International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 8 Issue 6, Nov-Dec 2024 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Artificial Intelligence Technology Disenchantment and Data Compliance Issues

Liu Guangpu, Bi Guoyi, Yang Wenhao, Ma Ping

School of Law, Beijing Wuzi University, Beijing, China

ABSTRACT

With the rapid development of artificial intelligence technology, China is playing an increasingly important role in the process of compliance in the development of artificial intelligence. In the past, people's understanding of the risks caused by artificial intelligence was limited to data leakage, which led to a series of questions such as personal information leakage, copyright infringement, data barriers, data fraud, illegal data trading, algorithm bias, intervention in social affairs, increased psychological pressure on trainers, generation of false and harmful information, academic fraud, criminal tools, laziness of users, social division, etc. Unlike most articles that focus on AI application results, this report takes a unique approach. Through extensive literature review, rigorous data analysis and in-depth comparative research, it not only analyzes the core generation principles of AI technology, but also makes a profound interpretation of them. After interpreting the principles of AI technology, many legal compliance and security risks that may arise in the pre-training process are analyzed. It aims to provide valuable reference and guidance for relevant policymakers, researchers, and the public." Through this report, readers can fully deconstruct the principles of generative AI technology, as well as possible legal compliance issues and security risks, and contribute to the healthy development of AI technology.

KEYWORDS: *artificial intelligence, data leakage, data compliance*

How to cite this paper: Liu Guangpu | Bi Guoyi | Yang Wenhao | Ma Ping "Artificial Intelligence Technology Disenchantment and Data Compliance Issues" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-8 | Issue-6, December 2024, pp.352-362, URL: www.ijtsrd.com/papers/ijtsrd71579.pdf

IJTSRD71579

1. INTRODUCTION

1.1. Background Introduction

1.1.1. The development of artificial intelligence at the present stage

A. research status abroad

Globally, patent applications in the field of artificial intelligence are highly competitive. In November 2022, the US company OpenAI launched ChatGPT chatbot, which quickly accumulated a large number of users worldwide. As an outstanding representative of generative artificial intelligence, ChatGPT has led artificial intelligence from perceiving and understanding the world to generating and creating the world[1], marking the [2]of a paradigm shift in the development and application of artificial intelligence technology· Under its influence, Google, Microsoft, Baidu and other giants have also carried out generative artificial intelligence research and development work. These companies are actively engaged in research and innovation in various fields such as machine learning, self- natural language processing, and computer vision. According to data from the World Intellectual Property Organization, in the 10 years from 2014 to 2023, the global number of patent applications related to generative artificial intelligence reached 54000, of which more than 25% of the patents were published in 2023. Especially after the advent of the deep neural network architecture on which the large language model (LLM) is based, the number of patents related to generative artificial intelligence has grown rapidly.

Since its birth, the chatbot of OpenAI Company in the United States has marked the arrival of the era of strong artificial intelligence. Different from the simple data accumulation and search in the era of weak artificial intelligence, strong artificial intelligence has the ability of integrated analysis. OpenAI's ChatGPT can even pass the U.S. judicial exam and the final exam in colleges and universities. The follow-up GPT-4 is even more amazing. It can draw excellent paintings that conform to human aesthetics according to the user's instructions. It is even difficult for ordinary people to find AI paintings. Not only that, GPT-4 can pass the lawyer professional qualification examination in Japan, the doctor qualification examination, and even achieved excellent results in these examinations. Books independently written, translated and proofread by artificial intelligence in South Korea were officially published on February 2.

B. domestic research status

China has made remarkable progress in the 1 field of multi-modal generative AI. Multi-modal generative AI systems can process multiple input information such as text, sound, melody and visual signals, and integrate them for comprehensive understanding. This technology is expected to enrich the content and level of literary works, and bring a variety of sensory experience to the audience. In addition, China is actively promoting the combination of quantum computing and AI, using the special properties of quantum computers (such as quantum superposition and quantum entanglement) to accelerate machine learning and optimization algorithms, so as to achieve more efficient and accurate AI applications.

At present, the Chinese government continues to launch policies to support the development of digital economy, converged communications and artificial intelligence, and encourage the development and commercialization of related technologies. From finance, medical care, education to intelligent manufacturing, smart cities, etc., AI technology is penetrating into all walks of life. Data show that China's AI patent applications accounted for 64% of the world, ranking first in the world. Especially in the field of generative AI, the number of invention patents submitted by China far exceeds that of other countries.

