International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 8 Issue 3, May-June 2024 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

The Data User of the Data Training Legally Grabs the Data and Puts it Into Training

Xu Xinqi

Beijing Materials Institute, Beijing, China

ABSTRACT

With the continuous innovation of artificial intelligence, a variety of new artificial intelligence, a variety of forms, various functions, various roles of artificial intelligence has penetrated into every aspect of our lives, but with the development of artificial intelligence, many legal pitfalls also surfaced, such as to upgrade in the development of artificial intelligence at the same time, our relevant laws have not followed the update. In many cases at home and abroad, the judge in the adjudication process is still invoking other laws to adjudicate cases of artificial intelligence to capture data infringement. On the one hand, there is no accurate relevant legislation, on the other hand, many people who have been captured data do not know that they have been infringed upon their legitimate rights and interests, so it is crucial for everyone to have a deeper understanding of AI data capture and know the legal pitfalls involved in order to protect their legitimate rights and interests.

KEYWORDS: *generative artificial intelligence, data crawling, crawlers, datasets, sensor acquisition, personality rights approach, data security approach*

How to cite this paper: Xu Xinqi "The Data User of the Data Training Legally Grabs the Data and Puts it Into Training" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-8 | Issue-3, June 2024, pp.768-771, URL: www.ijtsrd.com/papers/ijtsrd64886.pdf

IJTSRD64886

➢ What is generative AI.

Generative artificial intelligence refers to a new type of artificial intelligence that generates new and original content by learning from large-scale data sets, and it is a technology based on algorithms, models, and rules for generating content such as text, images, sound, video, and code. Provide generative artificial intelligence products or services should comply with the requirements of laws and regulations, respect for social morality, public order and morality. The more mature generative artificial intelligence on the market now has the following types.

➢ ChatGpt:

The natural language processing tool driven by artificial intelligence technology, which uses the Transformer neural network architecture, has the ability of language understanding and text generation. In particular, it will train the model by connecting a large number of corpora, which contain dialogues in the real world, so that ChatGPT can understand the sky and the earth, and can interact according to the context of chat, To communicate with real human beings in almost the same chat scene. ChatGPT is not only a chat robot, but also can perform tasks such as writing emails, video scripts, copywriting, translation, and code.

➢ Med-Gemini:

Med Gemini is a family of multimodal medical models based on Google's powerful Gemini model. It integrates advanced reasoning, multimodal understanding and long text processing capabilities. Through self training and network search integration, Med Gemini can more accurately diagnose and reason, and has achieved the best performance in benchmark tests such as MedQA (USMLE). Through fine tuning and customizing the coder, Med Gemini can better understand and process a variety of medical data modes, such as text, image For video and biological signals, Med Gemini can effectively analyze and understand long medical information, such as electronic health records (EHR) and medical teaching videos, and has achieved the best performance in relevant benchmark tests.

➢ AlphaFold 3:

In 2018, Google DeepMind launched its first protein structure prediction model, AlphaFold, and won the first place in the international protein structure

prediction contest. In 2020, DeepMind released the second version of AlphaFold software. AlphaFold 2 integrated a sub network system into a single differentiable model and applied Transformer to predict complex 3D structures based on amino acid sequences. In CASP14 in 2020, AlphaFold 2 stands out. Its prediction is accurate to atomic precision. Even for proteins lacking templates, it can produce excellent results in a few minutes. Today, millions of researchers around the world have applied AlphaFold 2 in malaria vaccine, cancer treatment, enzyme design and other fields. AlphaFold has been cited more than 20000 times, and its scientific influence has been recognized by many awards, the most recent of which is the Breakthrough Prize in Life Sciences Awarded. Later, the AlphaFold Multimedia promoted the prediction of protein protein complexes.

