Anonymization of Personal Information in Legal Context

He Zhuolyu[1] and jiagnfeng hao[1]

[1]Macau University of Science and Technology

September 24, 2022

Abstract

In order to balance the relationship between the protection of personal information and the use of personal information, the Personal Information Protection Law has set anonymization standards. Although, technically, absolute anonymity is not achievable, but the legal goals are not the same as the technical goals. In order to maximize social welfare, it is necessary to reduce the quality of anonymization to a certain extent to ensure the universality and speed of the personal information governance program. Therefore, it becomes feasible to realize the anonymization of personal information in law. In order to facilitate the anonymization of judgments, different jurisdictions have established different rational people standards under the guidance of value. The EU standard overemphasizes the protection of the rights of information subjects, which is not conducive to the use of personal information, and it is difficult to maximize social benefits. The Anglo-American standard takes into account both "security" and "efficiency", while protecting the information subject, while leaving enough space for personal information processors. Among them, the United States compromises the adoption of abstract standards and specific standards, which overcomes the impracticality of adopting abstract standards alone. The shortcomings of non-predictability and other specific standards.

Anonymization of Personal Information in Legal Context

Author information:

Dr.He Zhuolyu, phd student of macau university of science and technology,law faculty.

Dr.Hao jiangfeng, phd student of macau university of science and technology,law faculty.(corresponding author)

Authors have no conflict of interest relevant to this article.

All listed authors have contributed to the manuscript substantially and have agreed to the final submitted version. The list of authors accurately illustrate who contributed to the work and how.

Title:Anonymization of Personal Information in Legal Context

Running head:"Anonymization of Personal Information in Legal Context" would be fine.

Abstract:In order to balance the relationship between the protection of personal information and the use of personal information, the Personal Information Protection Law has set anonymization standards. Although, technically, absolute anonymity is not achievable, but the legal goals are not the same as the technical goals. In order to maximize social welfare, it is necessary to reduce the quality of anonymization to a certain extent to ensure the universality and speed of the personal information governance program. Therefore, it becomes feasible to realize the anonymization of personal information in law. In order to facilitate the anonymization of judgments, different jurisdictions have established different rational people standards under the guidance of value. The EU standard overemphasizes the protection of the rights of information subjects, which is

not conducive to the use of personal information, and it is difficult to maximize social benefits. The Anglo-American standard takes into account both ”security” and ”efficiency”, while protecting the information subject, while leaving enough space for personal information processors. Among them, the United States compromises the adoption of abstract standards and specific standards, which overcomes the impracticality of adopting abstract standards alone. The shortcomings of non-predictability and other specific standards.

Keywords: Personal Information; Anonymization; Value Goals; Rational People

introduction

We are being watched all the time. In the past, the collection of a large amount of personal information was extremely difficult due to objective resource and capacity constraints, and lack of computing power and algorithm support made it difficult to exert its commercial value even if personal information was collected. Too much energy is put into personal information protection. With the popularization of smart devices and the development of digital technology, the collection, analysis and application of personal information have become very easy. For example, when a search engine is opened to search for a certain entry or sentence, the search engine will immediately identify the current needs of the searcher and push corresponding commercial advertisements; when using a food delivery platform APP, the platform will collect user points in the past Meal data, identify preferences, and accurately display dishes that meet the user’s taste on the user’s homepage. In fact, the current advanced data technology and mature data industry have brought great convenience to people’s lives. We no longer have to search for a certain commodity, and we don’t have to worry about what to eat tomorrow, and we can even buy a bus ticket. At the same time, actively push the destination hotels that meet their various needs and can afford the price. Big data seems to know itself better than we do. Gradually, people felt threatened. When we were chatting with our family members on a daily basis, we suddenly found that the mobile phone pushed a marketing advertisement of the product that we had not searched but just talked about; when we swiped a short video, no matter how the settings were set, we could accurately push ”people you may know”. It turns out that cameras record our every move, cell phones monitor our every word, and all of our domestic life seems to be exposed to others; Our desire to explore, develop and change.11See Manon Oostveen, translated by Cao Bo, ”Data Boundaries-Privacy and Personal Data Protection”, Shanghai People’s Publishing House, August 2020, 1st edition, p. 51.Therefore, we are reluctant to talk to people too much, and dare not speak on the short video platform anymore, and our personal behavior is restricted. More than that, our network traces determine the content we are recommended to see, which causes us to be influenced by certain types of information and make our minds live in the cocoon; past consumption records make the price of goods closer to consumption The highest price that the operator is willing to pay, so that the operator can obtain more consumer surplus and realize price discrimination. At this time, the demand for personal information protection by information subjects has become extremely strong. Although the formulation of laws always lags behind social progress, it is necessary to respond to public opinion. In 2018, the entry into force of the EU’s General Data Protection Regulation (GDPR) has attracted widespread attention from all walks of life, and research on personal information protection has increased significantly in academia, which has accelerated the birth of special laws on personal information protection in my country. On November 1, 2021, my country’s ”Personal Information Protection Law” came into effect, but research on legal issues should not stop with the formulation of laws. Article 4 of the Law stipulates that personal information refers to various information related to identified or identifiable natural persons recorded electronically or in other ways, excluding anonymized information. Clearly regard ”identified”, ”identifiable” and ”anonymized” as the boundaries of personal information protection, in order to protect the rights and interests of personal information, promote the rational use of personal information, and balance the interests of information subjects and personal information processors. However, ”In a limited and specific sense, people’s cognitive ability is subject to many limitations.”22Zhang Benxiang, Yan Zexian. Some limitations of human cognition from the perspective of complexity [J]. Science and Technology and Dialectics, 2006(02):20 .Academic circles are full of debates on the criteria for “identifiability” and “anonymity” and their rationality.

