Labour-Object-Product of the Labour Process, as well as (iii) labour and materials.[19] This summary is highly objective and precise. It describes the scenario of a carpenter procuring a quantity of timber with the intention of using it to construct chairs. In the process, he utilises various tools, such as hand saws, hammers, and electric drills, which serve as his instruments of labour, while the wood itself is the object of his labour. According to MTL's perspective, "art represents the pinnacle of creative endeavor; it constitutes an unfettered, imaginative pursuit and represents the most elevated form of labour".[20] The activity of writing for an author always has a definite purpose, even if the author claims that their work was completed without a clear objective. According to MTL's definition, purpose is closely linked to the activity itself. The writer's stated purpose is merely a minor goal within the overarching goal of writing, which is the creation of a literary work. The writer's aim is to find inspiration for their writing, although most authors rely on the intellectual accomplishments of their predecessors for inspiration. The autonomy of artificial intelligence is contingent on human autonomy, and the "purposefulness" of human behaviour is carried and expanded by machines, thereby objectifying this purpose into a new material existence.[21] The worker is driven by the original purpose to seek labour materials and labour objects, and to design products according to their own will, thereby owning the ownership of their labour. Although Marx did not specifically examine intellectual labour, it can be seen from the above discussion that the fact of intellectual labour is not detached from his conclusions.

Marx also examined the process of commodity exchange. Starting from the exchange value and use value of commodities, he believed that only through the process of exchange could the worker obtain the value of the commodity he sold, including his labour value, while the buyer of the commodity obtained the use value of the commodity.[22][23]The exchange of knowledge commodities also follows this logical paradigm. According to Hettinger's view, labour alone is insufficient to justify an author's claim to the market value of their work, as the work builds upon the contributions of predecessors. Nonetheless, it is essential to acknowledge that authors also incur costs for the labour and materials used in creating their work. They often purchase the intellectual contributions of their predecessors through books or education, thereby compensating for the labour of their predecessors. This perspective helps to clarify the tragedy of Vincent van Gogh, which exemplifies the unfulfilled desire to exchange his work. Despite the pioneering post-Impressionist elements in his work, his contemporaries failed to recognise its value, and he was consequently unable to obtain even basic necessities for living, ultimately living in poverty until his death. It was only after his passing that people recognised the immense artistic value of his work and were willing to exchange it at high prices.

Therefore, for knowledge products, labour input should accelerate reasonable and improved copyright protection.[24] The generation of creativity and originality occurs during the process of labour, with variations in the degree of content within each instance of labour. Furthermore, there exist numerous intellectual labours which are highly repetitive and paradigmatic, such as news writing which necessitates the inclusion of details pertaining to time, place, and characters involved. Despite this, it is reasonable for copyright law to prioritise the protection of creative works as human society has a greater need for them. However, if creativity were the sole principle by which copyright protection was determined, resulting in a legal system for intellectual property, it is likely that more "Van Gogh tragedies" would occur at the legal level.

[15] Karl Marx and Friedrich Engels, "Capital: A Critique of Political Economy" in *Collected Works of Marx and Engels*, Vol.23 (Beijing: People's Publishing House, 2006), p.471.
[16] P.S. Adler, "Marx, Machines, and Skill" (1990) 31(4) *Technology and Culture* 780–812, available at: *https://doi.org/10.2307/3105907.*
[17] Jane C. Ginsburg and Luke A. Budiardjo, "Authors and Machines" (2019) 34 Berkeley Tech. L. J. 343, available at: *https://scholarship.law.columbia.edu/faculty_scholarship/2323.*
[18] James Vincent, "How three French students used borrowed code to put the first AI portrait in Christie's" (23 October 2018), *The Verge*, available at: *https://www.theverge.com/2018/10/23/18013190/ai-art-portrait-auction-christies-belamy-obvious-robbie-barrat-gans* [Accessed 8 March 2023].
[19] Karl Marx, "Capital: A Critique of Political Economy", (Chinese version),translated by Xiaohe He (Chongqing Publishing House, 2013), p.34.
[20] S. Sayers, "Creative Activity and Alienation in Hegel and Marx" (2003) 11 *Historical Materialism* 107–128.
[21] Zelin Zhao, "Machine and Modernity: Historical and Logical Insights from Marx and Beyond" (2020) 4 *Philosophical Research* 46–54.
[22] Marx's theory of commodity exchange argues that in a capitalist system, the exchange of commodities is based on their exchange value, which is determined by the socially necessary labour time required to produce them. This exchange value is distinct from the use value of the commodity, or its usefulness to the individual who owns it. Thus, the exchange of commodities is primarily an economic process driven by the exchange value of the commodities.It should be noted that Marx had a critical stance on this matter. In Volume One of "Capital," Marx pointed out the problem of commodity fetishism. In simple terms, the form of commodity exchange alienates the relationship between people into a relationship between things. This kind of relationship is not normal. To overcome this alienation, it is necessary to thoroughly change the capitalist ownership of the means of production. Due to limited space and the constraints of the topic, we will not discuss this in further detail here. For more information, please refer to Marx's "Capital: Volume One," published by People's Publishing House in 2004, 2nd edition, page 90.Other versions are also available for reference.
[23] Marx and Engels, "Capital: A Critique of Political Economy" (Chinese version) in *Collected Works of Marx and Engels*, Vol.23 (2006),.People's Publishing House, Beijing, 2006, pp.54–77.
[24] Chris Witts, "Tragedy of Vincent van Gogh—Morning Devotions" (22 November 2019), *Hope 103.2*, available at: *https://hope1032.com.au/stories/faith/2019/tragedy*

-*vincent-van-gogh/* [Accessed 8 March 2023].

This is due to the fact that creativity is a characteristic inherent to labour and its recognition is subject to variation among individuals. If AIGC is not protected by copyright due to an alleged lack of creativity, it may lead to consequences such as: (i) the potential for arbitrary copying without any legal recourse, and (ii) an inability for the results of intellectual labour to be exchanged, preventing producers from recovering their costs and obtaining profits, thereby hindering the economic use of AIGC, which is inconsistent with economic principles. In viewing the labour process as a continuous state, we may observe that at the level of social commodity exchange, if the quality of AIGC has already reached a level that satisfies consumers and is difficult to distinguish from artificially created originals, then there may be no need for buyers to differentiate between them. Additionally, given that the experience of news generated by artificial intelligence has already approached that of news written by humans, this is particularly true for knowledge products which are more difficult to distinguish from physical objects.[25]

The labour creation of artificial intelligence and the copyrightable process of AIGC

In terms of public welfare, the emancipatory role of AI in labour may be deemed more consequential than other facets, owing to the emphasis laid by Locke and MTL on labour as the foremost human identity and the priority afforded to its safeguarding. We assert that the paradigm of labour encompasses the notion of "enslavement" by inanimate intelligent agents.[26] If AI has no right to not suffer pain,[27] we would pay more attention to its value to humans. We endorse the notion that neither animals nor machines can be incentivised by copyright law to create works.[28] Empirical studies have suggested that the establishment of copyright law may not have fulfilled its intended purpose of incentivising creative output.[29] As previously argued, the lack of legal protection for certain types of works may hinder authors' labour productivity. The justification for intellectual property rights lies in preventing the misappropriation of the outcomes of intellectual labour. Copyright law recognises that AI-generated content is a result of human effort and given its ubiquitous and affordable usage, its proliferation is nearly unavoidable. AI technology has the potential to relieve individuals from the strenuous task of generating written content, especially considering the significant number of "knowledge workers" in modern society.[30]

With respect to the issue of ease of access to creative output through the use of AI, it is useful to consider a comparison between physical and mental labour. In the MTL, labour is distinguished in terms of ownership and the external form it takes. Workers commodify their labour and offer it for sale, with the buyer acquiring use value while the worker obtains the value of their labour. In the context of physical labour, workers may consider costs (defined by MTL paradigm as the cost of obtaining labour materials and objects) as well as the energy expended during production (which Marx refers to as necessary labour time) when determining the price of their commodity. The buyer, in turn, assesses the acceptability of the price based on the use value of the commodity, and this pricing process occurs directly between the buyer and seller in the same temporal and spatial domain. Ultimately, the pricing of these commodities is subject to the direct constraints of the buyer. The production and trade of intellectual labour products involve considerations of protecting labour results from infringement and establishing reasonable pricing based on their value, which necessitates the setting of a copyright protection period from an MTL perspective. Such regulations are crucial for establishing fair pricing based on actual value, as the period delineates a boundary between reasonable and unreasonable pricing agreed upon by both parties, and indicates the duration of time during which copyright holders can earn profits. Clearly, producing an article generated by AI is much easier than writing one using traditional methods, and the necessary labour time is shorter. Given that intellectual property rights create a monopoly for copyright holders within a certain period, it may be appropriate to shorten the copyright protection period for AIGC and appropriately lower transaction prices, thereby enhancing the bargaining power of diverse and complex buyers. For example, Mauritz Kop proposed shortening the patent period for AI inventions to 3–5 years.[31]

It should be clarified that MTL paradigm does not only focus on the three factors of labour, but rather uses these three key elements to define the boundaries of each intellectual labour field. MTL paradigm revolves around Marx's famous statement that "[32]The living, active labour is the soul of the entire production". Only by

[25] G. Song, "Application of Artificial Intelligence in News Communication" in J.C. Hung, J.W. Chang, Y. Pei, W.C. Wu (eds), *Innovative Computing: Lecture Notes in Electrical Engineering*, Vol.791 (Singapore: Springer, 2022).
[26] Belinda Bennett and Angela Daly, "Recognising rights for robots: Can we? Will we? Should we?" (2020) 12(1) *Law, Innovation and Technology* 60–80, available at: *https://DOI:10.1080/17579961.2020.1727063*.
[27] W. Sinnott-Armstrong and V. Conitzer, "How Much Moral Status Could Artificial Intelligence Ever Achieve?" in Steve Clarke, Hazem Zohny, and Julian Savulescu (eds), *Rethinking Moral Status* (Oxford: Oxford University Press, 2021).
[28] Qian Wang, "On the Qualification of Copyrightable Subject Matter Generated by Artificial Intelligence" (2017) 5 *Journal of Legal Science (Journal of Northwest University of Political Science and Law)* 148–155.
[29] R.S. Ku, J. Sun, and Y. Fan, "Does Copyright Law Promote Creativity? An Empirical Analysis of Copyright's Bounty" (2019) 62 *Vanderbilt Law Review* 1667.
[30] Richard Florida, *The Rise of the Creative Class* (New York City: Basic Books, 2014), p.174.
[31] M. Kop, "AI & Intellectual Property: Towards an Articulated Public Domain" (2020) 28(3) *Texas Intellectual Property Law Journal* 297–342.
[32] Marx and Engels, "Capital: A Critique of Political Economy" (Chinese version), in *Collected Works of Marx and Engels*, Vol.49 , People's Publishing House, Beijing, 2006, pp.47–66.

framing these principles can the value of labour be better protected, including the mental, creative, temporal, physical, and intellectual efforts, thoughts, and personalities that scholars or judicial institutions require. Often, these individual principles only protect one aspect of intellectual labour, but MTL paradigm comprehensively addresses the unique challenges of intellectual labour in terms of time and space, including artistic creation as the highest form of labour.[33]

The Berne Convention for the Protection of Literary and Artistic Works, which was established to address copyright protection issues, is widely regarded as affirming human beings as the creators of work.[34] The Convention's text and the historical context in which it was signed both indicate that the general term of copyright protection is based on the author's lifetime, as stated in art.7(1) of the Convention. This provision also covers the protection of the author's moral rights, but it is important to note that it only applies to natural persons. The Convention does not provide a specific definition for determining copyright eligibility, and therefore, each member country establishes its own criteria for eligibility.