1.1.2. Technical disenchantment

With the introduction of the ChatGPT chatbot by OpenAI in the United States, major domestic giants have successively introduced distinctive ChatGPT variations such as Bean Bags, Kimi, Wenxin Yiyan, etc.. It can be seen that artificial intelligence has officially entered the era of strong artificial intelligence, and can greatly promote the development of productivity and create more considerable economic benefits in the near future. However, although most people know the power of artificial intelligence and even benefit a lot from using artificial intelligence, due to the lack of understanding of the working principle of artificial intelligence, there will be excessive myth that AI is omnipotent, or excessive contempt for AI that AI is just a weak artificial intelligence search engine that is not original. Therefore, it is necessary to first look at the principles of artificial intelligence. If artificial intelligence is to be put into practice, it needs to be pre-trained. Pre-training refers to the first stage of training using general tasks and large-scale data, so that the machine learning model can learn parameters [3]with strong generalization. After pre-training, the cornerstone model (foundation models) can improve the performance of the model and the degree of generalization of the model through the learning of common language knowledge, can be said to be the production of AI "ten years of work" 5[4]. In the following, I will conduct an in-depth perspective one by one in the order of the general artificial intelligence training process.

A. Get data

The current methods for obtaining data generally include: data sets, because they are public resources, so the data quality is high and the cost is low. As well as seeking outsourcing platforms, such as Ali crowdsourcing and Baidu crowdsourcing, the advantage is large scale, the lack is high cost. Can also own collection, high quality can be customized, but the efficiency is very low. Due to the limitations of the data obtained by the above methods, some companies will use crawler tools, which have the advantage of fast speed and low cost. The principle is to write a program to Baidu Google and other browsers to pick up picture audio. The disadvantage is that this method is easy to fall into the infringement storm.

B. Basic data processing

The basic processing of data is actually the specific work of the data labanner. There is almost no threshold for data labelers. Everyone can become data labelers with a little training. Moreover, data labelers need a lot of data labelers and their wages are not high, which is comparable to assembly line workers in the new era. The most common annotation methods used by data annotator are as follows: (1) pull box, which is suitable for text speech image. (2) Classification labeling is actually labeling. (3) Insertion annotation is more refined and is suitable for images such as faces or bones. (4) voice audio annotation. (5)3D cloud data annotation, the operation is slightly difficult.

After data processing, it is also necessary to carry out inspection, usually cross quality inspection and sampling sampling inspection, to check whether the labeling is wrong.

C. Feature Engineering

Feature engineering simply translates human language into a digital language that is easier for machines to recognize. after data processing, the steps of feature engineering are transformed into content that is easily recognized by the machine for gpt learning.

D. Machine learning

Machine learning, this process is the process of automatically analyzing the data to obtain a preliminary model and continuously using algorithms to optimize the model. Algorithm as a computer science in the core concept, refers to the problem-solving program accurate and complete description, which can according to the input data or conditions, through a series of calculations and operations, and ultimately get the desired output results, and generative algorithm is the core of the training sample distribution modeling, and then based on the model to generate new samples[5].

There are a variety of algorithms for different purposes, and the common algorithms are the following.

1. Supervised learning, supervised learning is the most common and basic way of learning. It is by inputting a lot of data to AI, and the next command, AI gets a result. AI will run its own results and constantly get checked to get the final results. But supervised learning is also divided into two types, the is regression and the is classification. For input continuous data, the results are targeted, simply a pair of , one data corresponds to a target result. For example, forecast house prices, every day there is a forecast house price data. The other is classification. His results are discrete. One data does not correspond to a target result. For example, it can predict whether a tumor is benign or malignant. Many data input will be divided into two categories. One category of data represents predicting benign tumor and the other category represents predicting malignant tumor. In general, both classes have a clear expectation of the outcome.

2. Unsupervised learning, unsupervised learning is different from supervised learning, that is, there is no predictive expectation value, no data input after the program runs up the evaluation feedback. Simply put, it is to give artificial intelligence sufficient autonomous learning ability, only provide data feeding, let artificial intelligence learn by itself, AI learn what it learns. At present, this way of learning mainly performs three tasks: clustering, association and dimensionality reduction. Unsupervised learning is widely used in image processing, text mining, data compression, clustering analysis and other fields. For example, cluster analysis is a commonly used method in unsupervised learning, which classifies data points into different categories based on their similarity, and can be used for text mining, image processing, etc. Application areas: consumer behavior analysis, astronomical data analysis, anomaly detection, etc. Robot training direction: suitable for scenarios that require autonomous exploration and discovery of environmental structures, such as environmental perception and mapping of driverless cars, or autonomous navigation and exploration of drones in unknown areas.

3. Semi-supervised learning, semi-supervised learning is between supervised learning and unsupervised learning, using a small portion of labeled data and a large amount of unlabeled data for training. This approach aims to improve the predictive power of the model by combining the guidance of the labeled data and the underlying structural information of the unlabeled data. In short, semi-supervised learning is to let the robot have an abstract goal after inputting a large amount of data, but there is only a goal, but the specific goal is not clearly given, AI needs to learn the data by itself, and there is no evaluation feedback function. At present, semi-supervised learning applications are text classification, image recognition, natural language processing and so on. The trained robot can be applied to scenarios that need to process a large amount of unlabeled data and combine a small amount of labeled data for learning, such as smart home robots, which can learn autonomously and adapt to different home environments under the guidance of a small amount of user instructions.