1. The boundary of personal information and the problem of anonymization

(1) Scope of protection of personal information

Information is diverse. In essence, it is a signal and information that reflects the state and laws of movement, development, and change in the real world.33See Qi Aimin: "Defending Property in the Information Society", Peking University Press, 2009, p. 48.When researchers write their papers, the literature search records are information, the takeaway records left to satisfy their hunger are information, and the content displayed on the takeaway platform is also information, and the extension of information is extremely broad. Legally, classifying them is a double test of language and legislative techniques.

As the regulated object of the Personal Information Protection Law, personal information must be distinguished from other non-personal information; looking at the regulations of various countries and regions, the scope of protection of personal information is mostly defined in the form of simple definitions or definitions and enumerations. GDPR is one of the most influential norms in personal information protection law. It defines personal information as "any information related to an identified or identifiable natural person." , to guide the public, GDPR divides "identifiable natural persons" into direct identification and indirect identification, and lists "any information", such as name, ID number, location, body, physiology, genetics, economy, culture, society identity, etc.44EUROPEAN GENERAL DATA PROTECTION REGULATION. Article 4.There are also the United States "Children's Online Privacy Protection Act (COPPA)", China's Macao "Personal Data Protection Law" and so on. However, my country adopts a simple definition method, which defines personal information as various information related to an identified or identifiable natural person recorded electronically or by other means. This definition method is the practice adopted by most countries or regions in the world, such as the British Data Protection Act, the French "Information, Archives and Freedom Act", the German "Federal Data Protection Act", the Singapore "Personal Data Protection Act", etc.55See Qi Aimin and Zhang Zhe: Identification and Re-Identification: Concept Definition and Legislative Choice of Personal Information, Journal of Chongqing University (Social Science Edition), 2018, No. 2, p. 121.Compared with the two definition schemes, when defining personal information, both "identified" and "identifiable" are used as the extension of personal information, and the difference lies in whether "relevant information" is listed. However, this distinction is not absolute. For example, although my country has not explained "relevant information" in the Personal Information Protection Law, it may lead to relevant personal information in normative documents such as the Regulations on the Protection of Personal Information of Telecommunications and Internet Users. The enumeration provisions are provided for reference. The two definitions are gradually converging.

In addition to the vague concept of personal information itself, "identifiable" is also difficult to define. The so-called identification is the process of identifying an individual through various characteristics related to the individual. One morning, in a surveillance-installed restaurant, a woman had breakfast with her friends. Based on this information, ordinary people cannot identify the woman's identity and the menu. The restaurant owner can only know the menu content, and the friend who had breakfast with the woman can clearly know the woman's identity and what dishes were ordered. On the one hand, almost every kind of personal information corresponds to a certain characteristic of an individual, and as long as it is a characteristic, it is possible to identify an individual's identity; Therefore, establishing an absolute "recognizable" standard is difficult. The EU Article 29 Working Group equates "identifiability" with the possibility of "singling out" a person,66cf. Article 29 Working Group, Opinion No. 2007/4 on Personal Data (2007) WP13613-14.and divide personal information into "directly identifiable information", "indirectly identifiable information" and "anonymized information". Among them, "directly identifiable information" means that an individual can be directly identified through a certain information, and "indirectly identifiable information" means that a certain information needs to be combined with other information to single out a person, and "anonymized information" will be The information of personal information is de-identified to the extent that it is impossible to identify the individual or it becomes extremely difficult to identify the individual. However, as mentioned in the previous example, for ordinary people, even if they are provided with information such as "female", "a certain morning", "breakfast", etc., it is difficult to identify the woman's identity, so the information is "anonymized information" ; For restaurant owners, because they have mastered menus (handwriting, dining preferences, etc.) and video surveillance (body shape, facial contours, etc.) compared to ordinary people, they can completely rely on other information (face recognition at police stations, street visits, etc.)) "singling out" a specific individual, so the information it holds is "indirectly identifiable information"; and for the female friend, because she directly knows the name of the woman and can immediately complete the identification, it constitutes "directly identifiable information" ". When "identified" or "identifiable" personal information is exposed to the public or placed in a place accessible to others, the information subject has a reasonable expectation that the personal information will be protected. The disclosure of such personal information may be suspected of offending the privacy of the information subject, infringing on the "private domain", affecting the interaction between the information subject and the external environment, and thus affecting the personality development of the information subject.77See Xie Yuanyang: "The Value of Personal Information from the Perspective of Information Theory: A Review of the Privacy Protection Model", Tsinghua Law, No. 3, 2015, pp. 99-103.Today, "data" has become the fifth largest factor of production in my country after land, labor, capital, and technology. As the basis of big data, personal information, and its effective use is one of the keys to realizing the marketization of data elements. The Personal Information Protection Law explicitly excludes "anonymized information" from the scope of protection.