Methodology

Method of Definition

Ballardini, Kan, and Roos conducted a literature review and identified three schools of thought on the topic, namely the revolutionary, romantic, and modernist schools. The revolutionary school believes that AI can be granted legal personhood, making it a rightsholder. In contrast, the romantic school disagrees with the revolutionary school and argues that intellectual property protection is intended to safeguard an author's thoughts, personality, and creativity,[35] and therefore opposes the protection of intellectual property for non-human entities. The modernist school insists that only natural persons can become authors; based on this foundation, it discusses granting intellectual property rights to potential authors in the process of AI creation and usage.[36] Based on prior discourse, it is suggested that the primary requirement for authorship should transition from a focus on personality to a focus on "human labour".[37] Essentially, the discussion is about "whose labour is the work essentially based on?" The views of the revolutionary school are inconsistent with the conclusion of protecting the labour of human agents based on MTL paradigm derivation in this article, and cannot follow the principle of irreplaceability between the subject and object of human and machine from the perspective of Kantian philosophy.[38] From an economic standpoint, AI lacks a tangible entity or subject that can have interests, and granting it legal status could lead to ambiguity regarding who benefits from AI outcomes. If intellectual property law were to recognise AI as an author, would it then be necessary to endow AI with constitutional rights, civil legal status, criminal responsibility, and other legal positions? This would undoubtedly cause significant legal upheaval and shake human ethical and moral norms.[39] It is clear that at present, AI does not meet the fundamental criteria for personhood that apply to both human beings and legal entities,[40] and it lacks the ability to autonomously decide on its own objectives and values.[41] Therefore, this article first eliminates the viewpoint of the revolutionary school.

Based on MTL paradigm, we believe that the copyright ownership of AIGC should highly overlap with that of the labour owner. First, we distinguish between two different types of labour involved in AI design and use. The purpose of computer scientists is to design an AI, which is accomplished by utilising materials such as computer theory, programming languages, algorithms, data structures, and hardware facilities. This already constitutes a complete labour process. The user can obtain the right to use through legal exchange with the AI designer, which is the process of acquiring production materials for the user. Specifically, for the user, an AI without a database is considered a labour material, while the database used by ChatGPT is the object of labour, and ChatGPT's model is more inclined to be labour material. Overall, AI is the user's production material and the labour result of the designer. Users may use AI to obtain a desired text by giving instructions to AI. AI, based on its database, learns and adjusts itself through the model, generates text,and outputs results.[42] This is the labour process of the user. In terms of the external form of the method, the "authorship transfer" theory is very similar to our argument, although it does not recognise the existence of labour. It believes that these works are essentially created by non-human AI and should be protected by copyright.

[33] S. Sayers, "Why Work? Marx and Human Nature" (2005) 69(4) *Science & Society* 606–616.
[34] Berne Convention for the Protection of Literary and Artistic Works (Paris Act of July 24, 1971, as amended on September 28, 1979).
[35] J. Grimmelmann, "Copyright for Literate Robots" (2016) 101(2) *Iowa Law Review* 657–682.
[36] R.M. Ballardini, H.Kan, and T. Roos, "AI-generated content: authorship and inventorship in the age of artificial intelligence" in Taina Pihlajarinne, Juha Vesala and Olli Honkkila (eds), *Online Distribution of Content in the EU* (Cheltenham: Edward Elgar Publishing, 2019), p.117.
[37] Barry Scannell, "When Irish AIs are smiling: could Ireland's legislative approach be a model for resolving AI authorship for EU member states?" (2022) 17(9) *Journal of Intellectual Property Law and Practice* 727–740.
[38] Yang Li & Xiaoyu Li, "Discussion on the Copyright Issues of Artificial Intelligence Generated Works from the Perspective of Kantian Philosophy" (2018) 9 *Journal of Law* 43–54.
[39] Shakuntla Sangam, "Legal Personality for Artificial Intelligence with Special Reference to Robot: A Critical Appraisal" (2020) 6(1) Indian J Law Hum Behav 15–22.
[40] S. Chesterman, "Artificial Intelligence and the Limits of Legal Personality" (2020) 69(4) *International & Comparative Law Quarterly* 819–844, available at: *https://doi:10.1017/S0020589320000366*.
[41] I. Gabriel, "Artificial Intelligence, Values, and Alignment" (2020) 30 *Minds & Machines* 411–437.
[42] Sai Vemprala et al, "ChatGPT for Robotics: Design Principles and Model Abilities" (2023), available at: *https://www.microsoft.com/en-us/research/uploads/prod/2023/02/ChatGPT___Robotics.pdf*.

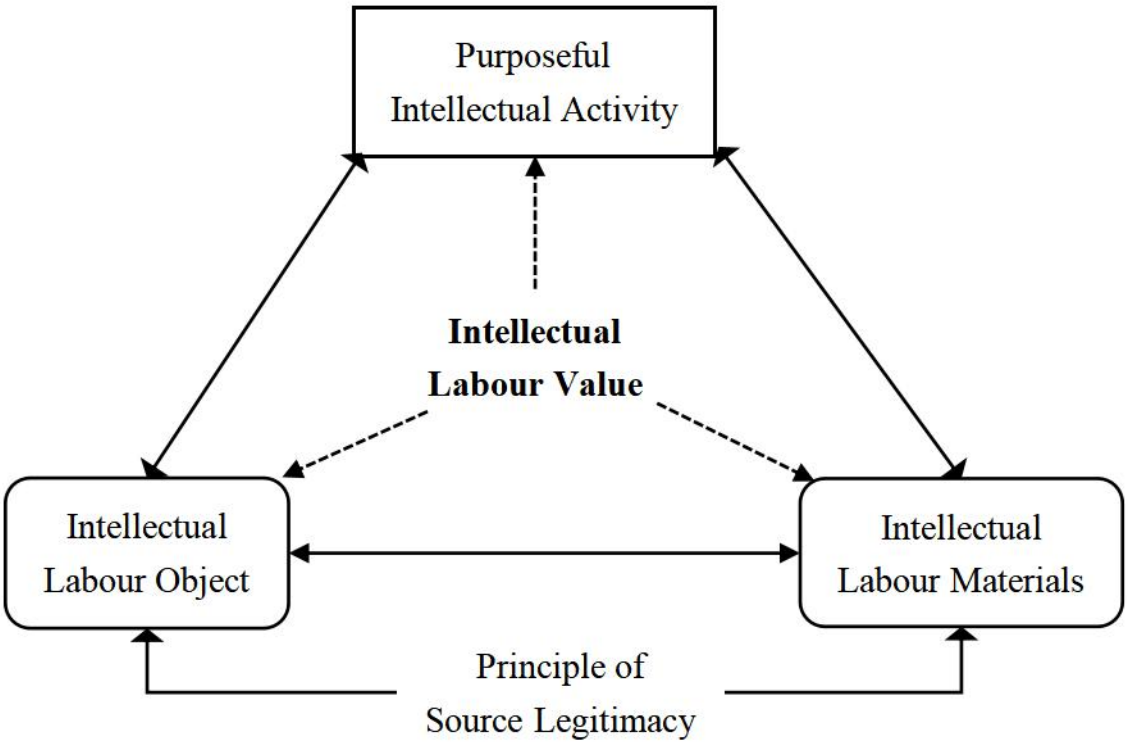


Figure 1: Criteria for copyrightability of the AIGC and determination of authorship[43]

effective dissemination and incentivisation of production of AIGC, the fully controlled person should be "constructed" (sometimes referred to as "fiction") as the author.[44] The purposes, labour materials, and labour objects of the designer and user are completely different. If we distinguish between two different types of labour, then the issue of ownership of labour results becomes clear, and there is no need to fabricate an author.

The preceding analysis is based on the assumption that designers and users can be easily distinguished. However, there are numerous potential claimants to intellectual property rights according to the AIGC, and current academic discussions revolve mainly around perspectives from artificial intelligence suppliers (designers or providers), programmers, end-users, artificial intelligence entities (referring to the view that artificial intelligence should have legal personhood), and joint rights bodies.[45] This article does not assert a definitive attribution of the AIGC to a particular entity, but rather underscores the significance of determining "whose labour product" the AIGC embodies, which may involve one or more agents. Once this matter is resolved, the ownership of the intellectual property for the AIGC and the authorship hierarchy can be ascertained. We have developed a rudimentary methodology utilising MTL paradigm and devised the "three elements and one principle" criteria for recognition, which is presented in Figure 1. In essence, a copyrightable AIGC necessitates the deliberate manipulation of the AI and the rightful acquisition of materials and objects required for intellectual labour by the author. Similarly, an adept author must satisfy these conditions.