4. Reinforcement learning, which learns the best decision-making strategy through interaction with the environment. The agent takes action in the environment and adjusts its behavior strategy based on the feedback (reward or punishment) of the environment to maximize the cumulative reward. Reinforcement learning does not rely on pre-labeled data, but learns through trial and error. In fact, to put it simply, it is similar to the way of bringing a child to make the machine run on its own. This is a continuous action rather than a flash action. It will observe whether the next step after the machine step is expected, and if it meets the reward, it will go back and go again

until it comes out of the expected behavior. It should be noted that reinforcement learning is a continuous action, and it is often not possible to judge that the program is wrong according to the previous two steps, which is different from supervising learning to directly judge whether it meets the expectation and make adjustments. Application areas of reinforcement learning: robot navigation, autonomous driving, game intelligence, etc. -Robot training direction: focus on cultivating the decision-making and adaptive ability of robots in dynamic environments. For example, robots trained through reinforcement learning can perform path planning, obstacle avoidance, and specific tasks in complex environments, such as automated picking robots in logistics warehouses. Generally in the majority of behavior.

5. Model evaluation, the model is actually the system that the machine obtains through various algorithms after receiving massive amounts of data, which enables the new data to be input and then make predictions about the data. Model evaluation is the accuracy, precision, recall and other data of the detection data prediction.

1.1.3. Research significance

The era of strong artificial intelligence has come impressively, and its tentacles extend to every corner of society, from performing basic repetitive tasks to assisting humans in creating copywriting and drawing paintings, showing unprecedented influence and creativity. However, with the rapid leap forward of AI technology, a series of challenges that were originally hidden in the depths of technology have gradually surfaced and become issues that cannot be ignored.

On the basis of in-depth research, this report deeply analyzes the generation mechanism and working principle of artificial intelligence, and pays special attention to the multi-dimensional problems that may be hidden in the pre-training stage. Through systematic integration, these issues are clearly divided into two core areas: the is the legal compliance issues arising from the AI training process, and the 2 is the security risks arising from the AI-generated content.

In order to effectively deal with these challenges, this report has conducted extensive and in-depth exchanges and discussions with expert teams within artificial intelligence enterprises and senior practitioners in the legal profession. We jointly analyze the feasible path under the current governance framework, aiming to explore a set of comprehensive governance strategies that not only conform to the law of technological development, but also effectively guarantee legal compliance and social security. Through these efforts, we hope to contribute wisdom and strength to the sustained and healthy development of artificial intelligence technology, and jointly promote it to become a powerful driving force for social progress and prosperity.

1.2. Domestic relevant legal regulation

1.2.1. Basic laws and regulations

Personal information protection: When handling personal information, artificial intelligence must abide by the the People's Republic of China Personal Information Protection Law, which clearly stipulates the right of individuals to know and decide on their personal information, and stipulates the principles and systems to be followed when handling personal information.

Network security: the the People's Republic of China network security law provides detailed provisions on the security protection obligations of network operators, personal information protection, network data security and other aspects, providing network security guarantee for the development of artificial intelligence.

Intellectual property rights: the the People's Republic of China Criminal Law, the the People's Republic of China Trademark Law, the the People's Republic of China Patent Law, the the People's Republic of China Copyright Law and other laws and regulations have severely cracked down on the infringement of intellectual property rights, protected the legitimate rights and interests of innovators, and promoted the healthy development of artificial intelligence technology.

1.2.2. Special legislation

Interim Measures for the Management of Generated Artificial Intelligence Services: The Measures came into effect on August 15, 2023, and put forward a series of compliance requirements for providers and users of generated artificial intelligence services, including regulations on algorithm filing, data security, content security, etc.

The Artificial Intelligence Law (Model Law) 2.0: The law clearly distinguishes between the negative list-based artificial intelligence licensing system and the filing system, creating a more relaxed and clear business environment for the AI industry. While the Act currently serves as a model law, it provides an important reference for possible future legislation.

2. Legal Compliance Issues in Artificial Intelligence Training Process

2.1. personal information leakage

The generation mechanism of artificial intelligence poses a fundamental challenge to the protection of personal information. In the process of database construction, which is the cornerstone of its development, it is difficult to meet its growing demand only by relying on public open resources. Especially in the pre-training stage, in order to pursue higher-quality and larger-scale data sets, data labors often have to resort to crawler technology or outsourcing services. In this process, data collection without the explicit consent of the data subject The phenomenon is not uncommon, which directly touches the principle of "clear purpose and minimized scope" emphasized in Article 6 of my country's "Personal Information Protection Law" and the European Union's "General Data Protection Regulations.