- **Sora:**

Sora, an artificial intelligence video model released by American artificial intelligence research company OpenAI, can create a lifelike video of up to 60 seconds according to the user's text prompts. The model understands the existence of these objects in the physical world, can deeply simulate the real physical world, and can generate complex scenes with multiple roles and specific movements. Inheriting the image quality and ability to follow instructions of DALL-E 3, Sora can understand the requirements put forward by users in the prompt. Sora brings infinite possibilities to artists, film producers or students who need to produce videos. It is one of the steps of OpenAI's plan of "teaching AI to understand and simulate the physical world in motion", It also marks a leap in AI's ability to understand and interact with real world scenes.

How AI learns to train

1. What is Artificial Intelligence Learning Training.

Artificial Intelligence training, also known as the learning process, refers to the training of a complex neural network model with large amounts of data to determine the values of the weights and biases in the network so that it can be adapted to a specific function. During training, the neural network weights need to be adjusted to minimize the loss function, and the training is performed by back propagation to update the weights in each layer. The training process requires high computational performance, a large amount of data, and some generalization of the trained network. This process can be likened to the process of learning on our own and using what we have learned to make judgments.

2. Artificial Intelligence Learning and Training Classification.

A. Supervised learning.

In supervised learning, the data has been labeled, which means you know the target variable. Using this learning method, the system can predict future outcomes based on past data. It requires at least input and output variables to be provided to the model in order to train it. The

B. Unsupervised Learning.

Unsupervised learning algorithms use unlabeled data to discover patterns from the data on their own. The system is able to recognize hidden features from the input data provided. Once the data is more readable, patterns and similarities become more apparent.

C. Intensive Learning.

The goal of reinforcement learning is to train an intelligent agent to accomplish tasks in an uncertain environment. The agent receives observations and rewards from the environment and sends actions to the environment. The reward measures how successful the action is in accomplishing the task goal.

D. Deep Learning.

Deep learning is a subset of machine learning that deals with algorithms inspired by the structure and function of the human brain. Deep learning algorithms can handle large amounts of structured and unstructured data. The core concept of deep learning lies in artificial neural networks, which enable machines to make decisions. Most of what we now call AI learning training refers to deep learning.

3. Legal Risks in AI Training.

At present, with the continuous innovation in the field of generative artificial intelligence, the continuous development and progress of science and technology have gradually exposed the contradictions between artificial intelligence and human beings, and at present, the most discussed aspects are also mainly focused on the copyright issues of generative artificial intelligence. From the perspective of the whole industrial cycle of generative AI, the copyright issue in the model training stage is at the beginning of the process, and thus has received extensive attention from all walks of life.

- **Case in point.**

According to incomplete statistics, as of May 2024, there are 19 actual litigation cases in the field of big model in the United States, 14 of which are copyright infringement cases. The core dispute is the unauthorized use of others' works for model training. In August 2023, the US company X said that Bright Data flagrantly violated the service agreement of the

platform, circumvented the risk control of the platform through technical means, and illegally captured the reply, like, forwarding and other data on X in batches. It believed that these illegal acts had a serious impact on X's servers and also damaged the user experience, so it asked for injunctive relief to prevent Bright Data from doing so. However, the court rejected X's request and gave reasons: social networks do not actually own user data, because the platform cannot enjoy the benefits of the safe haven principle on the one hand, and on the other hand, it emphasizes that data belongs to itself. This is tantamount to whether the legal principle of social platforms' sovereignty over user data has been established. Since X does not own data itself, but provides public data to users by other means, Bright Data's behavior of capturing public data is not illegal. From this case, we can see that there is no strict legislation and supervision standard for data crawling in the world. I will list several ways of data crawling and analyze the legal hazards involved in various ways.

1. Ways of acquiring data for AI training.

A. Crawler crawling.

Get data from the Internet using a web crawler. It is not illegal for a crawler to crawl data itself, but certain laws and ethics must be observed when using crawler technology to obtain data, so as to avoid infringing on the legitimate rights and interests of others or violating legal provisions. In January 2024, The UK Information Commissioner's Office (hereinafter referred to as the ICO) announced the launch of a series of research on generative AI. The first research has responded to the question whether there is a legal basis for the generative AI model trained by capturing data from the network. ICO pointed out that the legitimate interests of generative AI, which is trained by publicly capturing data, can become its legal basis, but the premise is that model developers should pass three tests, namely: purpose test, necessity test and balance test.[1] According to the explanation of ICO, we can see that there are three reasons for AI to capture network data reasonably: (1) data capture requires effective benefits; (2) Network data capture is necessary for AI training; (3) The capture process shall not infringe upon personal rights and interests.