The principle of source legitimacy pertains to both the legitimacy of the material origin and the functional origin of labour materials and objects. The legitimacy of the material origin requires that the author's labour materials and objects are procured through legitimate means, which may involve transactions or authorised, free use granted by the owner of the AI. The principle of legitimacy of sources addresses the legitimate acquisition and use of labour materials and objects, with a focus on both their material and functional sources. The material source legitimacy requires that the labour materials and objects utilised by the author are lawfully obtained, either through transactions or authorised, free use by the AI owner. Conversely, the functional source legitimacy necessitates a high degree of alignment between the purpose of the labour materials and objects and the author's creative intent. In this context, the AI model utilised by the author must demonstrate textual composition capabilities, as only such AI models can be considered legitimate functional source. For example, the AI tool ChatGPT cannot be viewed as a legitimate functional source, as it has not been marketed by OpenAI as a creative AI tool for writing purposes.[46] It cannot be assumed that OpenAI

[43] Figure 1 was created by the author.
[44] B. Lu, "A theory of 'authorship transfer' and its application to the context of Artificial Intelligence creations" (2021) 11(1) *Queen Mary Journal of Intellectual Property* 2.
[45] Robert Yu, "The Machine Author: What Level of Copyright Protection Is Appropriate for Fully Independent Computer-Generated Works?" (2017) 165 U. Pa. L. Rev. 1245.
[46] OpenAI, "Introducing ChatGPT", available at: *https://openai.com/blog/chatgpt* [Accessed 8 March 2023].

has sold the creative capabilities of ChatGPT, and furthermore, ChatGPT may not have yet met the criteria for being considered original content eligible for copyright in different countries. Consequently, works generated using ChatGPT may not qualify for copyright protection, given its inadequacy as a creative AI tool due to its high level of formal redundancy and relatively lower creative standards. Instead, creative AI platforms such as Deep Dream Generator, Wordsmith, and Flow Machines require advanced technologies to meet the necessary standards of creativity. The esteemed linguist Noam Chomsky has gone as far as to characterise ChatGPT as a form of "high-tech plagiarism".[47] Although current weak AI systems often fail to meet the demanding standards for high creativity and low redundancy that creative AI tools require, it would be premature to conclude that the copyrightability of AI-generated content should be dismissed solely based on the current limited creative output and high level of repetition in AI systems.[48] The German theory of "Kleingeld" has long recognized this issue, thus categorizing works based on creativity standards for different types of creations.[4950].Furthermore, it is equally essential to consider the legitimacy of the material source, as OpenAI frequently employs databases such as Wikipedia, CommonCrawl, and WebText.[51] If OpenAI were to use these databases without proper authorisation, it would be deemed a violation of the legitimacy of the material source. Francesco Marconi, a journalist for *The Wall Street Journal*, has accused OpenAI of using articles from media outlets such as *Reuters*, *The New York Times*, *The Guardian*, and the BBC to train ChatGPT without permission and without payment. With regard to the feasibility of determining the legitimacy of the material source, scholars have proposed a solution through the use of proprietary digital copyright management technology for legal reasoning on external data. This solution ensures compliance with regulations during the creative process, while also guaranteeing the legality of the collection and use of such materials.[52]

MTL-based revision

In the United States (US) telephone directory case, the court established the principle that works consisting of unoriginal, purely informational content are not protected by copyright law. To obtain telephone services within the plaintiff's service area, users must submit an application that includes their name and town. [53]The plaintiff then issues a telephone number to the user and compiles a directory of the collected user's names, towns, and telephone numbers in alphabetical order. [54]The court decided that the directory did not qualify as a creative work since it contained identical information provided by users without any modifications. As a result, it lacked originality.[55] It is believed that the copyright of a compilation does not extend to the materials compiled, thereby establishing that the "industrious collection" principle does not apply to the determination of the "originality" element of a work in copyright law. [56] Further more, the plaintiff did not decide to compile this directory on their own, but rather was required to do so by the regulations of the state company committee, which weakens its originality by following external instructions rather than internal, self-driven motives. [57]Lastly, the plaintiff's arrangement of compiling the directory in alphabetical order, although it is a coordinated arrangement that requires a certain amount of labour, is an arrangement that has been long established and considered a matter of course, and therefore not unique to the plaintiff and lacking the minimum "creative spark" required by copyright law to determine originality, thereby establishing that the "sweat of the brow" principle does not apply to the determination of the "originality" element of a work in copyright law.[57]

In the case of *Burrow-Giles Lithographic Co v Sarony*, the court's view can be summarised as follows: When humans use machines to create content, if the final product has the originality that human thought brings, it can be regarded as a work, and humans will be considered the authors.[58] At the time of the case, some argued that photos were simply mechanical copies of the physical characteristics or outlines of objects (whether living or not), and did not involve any originality of thought or intellectual effort related to the visible reproduction in the image's form.[59] As a result, photos were not considered works under US copyright law.[60]

[47] Jessica Stewart, "Noam Chomsky Says ChatGPT Is a Form of 'High-Tech Plagiarism'" (February 17, 2023), *My Modern Met*, available at: *https://mymodernmet.com/noam-chomsky-chat-gpt/* [Accessed 8 March 2023].
[48] He Tianxiang, "The sentimental fools and the fictitious authors: rethinking the copyright issues of AI-generated contents in China" (2019) 27(2) *Asia Pacific Law Review* 218–238, available at: *https://doi.org/10.1080/10192557.2019.1703520.*
[49] According to the "Kleingeld" theory, different types of works require varying levels of creativity. Literary and scientific works, for example, demand a higher level of creativity, whereas computer programs and similar works only require a moderate level of creativity.
[50] Marcel Bisges: Die Kleine Münze im Urheberrecht – Analyse des ökonomischen Aspekts des Werkbegriffs. Nomos-Verlag, Baden-Baden 2014, ISBN 978-3-8487-1775-0
[51] T.B. Brown et al, "Language Models are Few-Shot Learners" (2020) ArXiv, available at: *https://arxiv.org/abs/2005.14165.*
[52] Jesus Manuel Niebla Zatarain, "The role of automated technology in the creation of copyright works: the challenges of artificial intelligence" (2017) 31(1) *International Review of Law, Computers & Technology* 91–104, available at: *https://doi.org/10.1080/13600869.2017.1275273*.
[53] *Feist Publications Inc v Rural Tel Serv Co* 499 U.S. 340 (1991), available at:*https://www.law.cornell.edu/supremecourt/text/499/340.*
[54] Ibid.
[55] Ibid.
[56] Ibid.
[57] Ibid.
[58] *Burrow-Giles Lithographic Company v Sarony* 111 U.S. 53 (1884), available at:*https://supreme.justia.com/cases/federal/us/111/53/.*.
[59] Ibid.
[60] Ibid.

Amidst the case's backdrop, some argued that photographs were mere mechanical reproductions of physical features, lacking intellectual creativity and novelty related to visible replication.[61] However, the court disagreed and acknowledged the photographer's originality in posing, selecting, and arranging elements within the photograph, thus considering it an independent work of art under copyright law. The photographer is the author of this photo..[62]

There is currently no clear legal guidance in the United States regarding the degree of human thought activity required for AIGC to meet the requirement of "originality." Determining whether a work is "original" will depend on the specific circumstances of each case. The USPTO has proposed several eligibility criteria that can be used as reference points for identifying the "author" in cases involving human and AI collaboration. These criteria include factors such as whether the person designed the AI algorithm or process used to create the work, whether the person contributed significantly to the design of the AI algorithm or process, and whether the person intervened or selected the data used to train the algorithm or for other purposes.[63] Other factors include whether the person applied or guided the AI algorithm through personal choice and arrangement, such that the resulting output could be used for work, and whether the person engaged in multiple activities as described above.[64]

The principles of originality requirements, "sweat of the brow" principle, and necessary arrangements (degree of human involvement) are unclear and each of these theories provides a separate justification on its own.They neglect the labour of individuals as subjects and the division of their fields, resulting in the frequent confusion of several independent labour processes. MTL paradigm offers a process-based solution. The purpose of using AI for work has two meanings related to its intended output. The first aspect is what the user wants AI to produce, which can be framed in two ways:(i) what specific thing do I want AI to generate for me? and (ii) what kind of output do I want AI to generate for me? According to Scannell's example, even if someone only uses minimal effort to click a button and generate a work, as long as they legally obtained the AI and database, the intellectual property of the work should belong to them because the person's use was purposeful activity or labour. The law should not consider the amount of effort spent in making these arrangements, as it is still considered their labour. Ginsburg and Budiardjo attempted to establish the identities of two types of authors, the upstream creator who made some contribution during the creative process but did not ultimately result in the creation of the work, and the downstream creator, who did the actual work. Because AI may strongly express the will of the designer, it may weaken the originality and creativity of the user. Therefore, Ginsburg and Budiardjo believed that the creator of the AI would be the potential author. This view is essentially the same as the McCutcheon principle, which also did not consider labour, but McCutcheon believed that the unpredictability of AI's autonomous production would lead to the "the vanishing author in computer-generated works", requiring a new copyright rule to correct it.[6566]

In fact, the author as the subject of labour will never disappear. As long as labour exists, labourers will always be present. To address above commonly held misconceptions, there are several corrective points:

MTL paradigm emphasises the purposefulness of labour rather than creativity, and the designer's will imposed on AI's performance can be regarded as a characteristic of labour tools in the user's labour. No matter how strongly this characteristic is expressed, it will not affect the user's purpose of using AI. Therefore, strong will expression cannot be a reason for AI creators to be considered as authors of their works. Moreover, under the trend of increasingly strong "autonomy" of AI, the influence of the designer's will may become smaller and smaller, because the intelligent development of AI will only increasingly conform to the user's purposefulness, that is, the so-called intelligentisation, which enables the user's will to be better executed.

As for the potential for upstream and downstream authorship arrangements to lead to the exploitation of user labour outcomes by the designer. Typically, the designer obtains ownership or usage rights to AI in exchange for something, and the benefits derived from this exchange can be considered a form of labour remuneration for the designer. However, once the AI has been created, and its downstream use results in labour benefits that are satisfied, there is no legitimate reason for the designer to claim further rights in the user's labour field, particularly given that the user has already incurred usage costs. Unless the designer is also a user, any attempt to do so would be tantamount to an unjustified appropriation of the user's labour outcomes. In this context, creators stand to gain from users, while users perceive AI as a form of productive capital that yields labour outcomes, which they exchange for benefits. This symbiotic relationship not only satisfies Locke's

[61] Ibid.
[62] Ibid.
[63] USPTO, "Public Views on Artificial Intelligence and Intellectual Property Policy" (October 2020), p.22, available at: *https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf.*
[64] Ibid.
[65] J. McCutcheon, "The Vanishing Author in Computer-Generated Works: A Critical Analysis of Recent Australian Case Law" (2013) 36 *Melbourne University Law Review* 915–969 and J. McCutcheon, "Curing the Authorless Void: Protecting Computer-Generated Works Following IceTV and Phone Directories" (2013) 37 *Melbourne University Law Review* 46–102, UWA Faculty of Law Research Paper No.28.
[66] McCutcheon, Jani, Curing the Authorless Void: Protecting Computer-Generated Works Following IceTV and Phone Directories (2013). Melbourne University Law Review, Vol. 37, 46-102, UWA Faculty of Law Research Paper No. 28.

"sufficient and equal" criteria,[67] but also does not conflict with the requirement of Pareto efficiency espoused by the intellectual property rights framework.[68][69] Furthermore,this improvement occurs on a profound level of internal economic relations. The most important thing is that no one can encroach upon the labor of others. For instance, in the case of Beijing Feilin Law Firm versus Baidu Baijiahao for copyright infringement, the court recognized the plaintiff's copyright over the reports created through a legal information database, but did not consider the plaintiff's claim that the graphics constituted a work of art. This is because "[70]the differences in the shapes of the images are based on data variations rather than creative production". There is an interpretation of the MTL paradigm, according to which these images constitute a form of labour object as case data. In the new production process, the user's purposive selection space is exceedingly limited, lacking the utilization, synthesis, innovation, and presentation of elements that are typical of conventional artistic creation. In relation to the user, the copyright of the data-generated image fundamentally belongs to the AI developer who sold it as a production tool, as this image has not undergone essential changes through novel labour processes. This underscores the need for those employing AI for creative purposes not only to consider whether the AI possesses creative capacities, to avoid legal uncertainties, but also to attend to the extent to which the AI offers choice for the creator during the creative process. Such choice not only extends the creator's purpose but also satisfies the reconciliatory requirements of traditional copyright principles, such as originality, creativity, and necessary arrangements.