The tension between the pre-training phase's thirst for data volume and data protection regulations is particularly significant, as this phase typically requires massive and high-quality data to support the optimization and evolution of AI models, which undoubtedly conflicts with the legal framework for the protection of personal information.

Furthermore, even in the case of theoretical need to obtain personal consent, the actual operation also faces many difficulties. In the current Internet era of information explosion, users frequently encounter all kinds of lengthy and complicated privacy agreements. Whether they can fully understand the contents of the agreements, whether they are willing and have time to read the privacy clauses of each APP one by one, and how to ensure that each user can give consent in a timely and effective manner in the face of massive data demand are all problems to be solved urgently. , in the user agreement, it is usually the default that the user agrees that his information is used for data training. If the user has any objection, he needs to apply to the service provider separately. In this way, under the cumbersome content and operation, the "full knowledge and consent" condition stipulated by the law exists in name only[6].

This series of problems together lead to the phenomenon that is common in the pre-training stage of artificial intelligence: it is difficult for many trainers to fully follow the principle of personal consent in the process of data collection, while a large number of individual users lack sufficient cognition and attention on how their information is used. Therefore, how to build a more sound and efficient personal information protection mechanism while promoting the progress of artificial intelligence technology has become an important topic that needs to be explored by all sectors of society.

2.2. copyright infringement

In the pre-training stage, in view of the urgent demand of artificial intelligence systems for large-scale and high-quality data, these links not only hide the risk of personal information leakage, but also easily touch the sensitive areas of copyright. In order to reduce economic and time costs, artificial intelligence trainers tend to directly collect text materials on social media platforms such as Weibo and Zhihu, as well as rich multimedia content shared by users in applications such as WeChat, Little Red Riding Book and Douyin, including music, pictures and videos. However, while these practices accelerate model learning, they also increase the risk of unauthorized use of other people's original works, which constitutes a potential infringement of copyright.

In addition, when discussing the copyright of artificial intelligence-generated content (such as text, drawings, etc.), we are faced with a complex and cutting-edge legal issue. Within the framework of traditional intellectual property law, intellectual achievements are regarded as the product of human creativity, while artificial intelligence is a non-human subject, and the legal positioning of its creative behavior is not clear. However, it is worth noting that although artificial intelligence is a direct participant in the creation process, the algorithms, models and training data behind it all embody the wisdom and labor of developers and teams. Therefore, whether the concept of "intellectual achievements" should be appropriately extended, or the existing legal interpretation should be adjusted to accommodate and regulate the creative activities of artificial intelligence has become a topic that needs to be studied in depth.

Further, from the point of view of object judgment, the key to evaluate whether the work generated by artificial intelligence enjoys copyright lies in whether its expression is original. However, these standards are often controversial in practical application. On the one hand, artificial intelligence can independently complete some creative content, showing amazing creative ability; on the other hand, its creative process may also involve the integration and reference of existing materials, and it is difficult to distinguish the boundary between originality and citation. Therefore, how to scientifically and fairly evaluate the originality of artificial intelligence works and ensure that it not only promotes technological innovation but also protects the legitimate rights and interests of original authors is a major challenge faced by the legal profession and academia.

2.3. data barriers

Artificial intelligence, in the current business ecology, essentially plays role as a highly strategic tool, and its development is closely linked to business profitability goals. In this context, data, as a core driving force, has naturally become a valuable asset and secret that enterprises compete to protect. The pre-training stage is particularly critical. Major Internet companies have built high walls to prevent data leakage to competitors or external entities. This is not only a defense of commercial interests, but also a positive response to data privacy protection under the legal framework. However, this data blocking strategy, while restricting illegal data capture and protecting personal privacy, also virtually hinders the data exchange and fusion in the pre-training stage, limiting the breadth and depth of data required for artificial intelligence model training, thus affecting the overall progress and potential release of AI technology.

For a single enterprise, although the phenomenon of data island may maintain its competitive advantage in the short term, it is not conducive to the continuous innovation and leadership of AI technology from a long-term perspective. At the same time, this also constitutes a bottleneck for the development of the entire AI industry, slowing down the pace of technology iteration and popularization. Although this conservative attitude among enterprises stems from the desire for technological leadership and the consideration of commercial interests, it also virtually aggravates the barriers to competition in the industry.

At present, there is no perfect solution to the dilemma faced by artificial intelligence in the pre-training stage. However, it cannot be ignored that if this state continues, it will trigger a series of far-reaching negative effects. First of all, the risk of technological monopoly and "technological hegemony" will become increasingly prominent. Technological advantages may use their position to restrict data circulation, aggravate social inequality and information asymmetry, and hinder the release of innovation vitality. Secondly, the "digital divide" on a global scale will further widen, and technological powers and regions may control key technologies and data resources, making it difficult for developing countries and regions to catch up with technology, intensifying the phenomenon of information cocoon houses, and increasing the risk of social fragmentation and unsustainable development.