B. Sensor Acquisition.

Use various sensor devices (such as cameras, microphones, temperature sensors, etc.) to collect real-time data. With regard to sensor data collection, there is a case of illegal sound data collection in practice. On April 23, the first trial of the Beijing Internet Court announced the country's first case of infringement of the personality right of AI voice generation. The plaintiff, Yin Mou, is a voice dubber. According to a friend, the plaintiff found that his voice dubbing works were widely circulated in many well-known APPs. After sound screening and tracing, it is found that the voice in the above works comes from the text to speech product in the platform operated by an intelligent technology company in Beijing, the defendant. The user can realize the function of text to speech by inputting text and adjusting parameters. In the end, the Beijing Internet Court held that voice, as a kind of personal rights and interests, has personal exclusivity, and the voice of any natural person should be protected by law. The authorization of sound recordings does not mean the authorization of voice AIzation. Without the permission of the obligee, unauthorized use or permission of others to use the voice in the sound recordings constitutes infringement.[2]According to this case, we can learn that it is possible to use sensors to collect environmental data such as weather temperature, air humidity, air quality and other natural data, but if you want to take private data such as sound, you have to be careful not to infringe on the right to personality, and you have to seek the consent of the person being collected.

C. Manual labeling.

Processing data through manual labeling. For example, to build a speech recognition model, a large amount of recorded data needs to be labeled so that machine learning algorithms can learn how to recognize sounds.

D. Dataset Purchase.

Some data set providers can provide data sets in specific fields, such as medical, financial or social media data. In the early days of AI development, technology giants such as Google, Meta and OpenAI used a large amount of free data from the Internet to train generative AI models. The purchase and use of these data sets is legal and reasonable, but as time goes by, if you want to make more progress in generative AI, these free data sets can no longer be satisfied, and major manufacturers began to purchase use data in gray areas, which led to many cases of infringement of collected AI data, but data crawling and selling are recognized as gray area transactions, At present, there is no legislation or case in the world that can clearly point out which data to buy is illegal and which data to buy is legal.

E. Crowdsourcing.

Crowdsourcing platforms are used to hire people to perform specific tasks, such as labeling images, translating text, or categorizing data. When collecting AI data, there is a need to ensure the quality and

accuracy of the data. The data must accurately reflect reality and comply with privacy and security regulations. In addition, the diversity and quantity of data need to be considered to ensure that the AI models trained have broad applicability and high-quality predictions.

- **Conclusion.**

At present, there are no clear laws and regulations at home and abroad on the behavior of AI technology giants to capture data and train AI. To determine whether they are illegal, they can only use the Personal Rights Law, the Data Security Law and other laws and regulations as their standards. Personally, I believe that China should follow the legislation of the EU, Japan and other countries to adopt exemption rules for AI data capture training: that is, without harming personal interests, if they delete the illegally captured data in time, they will not be sanctioned. This can not only promote scientific and technological innovation in the field of artificial intelligence, but also protect the legitimate rights and interests of the authors. However, it must be noted that if personal interests are not harmed, severe punishment will be imposed if personal interests of others are harmed. Never appease.

References

[1] Is network capture data used for generative AI training? UK ICO: Legal interests are the key to feasibility. 21st Century Report April 25, 2024

[2] As a personal right, sound has personal exclusivity. The voice of any natural person should be protected by law. The authorization of sound recordings does not mean the authorization of sound AIzation. Without the permission of the obligee, unauthorized use or permission of others to use the voice in sound recordings constitutes infringement. Official account South China Sea United Front 2024.5.114