The requirement for copyrightability, originality, may disqualify AIGC due to the prevalent notion of an "algorithmic black box" and the lack of human interpretability of AI. [71] Humans cannot control what kind of content is generated by AI within this black box, leading many to argue that AI creations possess a certain level of autonomy. This perspective actually denies the second-order stochasticity of human labour production. When we are drawing a picture, we allow the influence of various factors such as the pen, paint, canvas, trembling of the hand, flow of paint, mental state, style, etc. on the final artwork. In fact, human society has always allowed this kind of randomness in creation, but for permanent labour, human intentionality comes first and human labour does not allow randomness to determine its course. However, the randomness of AIGC is also a kind of stochasticity guided by human intentionality. The materials (or databases) generated by artificial intelligence come directly from real materials, and the generation of these materials itself has randomness, although they have a fixed form when they are input into the database. Therefore, there is no difference between the randomness and unpredictability of the processing, combination, and application of input data used with these materials and that embraced in original creation. Since we have already recognized the first-order randomness of labour, what reason is there not to recognize the second-order?

There is a considerable body of scholarly opinion which advocates for the incorporation of AIGC into the public domain for its management. For instance, Mauritz Kop contends that patenting acts as a hindrance to innovation due to the obstacles presented by the intricate forest of rights, that ultimately results in market barriers and monopolies. As a result, Kop argues that public ownership of AIGC is a superior approach to fostering innovation. However, an opposing view is that the public ownership of AIGC, without regard to its classification or the time of its creation, amounts to the collective appropriation of individual labour by the public. This would constitute another form of privatisation and theft and could create impediments to new entrants in the market. Furthermore, it is believed that the incorporation of AIGC into the public domain at its point of inception in a general sense, particularly in light of the prevailing market economies of most nations, is not a tenable system design. Nevertheless, it is agreed that goods that have already yielded reasonable labour benefits could enter the public domain to prevent extended monopolies, and this is indeed the practice as stipulated by the current law that ceases to protect intellectual property rights 50 years after the death of the author.

In the current legal framework, AIGC's copyrightability process could encounter challenges, as AI users may not satisfy the minimal authorship requirement of "intellectual labour" contribution as stipulated by the Berne Convention. We contend that the notion of "intellectual creation" can be broadened to encompass the increased flexibility of AIGC users through intelligence. Specifically, the second layer of intention mentioned earlier, namely "what kind of output do I want AI to generate for me" could be subsumed under "intellectual creation". This is particularly consistent with the notion of intellectual creation, since, with ample space for choice offered by AI to users, the user's ideas assume a preeminent role in the creative process, albeit with AI replacing human execution and AI is generally better equipped than human labour to handle processing tasks in the trajectory of intelligence.

[67] J. Locke, *Two Treatises of Government*, translated by Ye Qifang and Qu Junong (Beijing: The Commercial Press, 1996), pp.5–12.

[68] Adam D. Moore, "Toward a Lockean Theory of Intellectual Property" in A. Moore (ed), *Intellectual Property: Moral, Legal, And International Dilemmas* (New York: Rowman & Littlefield, 1997), p.81.

[69] It should be noted that our method is based on the labour theory, where everyone receives labour income that justifies intellectual property rights based on the labour theory. This can be a legitimate reason in jurisprudence rather than an optimized choice, but it also happens to meet the requirement of the pareto efficiency of "allowing someone or a few people to benefit without harming the interests of others", at least no one's interests are harmed by this distribution plan, and everyone gets what they deserve.

[70] *Beijing Feilin Law Firm v Beijing Baidu Netcom Technology Co Ltd* Judgment of the Civil Case No. (2018) Jing 0491 Min Chu 239, by the Beijing Internet Court, on the Copyright Infringement of Written Works Lawsuit.

[71] L.H. Gilpin et al, "Explaining Explanations: An Overview of Interpretability of Machine Learning", 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (Turin, Italy: 2018), pp.80–89, available at: *DOI:10.1109/DSAA.2018.00018*.

Case review

In the case of copyright infringement brought by Beijing Feilin Law Firm against Baidu Baijia self-media platform, the plaintiff utilised the Wolters Kluwer China Law & Reference database to conduct a keyword search and generate an analysis report consisting of both charts and textual content via the database's "visualization" function. This report was subsequently published on the plaintiff's WeChat public account. In due course, the defendant, Golden Touch, published the allegedly infringing article on Baidu Beijing's self-media platform without permission, omitting the plaintiff's signature and certain portions of the report. As a consequence, Feilin Law Firm filed suit against Baidu Baijia self-media platform for violating its right to attribution, integrity, and information network dissemination. In response, the defendant argued that the report in question lacked originality and was therefore not protected by copyright law, and further asserted that the plaintiff, as a legal entity, could not serve as the author and should accordingly be exempt from any liability for infringement. Figure 2 is for a more detailed account of the case.

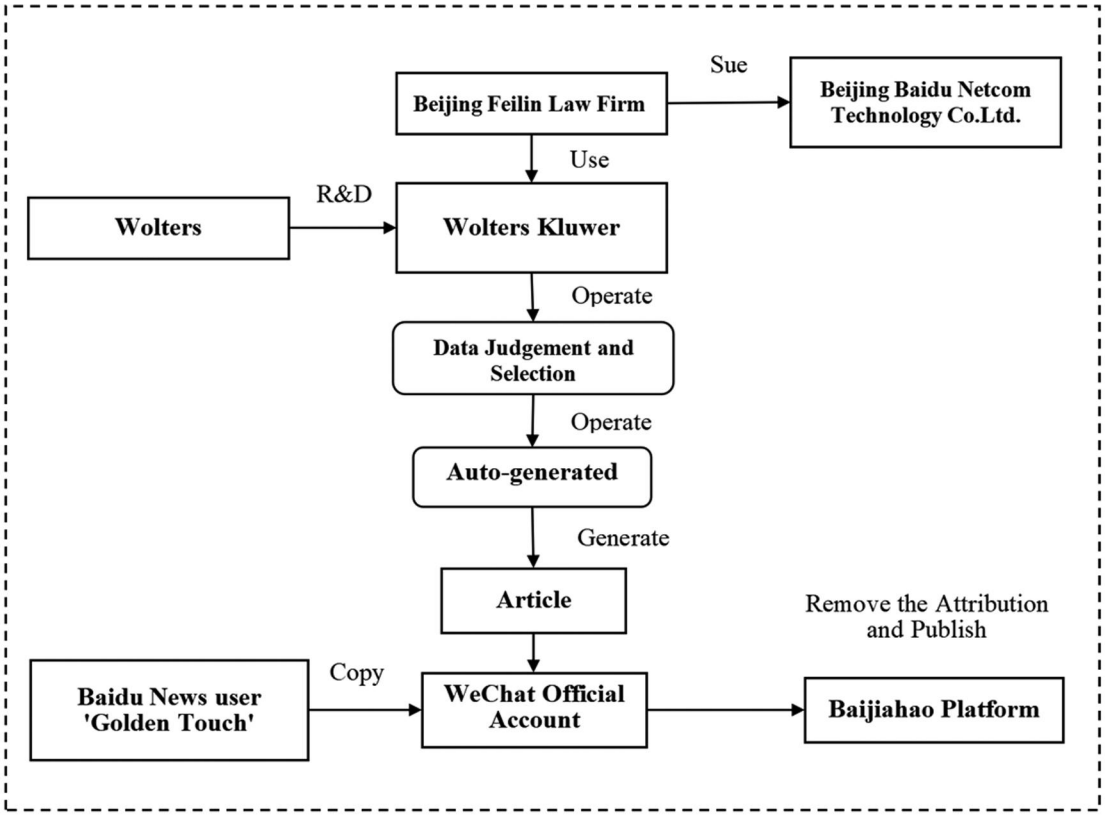


Figure 2: Diagram of the case filed by Beijing Feilin Law Firm against Bejing Baidu Netcom Technology Co.Ltd.[72]

In general, the approach taken by the first-instance court closely resembles the "authorship transfer" theory, which consistently seeks to identify an author who is most in line with the current legal framework. The court initially determined that the report in question was essentially generated by the Wolters Kluwer China Law & Reference, as the "analysis report is formed through the combination of input keywords, algorithms, rules, and templates by the Wolters Kluwer China Law & Reference, which can be considered to have 'created' the analysis report to some extent." Given that the Copyright Law requires the author to be a natural person, the court abandoned the attempt to grant copyright to Wolters Kluwer China Law & Reference and instead sought to identify a potential natural person author. The Beijing Internet Court considered both the research and development (R&D) and use stages of AI to be part of the creative process of the work, which appears to differentiate between the distinct labour fields of usage and development. However, the court's decision to exclude the authorship of both software developers and users based on the originality principle was disappointing, as it determined that "the analysis report did not convey the software developer's (owner's) original expression of thoughts and feelings because the software developers did not search for keywords according to their needs". The MTL paradigm suggests that the report in question cannot be attributed to the labour of the developer, as it did not serve the developer's purpose. The court also held that the analysis report generated automatically by the "visualization" function did not reflect the software user's original expression of thoughts and feelings, and therefore, it could not be recognised as the user's work. However, the court acknowledged that the analysis report generated using the "visualization" function involved a certain degree of originality, as it required the selection, judgment, and analysis of relevant data. In our view, the user's demand for the uniqueness of the data

[72]Figure 2 was created by the author.

during the labour process satisfies the user's exclusive labour purpose, making the user the author of the report. Nevertheless, the court did not adopt this approach. As software users, they have invested by paying to use the software, meeting the principle of legitimacy of the source. They have set keywords based on their needs and generated an analysis report, which motivated them to use and disseminate the report, fulfilling the MTL's purposefulness requirement. After excluding the identities of the developers, users, and the WeChat group, the court also refused to push the report involved into the public domain because "the production of the analysis report not only embodies the input of the software developers (owners) but also the input of the software users, and it has dissemination value. If the input contributors are not granted certain rights protection, it will be detrimental to the dissemination of the input results (i.e., analysis report), and the utility of it cannot be fully utilized. For software developers (owners), their interests can be obtained through software usage fees and other means, and their development inputs have already received corresponding returns. Moreover, the analysis report is generated by software users according to different usage needs and retrieval settings, and software developers (owners) lack the motivation to disseminate it." Here, the court expressed a typical tendency of pushing copyright law towards investment law.[73] The court believed that "the production of analytical reports not only embodies the investment of software developers (owners), but also the investment of software users, which has dissemination value. If the investors are not given certain rights protection, it will be unfavorable for the dissemination of investment results (i.e. analytical reports) and cannot exert their utility." Based on incentive theory, the court determined that the users have relevant rights but no right to attribution as their investment has not been rewarded. Finally, the court found out whether there was any direct creation based on an automatic generation system in the report involved, and officially attributed the copyright to Feilin Law Firm, which was the user. The second-instance Beijing Intellectual Property Court basically supported the first-instance view.[74] We have summarised the trial logic of the first-instance Beijing Internet Court, which can be found in fig.3 below.