Therefore, in the face of the problem of data acquisition in the pre-training stage, timely intervention and guidance at the legal level is particularly important. Building a data compliance framework, encouraging and regulating data sharing mechanisms, such as establishing data sharing communities or platforms, and balancing the relationship between data protection and utilization through legal means, is the key to promoting the healthy and compliant development of artificial intelligence. At the same time, it also requires the joint efforts of the government, enterprises, scientific research institutions and all sectors of society to form a consensus and jointly explore a new path that can not only protect data privacy, but also promote data circulation and sharing, in order to achieve the vigorous development of artificial intelligence technology and the overall progress of society.

2.4. data fraudulent

In the key link of pre-training, facing the huge challenge of massive data processing, some outsourcing agencies or artificial intelligence trainers may hide the mystery, quietly mixing a large amount of low-quality or even completely non-compliant forged data, and even deliberately implanting malicious data. These seemingly harmless data actually hide misleading information, and their ultimate goal is to confuse AI's learning path, making it lost in the ocean of information, unable to accurately identify the true and false. This behavior undoubtedly poses a threat to the healthy growth of artificial intelligence that cannot be ignored. The quality defects of training data are like the sword of Damocles hanging on the top of ChatGPT performance. Once false content penetrates into it, the model may deviate from the right track, mistakenly absorb data distribution, and then output misleading information when generating new content, seriously eroding its utility and the cornerstone of public trust.

In addition, malicious attackers are also good at using adversarial data, which is a hidden weapon, to carefully construct seemingly harmless but hidden data samples. Through minor changes, AI's recognition and classification capabilities can be easily disintegrated, resulting in confusion of algorithm logic and wrong judgment. What is more worrying is that they may also quietly integrate carefully forged data into the database, inducing AI to establish a wrong learning mechanism in the learning stage, making the AI system unable to operate effectively in practical applications, and the performance of the algorithm is greatly reduced, seriously affecting the user experience and application effect.

In view of the rapid development of AI technology, we have clearly realized that any neglect of data security is a hidden time bomb, and the consequences of its outbreak will be incalculable. Therefore, it is an urgent moment for us to accelerate the construction of a sound legal system and comprehensively enhance the awareness of AI developers' data security and privacy protection. If we miss this opportunity, we may have to face the painful reality in the future and make up for today's negligence at a high price. To this end, it is urgent to enhance data security management to a strategic height, aiming to ensure that technological progress can truly benefit individuals and society, rather than at the expense of security and stability, and jointly protect the healthy future of AI technology.

2.5. illegal data transactions

Similarly, the scarcity of data resources in the pre-training phase has unfortunately given rise to the proliferation of illegal data transactions. Some enterprises and individuals, driven by interests, do not hesitate to use illegal means to seize and sell personal information. Because of its concealment and complexity, such behaviors are often separated from effective tracking and supervision. The invisibility and easy replication of the data itself poses a challenge to the traditional trading rules and opens up a gray area for illegal activities. Some enterprises or individuals, in the face of the temptation of interests, ignore the law, take advantage of the loopholes of data characteristics, wantonly carry out illegal transactions, which also exposes the shortcomings of the current legal system in related fields.

In the context of the vigorous development of AI technology, illegal data trading has become a major problem to be solved urgently. It is not only a gross violation of personal privacy, but also a profound threat to national security and market economic order. Such transactions often involve the illegal collection, trading and abuse of sensitive personal information, such as telephone numbers, residential addresses, financial accounts and other private information, the impact of which is far-reaching and cannot be ignored.

The Research Report on Global Data Transaction Practice, Industry Norms Status and Policy and Legal Issues profoundly reveals the global disputes and discussions on data property rights. However, the current global legislation on data property rights is still at the level of theoretical discussion, and no country has yet Can clearly define and protect data property rights in law, which undoubtedly sets up many obstacles for the effective promotion of data protection. Therefore, strengthening the legislation of data property rights and improving the supervision mechanism of data transactions have become an urgent need to protect personal privacy, maintain national security and promote the healthy development of the digital economy.

2.6. algorithm bias

The so-called algorithm discrimination refers to the systematic errors in the decision-making, operation and other processes in the algorithm field due to the bias of the algorithm designer or the deviation of the training data, and then the unfair algorithm conclusion[7]. The so-called algorithm bias refers to the kind of biased processing [8]of data according to the preferences of the data subject in violation of recognized norms and ethics.

First, in the pre-training phase of the data selection process, which naturally implies inherent bias. This difference mainly stems from the difference of Internet information ecology at home and abroad. Foreign countries tend to rely on widely recognized databases such as Wikipedia, while China integrates more data resources of Baidu, news and various Chinese websites. This tendency of data selection inevitably leads to deep differences in cultural understanding and value orientation of artificial intelligence systems at home and abroad, which has become an irreconcilable gap. In short, although artificial intelligence is essentially the absorption and application of knowledge by machines, the humanistic knowledge it carries inevitably permeates the unique educational and cultural background and value orientation of their respective countries and regions. thus inherently carry a certain cultural and value prejudice.