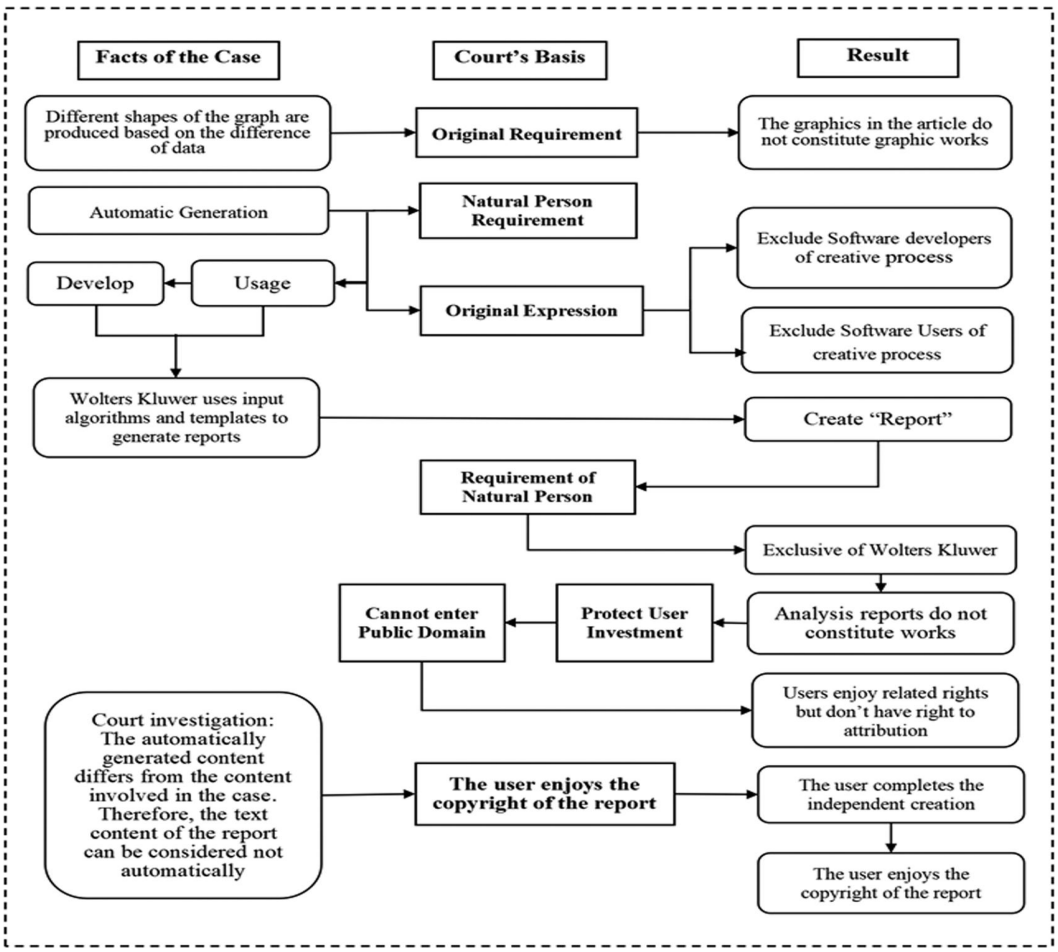


Figure 3: The trial logic of the first-instance court in the Feilin Law Firm vs. Bejing Baidu Netcom Technology Co.Ltd. case heard by the Beijing Internet Court[75]

[73] Ruth L. Okediji, "When is Intellectual Property an Investment?" in Christophe Geiger (ed), *Research Handbook on Intellectual Property and Investment Law* (Cheltenham: Edward Elgar Publishing, 2020), p.94.

[74] *Beijing Feilin Law Firm v Beijing Baidu Netcom Technology Co Ltd Judgment of the Civil Case No.* (2019) Jing 73 Min Zhong 2030, by the Beijing Intellectual Property Court, on the Copyright Ownership and Infringement Dispute Appeal.

[75] Figure 3 was created by the author.

Although this case made progress in the AIGC copyright result,[76] the requirement of originality has been repeatedly used by judges, apart from the final circumstances, and the court was "lucky" to find that the report in question had the plaintiff's direct creation, which led to the recognition of the plaintiff's copyright in the report. However, what if the automatically generated report has not been directly processed by human labour? And if the originality principle excludes all natural persons in this case from copyright, it will lead to "[77] the vanishing of the author", as described by McCutcheon. In fact, ignoring the labour theory of law often leads to such judicial views, which, if not revised more comprehensively, will lead to a series of judicial dilemmas regarding AIGC copyright issues. The *Tencent v Yingxun Technology* copyright infringement dispute case also has such a coincidence. Tencent, as both the developer and user of the intelligent writing software Dreamwriter, was recognised as the copyright owner of the article in the court's support of the originality view.[78] This seems reasonable and legal, but the court did not consider whether the data source used by Dreamwriter for creation was legitimate or whether the relevant rights holders of the data were consulted. Moreover, the developer cannot always be the user at the same time. The *Tencent v Yingxun Technology* copyright infringement dispute case discussed the copyright issue of AIGC in the situation where the user and developer are mixed together, which does not have strong persuasiveness under the trend of widespread use of AI creation in the future. Therefore, under the current copyright rules, the copyright issue of AIGC has not been fundamentally resolved. At the same time, we do not exclude the provisions of the existing copyright rules. In fact, MTLs can be included in different aspects under the existing copyright rules, which may be a good solution to avoid systematic legal turbulence in the copyright law system. For example, originality can be understood as the unique choice of the author when using AI to create, because the operation of the machine itself is an extension of human purpose.

Concluding remarks

This article is situated between two labour theories, those of Marx and Locke, and creates a methodology for defining the copyright of AI-generated works. The discussion in this article is not primarily concerned with the question of whether AI-generated creations need intellectual property protection, or how to protect them, from a doctrinal perspective. Rather, it examines how changes in production relations resulting from further technological advancement will affect legal and public policy, which are superstructures. The correct determination of the ownership of intellectual property and its attribution to the general rightsholder has liberating economic significance for intellectual labourers, which will benefit economic development and social progress. The number of people who use AI for creative or work purposes is increasing, as evidenced by the rapid evolution from ChatGPT 3.5 to 4.0. We are currently in a critical period of the scientific control theory revolution, at the "primary modernization stage" of new technology, which involves the diffusion and strengthening of production principles.[79] Our understanding of production principles cannot remain in the traditional industrial era. If new things like AIGC cannot receive effective copyright protection and a proper allocation of rights that conforms to actual production relations, and if a standardised legal system cannot be produced, then we will face significant challenges.[80] This will hinder the benefits that new technologies can bring to society as a whole. Therefore, this article emphasises the significance of AI as a means of production further socialising, such as its role in improving production efficiency, output quality, employment conditions, and so on, based on the labour-based copyright treatment of AIGC.

[76] J.Y. Lee, "Artificial intelligence cases in China: Feilin v. Baidu and Tencent Shenzhen v. Shanghai Yingxin" (2021) 7(1) *China and WTO Review* 211, 222.

[77] Ibid.at 65.

[78] *Tencent Computer Systems (Shenzhen) Co Ltd and Shanghai Yingxun Technology Co Ltd* Judgment of the First Instance of Civil Case No. (2019) Yue 0305 Min Chu 14010, by the People's Court of Nanshan District, Shenzhen City, Guangdong Province, on the Copyright Ownership, Infringement Dispute, Commercial Bribery, and Unfair Competition Dispute.

[79] Leonid E. Grinin and Anton L. Grinin, "The Sixth Kondratieff Wave and the Cybernetic Revolution" in *Kondratieff Waves: Juglar — Kuznets — Kondratieff Yearbook* (Volgograd: Uchitel Publishing House, 2014).

[80] L. Floridi, "Soft Ethics and the Governance of the Digital" (2018) 31 Philos. Technol. 1–8.

3D Laser Scanning Acquisition and Modeling of Tunnel Engineering Point Cloud Data

3D Laser Scanning Acquisition and Modeling of Tunnel Engineering Point Cloud Data

Fangzhe Shi[1,*], Jingxin Yang[2], Qiuyi LI[1], Junjie He[1], Boning Chen[1]

[1]Guangzhou Maritime University, Guangzhou 510725, China
[2] South China Agricultural University,Guangzhou 510725, China
Email: 476057168@qq.com

Abstract. For tunnel deformation analysis using traditional measurement methods to obtain tunnel section data, there are problems such as small data coverage and low efficiency. 3D laser scanning technology has the advantages of automatic, high precision and high efficiency in collecting the point cloud data of the target object, and can completely and accurately express the target entity. Based on the tunnel point cloud acquired by 3D laser scanner, the tunnel engineering modeling research is carried out in this paper. Firstly, the engineering survey and 3D reconstruction technology of shield tunnel were carried out based on 3D laser scanning technology. The total station layout control was used to scan the subway platform and tunnel interval, and the high-precision laser point cloud data were obtained. Secondly, a random sampling consistency algorithm is proposed to extract engineering measurement results such as tunnel axis and cross section. Finally, the 3D modeling of the tunnel is established by using the stretching setting out modeling method. Taking Yuzhu Tunnel planning and acceptance project as an example, the experimental results show that the method can effectively visualize the overall deformation of the tunnel, and provide accurate and scientific spatial data for tunnel engineering measurement, operation and maintenance.

Keyword. 3D laser scanning technology: Point cloud data; 3D reconstruction; Tunnel engineering

1. Introduction

Traditional measurement method is the use of a single point of access to target the 3D position information, the traditional tunnel engineering measurement method can be collected high-precision tunnel cross section data, but to collect data of low work efficiency and point to point control point such outstanding problems as easily damaged in the process of tunnel in construction and operation are responsible for a large number of economic losses. 3D laser scanning technology has the advantages of high data accuracy, efficient data acquisition, real-time, non-contact, full automation, panoramic scanning and so on. 3D laser scanning technology has gradually replaced the traditional measurement methods in the field of tunnel engineering[1].

Since the emergence of 3D laser scanning technology, many countries have done a lot of research on this technology. One of the world's first 3D laser scanners was based on lidar technology. In terms of 3D laser point cloud data processing, Gaopeng Hou[2] discussed and studied the downsampling algorithm and point cloud denoising method in view of the data reduction problem of tunnel point cloud model. Zhihai Liu[3] introduced a measurement data processing method for tunnel laser 3D scanning monitoring, which solved the problem of reducing the monitoring accuracy of the conventional point cloud thinning method. In terms of tunnel point cloud section extraction, Min Liu[4]

selected a station and interval of Hangzhou Metro Line 5 for measurement and analysis, and showed that the mobile 3D laser scanning technology can better meet the requirements of the full section scanning of metro tunnel. Shuhao Cheng[5] developed the software of continuous section extraction, and combined with the 3D laser scanning data of Wuhan subway tunnel to carry out the test, and achieved good results. Ge Ding[6] used the method based on the central axis to extract the boundary of the tunnel point cloud, which provides a reference for the production practice of 3D laser scanning. Jianyong Yang[7] carried out operations such as structural section extraction, point cloud data thinning, image matching, splicing and denoising on the collected data and combined with the actual situation, he elaborated the processing method of 3D laser scanning technology in the measurement of structural section of urban rail transit engineering. Yuxiang Hu[8-9] truly showed the spatial form of the subway tunnel.

Taking Yuzhu Tunnel planning and acceptance project as an example, this paper studies the engineering survey and 3D reconstruction technology of tunnel by using 3D laser scanning technology. The method of preprocessing the collected point cloud data and how to reconstruct the 3D model are introduced in detail. The technical route is shown in figure 1.

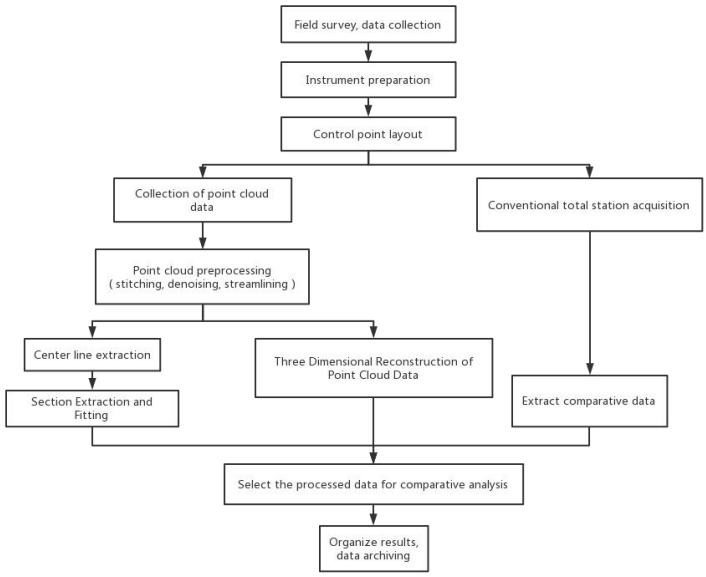


Figure 1. Technology roadmap

2. 3D Laser Scanning Technology and Point Cloud Data Processing

2.1. Working Principle of 3D Laser Scanning Technology

The main part of 3D laser scanner is composed of laser emitter, receiver, prism and so on. It is driven by a rotatable motor and emits laser at the same time， the laser is reflected back to the system through the target object. The scanner receives the laser and analyzes and calculates the 3D coordinate information and reflection intensity information of the surface of the target object.

The calculation principle is through the oblique distance S between the scanner and the target object, according to the known laser pulse horizontal measurement Angle α and vertical measurement Angle β. This is shown in figure 2.

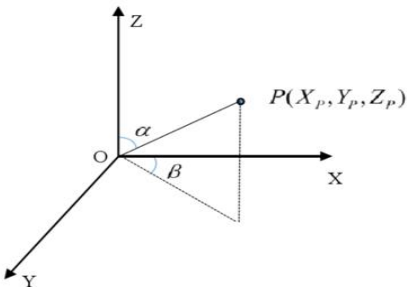


Figure 2. 3D laser scanning positioning principle

Thus, the coordinates of a point in the scanning coordinate system can be calculated. The calculation process is as follows:

$$\begin{aligned} X_P &= S \times \cos\alpha \cos\beta \\ Y_P &= S \times \sin\alpha \cos\beta \\ Z_P &= S \times \cos\beta \end{aligned} \tag{1}$$

2.2. Data Acquisition and Data Processing of 3D Laser Scanner

Patch cloud data using the method of this article is introducing the external reference[10,11], measured with conventional total station instrument construction coordinates of the center of the target coordinates, and then USES the 3D laser scanner data, using software to point cloud data and target coordinate registration for coordinate conversion, and some data to be included in the coordinate system of construction. This is shown in figure 3. Data processing includes point cloud stitching, point cloud denoising and data reduction.

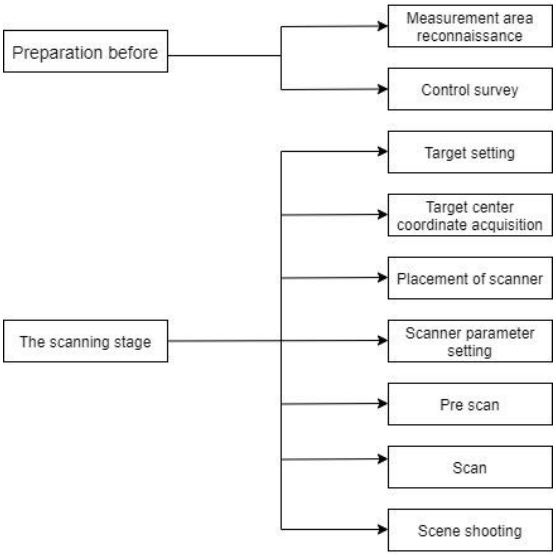


Figure 3. 3D laser scanner operation flow

3. Tunnel Visualization and Precision Analysis

3.1. Extraction of Tapped Lines in Tunnels and Tunnel Section and 3D Modeling

In general, the central axis is extracted according to the characteristics of the tunnel[12,13]. The tunnel axis extraction is shown in figure 4. At the beginning, the point cloud data of the tunnel should be adjusted to be orthogonal to the direction of the Y-axis, and then the point cloud data should be projected onto XOY and YOZ planes to obtain two groups of point clouds, and then the central axis coordinates of the two planes can be obtained according to the geometric principle. The central axis of the tunnel can be obtained by curve fitting the coordinates of the two groups. The disadvantage of this method is that the midline is fitted in both directions, so the integrity of the midline is lost. In this paper, the project is used to extract the central axis of the tunnel with a line to replace the curve, the tunnel is segmented, the central axis of each section of the tunnel is considered to be a straight line, and then the establishment of the tunnel transverse and longitudinal normal plane, two normal plane intersection line is the spatial axis of the tunnel. The detailed process is:

First, the point cloud data of the tunnel shall be adjusted to be consistent with the positive direction of the Y axis. Two groups of point clouds can be obtained by projecting the point cloud data on the xoy and YOZ planes. The boundary point coordinates can be fitted by the linear equation:

$$\begin{cases} a_1x + b_1y + c_1z = 0 \\ a_2x + b_2y + c_2z = 0 \end{cases} \quad (2)$$

A1, B1, G1 represent the coefficients of the linear equation of the boundary point of the plane xoy, and A2, B2, C2 represent the coefficients of the linear equation of the boundary point of the plane YOZ.

The normal vector of the center line of the point cloud projection on the plane xoy and the plane of the two boundary lines can obtain the normal vector m including the center line of the tunnel and perpendicular to the point cloud plane. According to this principle, the normal plane n of the point cloud perpendicular to the YOZ plane and including the center line of the tunnel can also be obtained. The intersection line of these two planes is the space center axis of the tunnel.

Let the two normal vector equations including the central line of the tunnel be:

$$\begin{cases} A_1 + B_1 + C_1 + D = 0 \\ A_2 + B_2 + C_2 + D = 0 \end{cases} \quad (3)$$

cross multiply the two normal planes to calculate the intersection vector of the central axis of the tunnel, m= {A1, B, C}, M = {A2, B2, C2}, then:

$$\begin{aligned} m \times n &= \begin{vmatrix} l & j & k \\ A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \end{vmatrix} = \begin{vmatrix} B_1 & C_1 \\ B_2 & C_2 \end{vmatrix} l - \begin{vmatrix} A_1 & C_1 \\ A_2 & C_2 \end{vmatrix} j + \begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix} k \quad (4) \\ &= (B_1C_2 - B_2C_2)l + (A_2C_1 - A_1C_2)j - (A_1B_2 - A_2B_1)k \\ &= (E, F, G) \end{aligned}$$

Find the intersection point of the plane xoy and the central axis of the tunnel, and take Z = 0, The two equations are solved simultaneously:

$$\begin{cases} A_1x + B_1y + D_1 = 0 \\ A_2x + B_2y + D_z = 0 \end{cases} \quad (5)$$

It can be calculated that:

$$\begin{cases} x = \dfrac{B_1D_2 - B_2D_1}{A_1B_2 - A_2D_1} \\ y = \dfrac{A_1D_2 - A_2D_1}{A_2B_1A_1B_2} \end{cases} \quad (6)$$

It can obtain the intersection straight line of the normal plane, that is, a point on the central axis of the tunnel is:

$$\left\{\frac{B_1D_2 - B_2D_1}{A_1B_2 - A_2B_1}, \frac{A_1D_2 - A_2D_1}{A_2B_1 - A_1B_2}, 0\right\} \quad (7)$$

the coordinates of the three points are denoted as (h.po).

The straight line equation passing through this point and the direction vector is (E, F, g) is:

$$\frac{x - H}{E} = \frac{y - P}{F} = \frac{z}{G} \quad (8)$$

connecting the central axes of all sections of the tunnel can obtain the 3D axis of the tunnel as a whole. And extraction results of tunnel section point set is shown in figure 5.

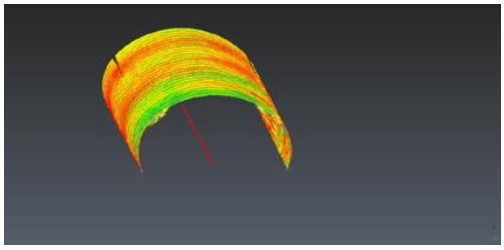


Figure 4. Tunnel axis extraction

Figure 5. Extraction results of tunnel section point set

3.2. 3D reconstruction of Tunnel

In this paper, the tunnel modeling adopts the triangular grid modeling method. Use a triangular grid to connect discrete point cloud data to establish a topological relationship between data. The established triangular grid model can represent the surface structure of the target object, which can achieve the purpose of modeling. The advantage of this method is that it can well show the structural changes of the target object and has a relatively high level of automation. Some software based on this method can be used to build models. The quality point and integrity of point cloud data have an important impact on 3D modeling.