Furthermore, the human intervention in the pre-training stage aggravates the complexity of the biased treatment. In key links such as data cleaning and labeling, due to individual differences in understanding, cognitive limitations, and the intervention of subjective preferences, even in the face of the same data set, may produce very different processing results. These differences not only directly shape the output characteristics of AI models, but may also amplify or introduce new biases in some cases. It is particularly noteworthy that although some biased processing can be avoided through rigorous process design, in practice, the personal will and preferences of data labelers are often difficult to completely eliminate, and even in some cases are unconsciously magnified.

In addition, the harmful elements such as errors, outdated information, biased content, hate speech, violence and pornography hidden in the training data are not only the erosion of data quality, but also a serious challenge to the fairness and accuracy of the algorithm. These negative factors may quietly affect the understanding framework, feature selection and problem construction process of the algorithm, making the decision-making and judgment of the AI system deviate from the objective and fair track. Therefore, in the pre-training stage, we must attach great importance to and effectively deal with the bias caused by these human factors and the data itself, so as to ensure that the healthy development and wide application of artificial intelligence technology can really benefit the society.

3. Security Risks in the Development of Artificial Intelligence

3.1. intervention in social affairs

The role of artificial intelligence trainers is irreplaceable. They are close-fitting guides from immature to mature in AI's growth journey, shaping AI's intelligence and ability like carefully cultivating children. However, in this stage full of hope and potential, there are also risks and challenges that can not be ignored. The pre-training period, originally a golden period for AI to learn the foundation and build a cognitive framework, may also become a hotbed for some unruly people. They take the opportunity to pour a large amount of bad data into AI in an attempt to use it for espionage activities, undermine election justice, manipulate political public opinion, and even steal state secrets, seriously threatening the country's political stability and social public interests. The precedent of using data analysis to interfere with US elections on Facebook platforms is a warning light of this risk.

Further, in-depth investigation shows that many developers in the field of AI generally recognize that AI trainers, through carefully designed training data sets, actually have a strong ability to intervene in political ecology, economic order, cultural landscape and even broader social affairs. If this ability falls into the hands of users who lack a sense of responsibility or have ulterior motives, it is tantamount to giving them the key to manipulate the pulse of society, which may lead to a series of far-reaching negative effects. Specifically, through in-depth mining and analysis of subtle patterns and trends in AI training data, malicious users can accurately locate the weak links of political propaganda or economic strategies, and implement efficient information manipulation strategies to guide the direction of public opinion and even influence market fluctuations. At the cultural level, deep insights into data can also be misused to promote the widespread dissemination of specific ideologies or values, quietly changing the cognitive framework and behavioral paradigm of society.

Therefore, ensuring the transparency, morality and legitimacy of the AI training process is not only an inevitable requirement of technological progress, but also the key to maintaining social stability and promoting healthy development. We need to establish a sound regulatory mechanism, strengthen the review and management of AI training data, and at the same time raise public awareness of AI ethics and safety, and jointly protect the health ecology of this emerging technology field.

3.2. trainers increased psychological pressure

In the pre-training stage, data labelers shoulder heavy and sensitive tasks. They need to process tens of thousands of data samples in detail. These data contents are numerous and complicated, including intuitive and easy-to-understand positive images such as cars and fruits, as well as uncomfortable and even shocking elements such as violence, racial discrimination and gender bias. Every day, data labelers need to constantly shuttle in the ocean of these multimedia information, including pictures, audio, video and text, and so on. The heavy workload and sensitive content are undoubtedly a severe test of their psychological quality and professional ethics.

Long-term exposure to such sensitive and harmful information makes data labors vulnerable to mental health erosion, resulting in mental health problems such as mood swings, anxiety, depression and even post-traumatic stress disorder, which are like invisible shackles and seriously hinder their right to enjoy a healthy life. In addition, the diversity of this information may also bring profound cultural impact, forcing developers to face up to ideas that conflict with their own values and cultural background, causing inner cognitive conflicts and emotional struggles.

What is particularly worrying is that when data labsters experience psychological discomfort due to long-term stress, they may not be aware of the changes in their own state, and then inadvertently reflect the distortion of negative emotions and personal values into the AI training process. As a learning machine, once AI absorbs this bad information, its output will inevitably be biased and misleading, which not only damages the fairness and reliability of AI system, but also may in turn aggravate the psychological burden of data labors, forming a vicious circle and causing continuous negative impact on both sides.

3.3. generate false harmful information

In the pre-training stage, the recruitment of data labors covers a wide range of sectors of society. This measure significantly reduces the entry threshold for this position, but it may also bring about uneven data quality. Due to the different professional backgrounds and skill levels of the data lablers, the data they process is not completely accurate and of high quality, which undoubtedly inputs potential impurities into the AI system.