4. Engineering Application

The practical part of this paper takes the Guangzhou Yuzhu Metro Station Tunnel Project as an example. The Yuzhu Tunnel is located in Guangzhou.The three districts of Haizhu District, Tianhe District and Huangpu District start from Xingang East Road, go northeast through the Pearl River, connect to the planned Zhuji Road on the north bank of the Pearl River, go north along the planned Zhuji Road corridor, go down through Huangpu Avenue, and stop south of Zuozhi Chung, Shen Chung. This is shown in figure 6. And the station view inside tunnel is shown in figure 7. The road grade is the main road of the city, with a design speed of 60km/h (locally 50km/h), the planned red line width is 60m, and the two-way six-lane standard. The total length of the main line of the project is 2.594km, of which the tunnel section is 2.454km long.

Figure 6. Yuzhu Tunnel

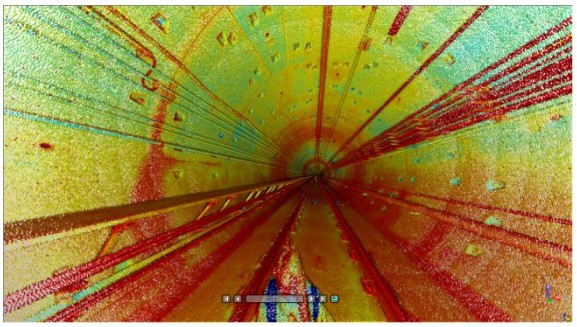


Figure 7. Station view inside tunnel

The point cloud data obtained by 3D laser scanning in the tunnel is based on an independent coordinate system. The coordinates are converted into control points for elevation control measurement and plane control measurement, and then the high-precision control point coordinates are calculated. Using conventional total station acquisition, using the attached wire control network, the control interval is controlled within 30 meters. A total of 6 control points are laid out in the text, of which 4 are encryption points, and 6 target points with construction coordinate system are collected on the encryption points. This is shown in figure 8.

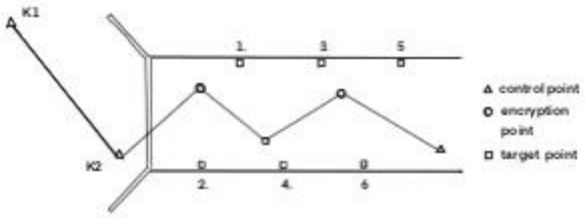


Figure 8. Tunnel control point layout

The tunnel has a long and narrow feature, so when collecting point cloud data in the tunnel, it is necessary to set up multiple scanning stations. The distance between the stations will directly affect the scanning accuracy. According to the characteristics of the 3D laser, the long scanning distance will lead to low accuracy and low density of the point cloud, resulting in the distortion of the geometric information of the point cloud and the quality of the point cloud data. Therefore, selecting the appropriate interval can ensure the quality of the scan. The tunnel target ball layout diagram is shown in figure 9. The inner diameter of the tunnel and the maximum incident angle determine the distance between the scanner stations, assuming that the station is on the central axis of the tunnel and that the measurement point with the largest incident angle is located at location B. According to the geometric relationship, we can draw the following conclusions :

$$\theta_{\max} = \arctan \frac{S}{D} \tag{9}$$

$\theta_{\max}$: The maximum incident angle of the scanning range of the measuring station.

S: Station spacing.

D: Tunnel inner diameter.

Studies have shown that errors increase dramatically when the incident angle is greater than 65°. When $\theta_{\max}$ = 65°, S = 2.14D. The point cloud data of the tunnel in this paper uses iterative closest point (ICP) registration with at least 20 % ~ 30 % overlap. Due to construction, the station may not be accurately erected on the central axis of the tunnel, so S = 1.3D.

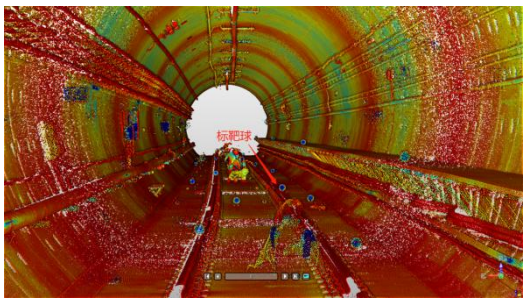


Figure 9. Tunnel target ball layout diagram

Finally, the layout of control points and control network is carried out to check whether the scanner equipment has electricity. The 3D laser scanner is for panoramic scanning, the scanner is mounted on the tripod and the instrument is leveled, the resolution is set, and the panoramic scanning is carried out. After scanning at one site, store the point cloud data and move the 3D laser scanner to the next site for scanning.

4.1. Fish bead tunnel point cloud data preprocessing

Import the collected data into the computer, open the Trimble RealWorks 11.3 software, establish a project folder, import the data scanned by the target ball into the software and use the registration function to extract the target ball. This is shown in figure 10. Each measured site coordinate system has its own independent coordinate system, so you need to collect 3D coordinate information for each site by splicing in the same coordinate system. This is shown in figure 11.

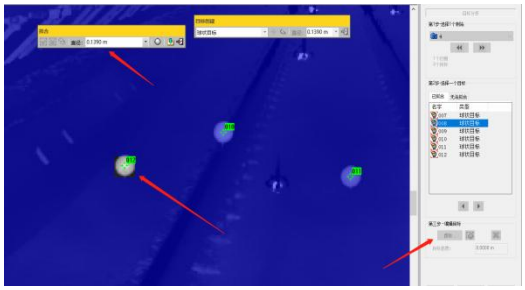


Figure 10. Extraction and registration of target ball

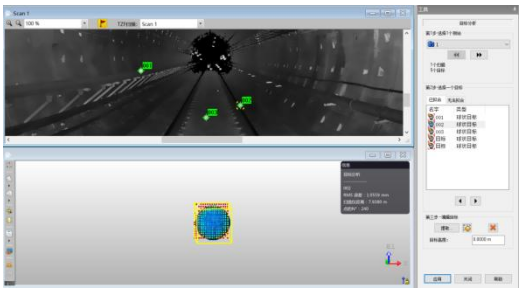


Figure 11. Extraction and registration of target ball

This project uses software automatic splicing to splice the data of each site, arrange a certain number of target balls evenly in the scanning area, and then collect the point cloud data of each site. This is shown in figure 12. Finally, the point cloud data processing software Trimble RealWorks 11.3 is used for splicing processing. This is shown in figure 13.

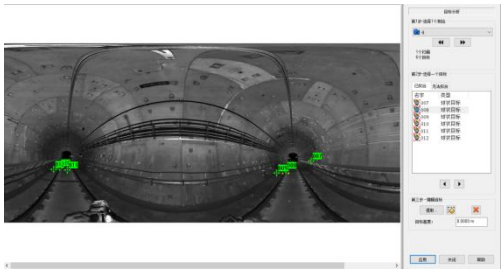


Figure 12. Extraction of target ball registration

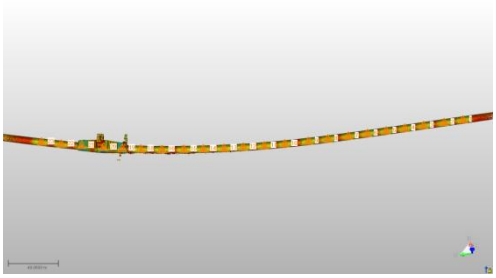


Figure 13. Point cloud effect diagram after splicing

Scanning the tunnel during construction, the dust in the tunnel and the surface of the instrument will undergo mechanical vibration, and the point cloud data will generate noise. There are many temporary facilities in the tunnel. The collected point cloud data includes unnecessary information such as machinery, trolleys, ducts, wires and cables in the tunnel, which will seriously affect the point cloud extraction. To obtain the accuracy of the target data, the original data must be denoised. The noise points can be divided into drift points, isolated points, mixed points and redundant points. The isolated points generated by mechanical vibration in the tunnel and the redundant points generated in the temporary facilities can be manually deleted. However, manual deletion will delete data incorrectly because of human subjective judgment. Therefore, some specific algorithms need to be used for deletion. In this paper, the point cloud data processing software Trimble RealWorks 11.3 is used to delete the noise points in the tunnel, but the mixed points need to be deleted by the corresponding algorithm.

When scanning the tunnel during construction, mechanical vibration will occur on the dust inside the tunnel and the surface of the instrument, and noise will be generated from the point cloud data. There are many temporary facilities in the tunnel, and the collected point cloud data includes unnecessary information such as the machinery, trolley, air duct and electrical cable in the tunnel, which will seriously affect the point cloud extraction. In order to obtain the accuracy of target data information, the original data must be denoised. Noise points can be divided into drift points, isolated points, mixed points and redundant points. Isolated points in tunnels caused by mechanical vibration and redundant points in temporary facilities can be manually removed. However, manual deletion may result in incorrect deletion due to subjective judgment. Therefore, you need to use some specific algorithms to delete. In this paper, Trimble RealWorks 11.3, a software matching with 3D laser scanner to process point cloud data, is used to delete the noise points in the tunnel, but the hybrid points need to be removed by the corresponding algorithm. This is shown in figure 14.

After processing the point cloud data, it is necessary to check the errors generated in the process of data processing, check the stitching accuracy of two adjacent stations, and check whether the sections between local point clouds are fused through slices. Figure 15 and figure 16 are the schematic drawings of cross-sectional sections.

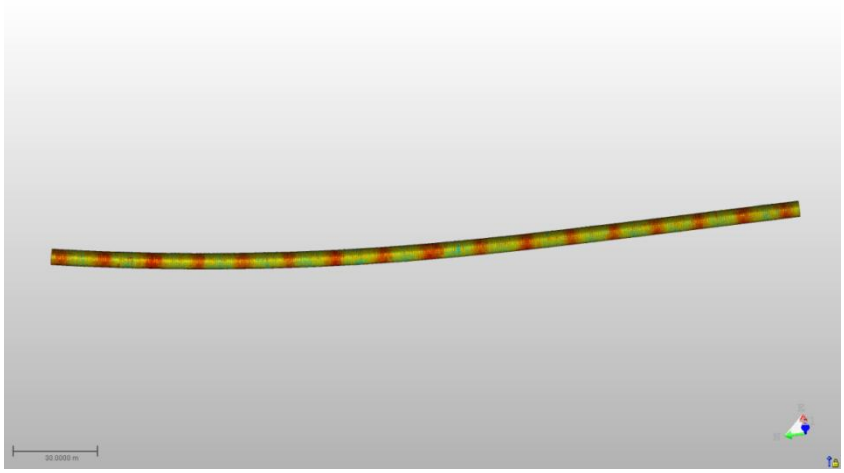


Figure 14. Point cloud data with noise removed

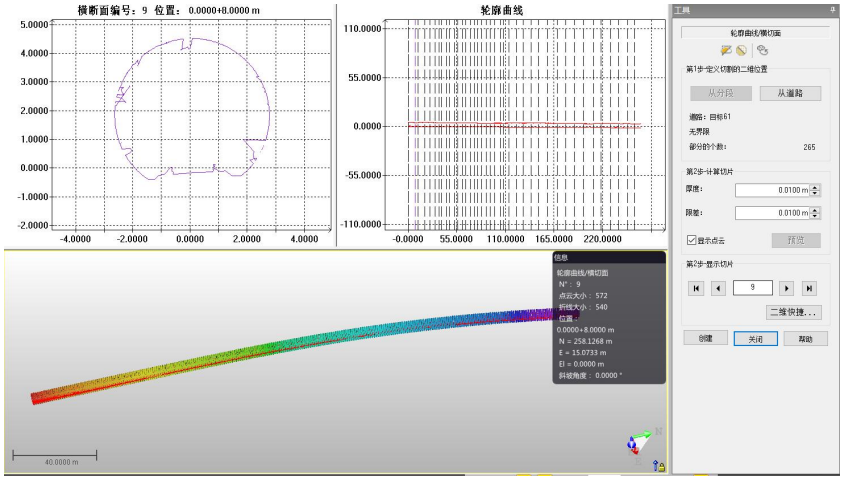


Figure 15. Cross section intention(multiple sections)