Further, the complexity of AI technology and the existence of algorithmic black boxes make it difficult to intuitively and comprehensively understand how AI absorbs knowledge and generates output. This technical opacity virtually increases the risk of AI generating false or harmful information. When users rely too much on AI-generated data, they may be misled and affect the correctness of decision-making.

Take the AI system launched by Google in 2023 as an example. When asked, "What are the new discoveries of the Webb Space Telescope that are suitable for sharing with 9-year-olds?", its wrong answer-"The Webb Space Telescope has taken the first pictures of planets outside the solar system"-is a warning. In fact, these historic achievements should be attributed to the European VLT platform. If such misleading information is spread without verification, it will undoubtedly have a negative impact on the public's scientific cognition, especially children's astronomical enlightenment education, and hinder the healthy development of their scientific thinking.

Therefore, while enjoying the convenience and efficiency brought by AI, we must maintain a cautious attitude and verify and screen the information generated by AI to ensure its authenticity and reliability, so as to minimize misleading and promote the harmonious coexistence of science and technology and society.

3.4. academic fraud

The core of returning to the creative origin of AI is to use the vast ocean of data, rich academic works and profound research papers as nourishment, and to build a unique logical intelligent system through the learning and refinement of sophisticated algorithms. In process, AI will inevitably absorb and cite the best of these documents to form part of its output.

In January 2023, "Study.com" released remarkable data from a survey of students over the age of 8 in the United States: among the 1,000 students interviewed, as many as 89% admitted to using ChatGPT to assist in completing their homework, more than half of them even used it directly to write papers, and even 22% of the students relied on the tool to generate the outline of their papers. This flood of data not only reflects students' strong dependence on emerging technologies, but also quietly uncovers the serious challenges facing academic integrity, pointing to the core of the complex issue of accountability.

With its excellent text generation ability, ChatGPT has opened up a shortcut for students to quickly obtain information and construct articles. However, behind this convenience lies the risk of potential erosion of independent thinking, original writing skills and deep understanding of knowledge. In the long run, the quality of education may be impacted, and the yardstick for assessing students' true abilities and knowledge levels may also shift.

Of particular concern is that when AI tools go deep into academic fields, the boundaries of responsibility become blurred and complex. Who is responsible for this "smart" output-the students, the developers, or the education system itself? The raising of these issues forces the community to re-examine the boundaries of academic integrity, responsibility and technical ethics.

3.5. tools of crime

In the pre-training stage of AI, although the intake of massive data information is an indispensable part of its learning and growth, "information overload" is not always beneficial, especially when the quality of data is uneven. In addition to the obvious violence and pornography that need to be strictly eliminated, the information hidden in it that may be used by criminals cannot be ignored. If these potential risk information is directly used for AI training without careful screening, it is tantamount to providing an intelligent "accelerator" for criminal activities ".

Unfortunately, even in the highly specialized AI training process, it is difficult to completely avoid human negligence, resulting in some information beneficial to criminals being inadvertently incorporated into the AI learning category. In this case, AI may not only fail to become the guardian of society, but become an "accomplice" of criminal behavior ". For example, when criminals ask AI how to construct fraudulent emails, if AI provides detailed steps and techniques without screening, it undoubtedly invisibly facilitates the planning and implementation of criminal acts, and may even further upgrade criminal means by providing "optimization suggestions.

What's more serious is that if the pre training stage fails to effectively distinguish and eliminate such harmful information, the wide application of artificial intelligence may become a hidden danger of disturbing social order and endangering public safety. In recent years, there have been many cases to warn us of the real existence of these risks. For example, a ChatGPT in a community in Hangzhou was used to fabricate false news events, which not only misled the public, but also caused unnecessary panic and confusion. In addition, the "network water army" uses ChatGPT to produce and disseminate false and even terrorist information at low cost and high efficiency, which poses a serious threat to the network environment and social stability. Although the successful crackdown by the Shaoxing police demonstrates the zero tolerance attitude of law enforcement agencies towards cyber crime, it also highlights the urgency of strengthening the supervision of AI data and preventing its abuse.

3.6. users lazy psychology

In the pre-training phase of artificial intelligence, trainers often have the vision of making AI output more reliable and comprehensive answers, pouring large amounts of data into it, while the algorithm strives for perfection in its conclusions. Such a pursuit is certainly worthy of recognition, however, its applicability is not universal. When AI's answers generally tend to be stable and accurate and comprehensive, a potential risk quietly emerges: users may fall into a whirlpool of over-dependence and breed a lazy mentality.