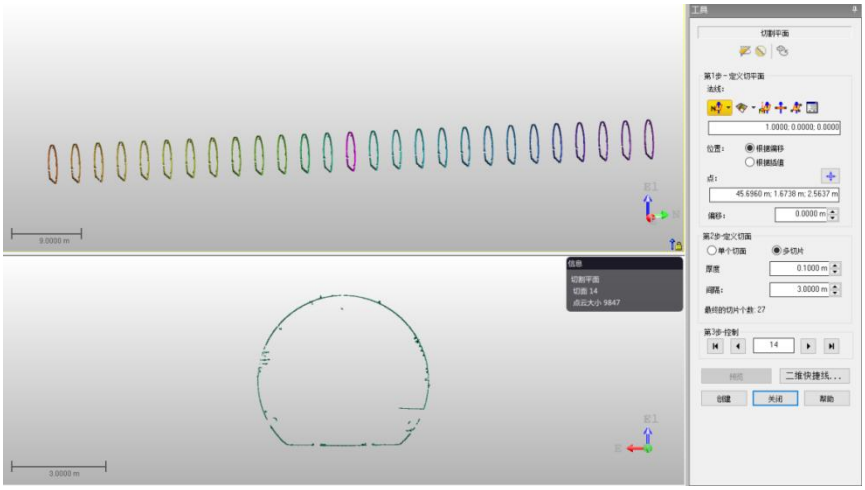


Figure 16. Cross sectionintention (single section)

4.2. 3D modeling of point cloud data

The main purpose of point cloud data acquisition by laser scanning technology is to construct digital and accurate 3D models of buildings. 3D laser scanning technology can improve the efficiency and accuracy of point cloud data modeling. The data used in tunnel reconstruction in this paper are point cloud data acquired by 3D laser scanner, and the point cloud data modeling method adopted is based on the extraction of feature parameters. Figures 17 and 18 are the application of the data in the StetchUp. The 3D model of the software is shown in figure 19.

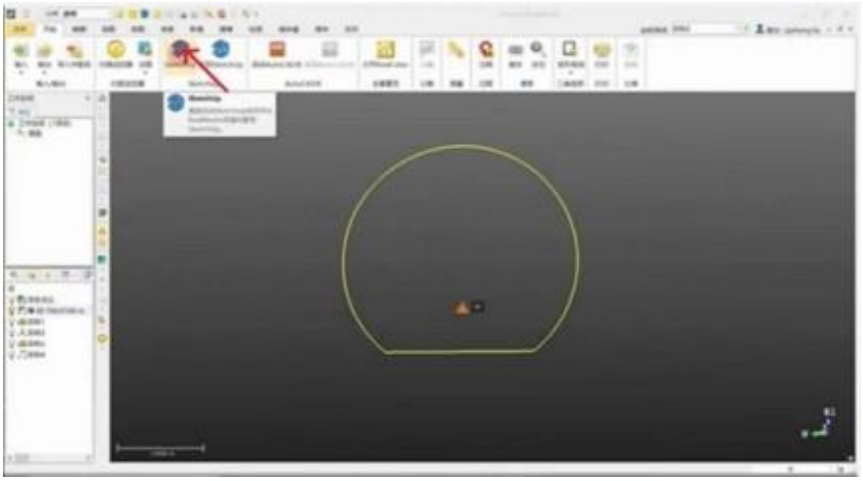


Figure 17. Import the StetchUp tunnel

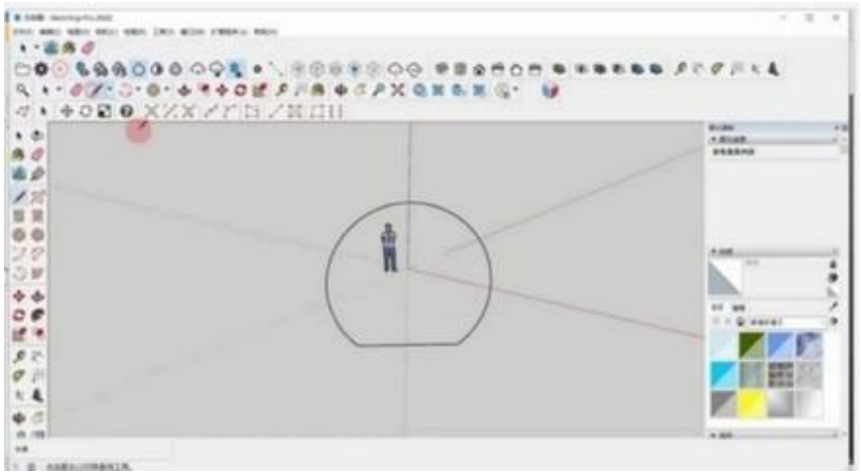


Figure 18. Section of the tunnel in StetchUp

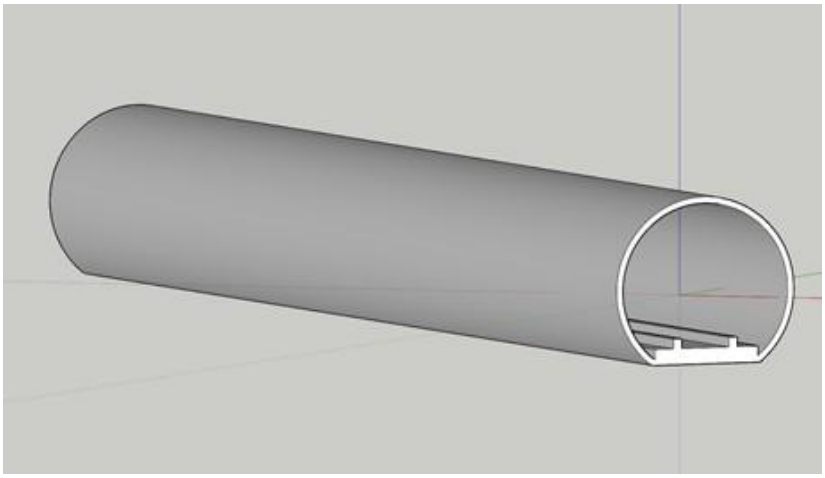


Figure 19. 3D model of software

5. Conclusion

This paper studies the tunnel engineering of point cloud data acquisition and modeling of 3D laser scanning technology, through the trimble X7 3D laser scanner to scan tunnel to get point cloud data, and to the point cloud data splicing, denoising and streamlining, such as pretreatment, extraction and fitting and tunnel sections of 3D reconstruction, to build a tunnel model. The main research results of this paper are as follows : 1) The ICP algorithm is used to splice the point cloud data acquired by the 3D laser scanner, which solves the problem that the missing point cloud of the target sphere and target boundary leads to the deviation of the fitting center point, which affects the splicing accuracy. The point cloud data are denoised and reduced. 2) Two normal plane equations including the central axis of the tunnel can be obtained by projecting the point cloud on the YOZ plane and XOY plane. The intersection lines of the two planes can be used to obtain the spatial axis of the tunnel, and the continuous extraction of sectional point cloud can be realized according to the central axis. 3) Study the application of 3D laser scanner in 3D data acquisition of tunnel engineering, preprocess the collected data with point cloud data, and analyze and verify the complete 3D geographic information data of subway station.

Acknowledgments

This research content of this paper is supported by Guangdong Science and Technology Innovation Strategy Special Funds(pdjh2022b0392, pdjh2022b0394)and College Students' innovation and Entrepreneurship Projects(C2106001310, 202211106008).

Reference

[1] Lin Jingfeng, Yu Jiayong, Tian Maoyi, Xu Fei, Zhou Maolun. Central axis extraction and overall deformation analysis using tunnel laser point cloud[J]. Remote Sensing Information, 2021, 36(01): 94-101.
[2] Hou Gaopeng. Research on processing method of tunnel deformation point cloud data based on 3D laser scanning [D]. Southwest Jiaotong University, 2021.
[3] Liu Zhihai. Research on Data processing method of tunnel laser 3D scanning monitoring measurement [J]. Railway Construction Technology,2020(08):145-148.
[4] Liu Min. Application research of mobile 3D laser scanning technology in tunnel scanning [J]. Geomatics and Spatial Information, 2022, 45(02): 198-202.
[5] Cheng Shuhao, Zou Jingui. Continuous extraction of metro tunnel cross section [J]. Bulletin of Surveying and Mapping,2020(S1):45-50.
[6] Ding Ge, Yan Lishuang, Peng Jian, Zhu Chuanguang, Huang Xueting. Cross-sectional extraction

of tunnel point cloud based on RANSAC algorithm [J]. Bulletin of Surveying and Mapping,2021(09): 120-123.
[7] Yang J Y. Research on 3D laser scanning technology in urban rail transit tunnel section measurement [J]. Southern Agricultural Machinery, 222,53(06): 128-130.
[8] Hu Yuxiang, Fan Shanshan, Wang Zhi, Meng Qingnian. Research on application of station 3D laser scanner in metro tunnel section detection [J]. Urban Survey, 2020(01): 130-134.
[9] Ding Xiaobing, Gao Zhiqiang, Yang Kun. 3D Point Cloud Reconstruction of Subway Tunnel Based on Inertial Navigation System and CP iii Control Points [J]. Bulletin of Surveying and Mapping, 2021(09): 112-115+129.
[10] BAO Z Y. Visualization analysis of 3D laser scanning in subway tunnel monitoring [J]. Geomatics and spatial information,2021,44(03):183-187.
[11] Chen Junming. Research on Building Elevation Extraction and Modeling Based on Ground Li DAR Data [D]. East China University of Technology, 2017.
[12] Zhu Wenhuan. Research and Application of Point Cloud Data Preprocessing Optimization Algorithm [D]. Guangdong University of Technology, 2016.
[13] Wang Longfei, Hu Haifeng, Lian Xugang. Rapid Cross-sectional Extraction and Deformation Analysis of Railway Tunnel based on whole-station scanning point cloud [J]. Bulletin of Surveying and Mapping, 2019(05): 55-59.