In the daily interaction between users and AI chat robots, a challenge that cannot be ignored is that with the establishment and deepening of trust, users are increasingly inclined to transfer the initiative of thinking to AI, leading to the breeding of "thinking inertia. This inertia manifests itself not only in the blind acceptance of AI's answers, but also in the abandonment of the process of personal thinking and judgment in favor of relying on mechanical conclusions that lack human emotion and insight and are purely algorithm-driven.

What is more serious is that when the AI model lacks the necessary warning mechanism when providing information, it can easily answer uncertain or wrong questions without screening. This lack of transparency and the weakening of sense of responsibility will undoubtedly erode the cornerstone of users' trust in AI. In the long run, the information received by users may become one-sided and fragmented, and it is difficult to form a complete and accurate cognitive framework. This is not only about the risk of degradation of personal thinking ability, but also a profound contest about how to keep a clear head and stick to independent thinking ability in the wave of AI.

3.7. social division

The generation of harmful information risk is also users and developers need to pay attention to the problem, when ChatGPT generate content may contain information that violates ethical standards or laws, which may not only harm the rights and interests of users, but also may cause social order and public interests. Negative impact. Such risks can have a profound negative impact on social fragmentation, undermine public trust, affect overall cohesion, exacerbate social divisions, make understanding and communication between different groups more difficult, and even promote individual extremist ideas and weaken social stability. This has also led to an increase in the cost of social governance, affecting economic stability and social development.

For education, this is to reduce the value of education, education on social harmony contribution to the potential harm. As students gradually rely on the AI dialogue model, the traditional teacher-student interaction decreases, representing the decline in the utilization rate of the most basic direct interaction mode of human beings, which not only alienates the emotional connection between teachers and students, but also weakens the inheritance of values.

4. Summary

In today's information and digital age, the development of generative artificial intelligence is even more powerful, and the continuous update of algorithms and the continuous expansion of data provide a strong impetus for its progress. This not only shows the great potential of social benefits behind science and technology, but also becomes a hot topic in the field of science and technology. People are full of expectations for the application prospects of artificial intelligence technology in all walks of life. This is not only because of the performance of its revolutionary technology, but also because all sectors of society are full of longing for the changes brought about by AI technology.

This paper analyzes the generation mechanism of artificial intelligence in depth, aiming to uncover the mystery of AI and disenchantment of technology. Through in-depth analysis and practical exploration, especially in the pre-training stage, we found that there are many compliance risks in artificial intelligence. Many AI development companies act recklessly in this unregulated field. At the same time, due to the rapid development of artificial intelligence and the inherent lag of the law, the actual use of artificial intelligence has also caused a series of problems, such as the generation of harmful information, academic fraud and so on.

This article is intended to disenchant through technology, so that more people understand the generation mechanism of artificial intelligence, for those who are interested in this field of practitioners to provide a clear technical understanding, and remind artificial intelligence trainers in the pre-training stage must be more cautious, strict control of data collection sources, to prevent data leakage caused by data compliance issues. At the same time, we should pay close attention to domestic and foreign legislative trends. With the development of artificial intelligence, the law will be supplemented and improved in terms of copyright and data leakage. In addition, this paper also points out the current and possible future data compliance issues for legislators, and provides a reference for them to formulate relevant laws.

Through the in-depth analysis of this article, we hope to provide valuable insights into the future development of artificial intelligence, promote the healthy and compliant development of technology, and bring positive impact on society.

References

[1] Bi, W. (2023). Generative Artificial Intelligence Risk Regulation Dilemma and Its Solution: From the Perspective of ChatGPT. Comparative Study, 2023(03), 155-172.

[2] Cai, S., & Yang, L. (2023). Research on ChatGPT Risk and Collaborative Governance of Intelligent Robot Application. Theory and Practice of Information, 46(05), 14-22.

[3] Zhao, C. Y., Zhu, G. B., & Wang, J. Q. (2023). ChatGPT's inspiration to linguistic macromodeling and new development ideas of multimodal macromodeling. Data Analysis and Knowledge Discovery, 1-12. Retrieved March 31, 2023, from http://kns.cnki.net/kcms/detail/10.1478.G2.20230320.1508.004.html.

[4] Chen, R., & Jiang, Y. (2024). Research on governance of generative AI: A case study of ChatGPT. Science of Science Research, 24(1).

[5] Chen, J., & He, L. (2023). Generative Artificial Intelligence Risk Governance Urgently Needs "Ethics-Law" Integrated Framework. Science and Technology Review of Zhangjiang, 2023(02), 8-10.

[6] Chen, H., & Shifeng. (2024). Research on risk and governance of training data for generative artificial intelligence. Journal of Hainan Open University, 1.

[7] Perel, M., & Elkin-Koren, N. (2016). Accountability in Algorithmic Copyright Enforcement. Stanford Technology Law Review, 19, 473-533.

[8] Zhang, T. (2020). Legal Regulation of Algorithmic Bias in Automation Systems. Journal of Dalian University of Technology (Social Science Edition), 4.