(2) The problem of anonymization of personal information

In general, there are at least three difficulties in the anonymization of personal information: science and technology, legislative technology and balance of interests. The advancement of science and technology can provide more comprehensive, convenient and practical technology for anonymization, but at the same time, re-identification technology is also developing, and the two are contradictory, making it difficult to realize the anonymization of personal information in science and technology. The effective distinction between personal information and anonymized information in legislation can give the relevant public expectations, but this distinction is often difficult to achieve. This ambiguity is the most challenging and complex area of personal information judgment. The information subject and the personal information processor have different interests and needs for personal information. Complete or excessive anonymization makes the personal information of the information subject well protected, but for the personal information processor, this anonymized data is almost useless; On the contrary, if the degree of anonymity is too low and the personal information of the information subject is easily infringed, it is really complicated where the interests of the two parties should be balanced.

1. Science and technology are difficult to be anonymous

"In the era of big data, through data fusion and cross-validation, the information subject can be identified through multiple pieces of information that cannot identify the information subject."88Liu Yingshuang: "Rethinking the Protection of Personal Information in the Era of Big Data", Social Science, No. 3, 2019, p. 103.In 2010, a study by Paul Ohm showed that existing anonymization technologies could not live up to the expectations of anonymity, and anonymity had become a broken promise. Therefore, he advocated that the anonymization of personal information is technically impossible and the concept of "anonymity" should be abandoned.99See PAUL O.Broken promises of privacy:responding to the surprising failure of anonymization [J].UCLA Law Re-view, 2010, 57(6):1701-1777.In 2019, Imperial College London and KU Leuven Using just 15 demographic characteristics and some machine learning, the team of researchers was able to get 99.98% of Americans to be correctly re-identified in any dataset. The researchers say that even heavily sampled anonymized datasets are unlikely to meet modern anonymization standards mandated by the GDPR and seriously challenge the technical and legal adequacy of de-identification "release and forget" models.1010See Rocher L , Hendrickx J M , Montjoye Y . Estimating the success of re-identifications in incomplete datasets using generative models [J]. Nature Communications.Over the past 10 years, views on technical anonymization have remained relatively negative. In other words, the technology does not meet the legal requirements. Moreover, capital will gather in places of interest. For personal information processors, if they have to pay a huge cost to realize information anonymization, it is not in line with their goals of commercial interests, and it will also hinder the personal information processors. The utilization of, therefore, there is no driving force for anonymization technology updates in the industry. If the reverse analysis and re-identification technology can bring more benefits to personal information processors, they are willing to invest capital to pursue greater benefits. To take a step back, even if there is an absolute personal information anonymization technology objectively, it can only be realized when the benefits brought by anonymization to the personal information processor are greater than the use of personal information.

2. Difficulties in Legislative Technology

Personal information is an intangible object that is not as real and tangible as real objects. Its intangible characteristics make its protection scope impossible to be naturally delimited like tangible property rights. Its ”shape”, ”dimension”, ”boundary” and ”size” need to be constructed artificially to determine the connotation and extension of the corresponding rights, so as to make Personal information is properly protected. at the same time,

Humans’ understanding of objective things is always insufficient, especially with the development of science and technology and the complication of interpersonal network relationships, it is a test of human wisdom to identify the essence, connotation and extension of emerging things, and the protection of personal information is like this. .

”Identification” and ”anonymization” are two aspects of personal information protection. To make ”anonymization” meaningful in personal information protection laws, the scope of ”identifiable” must not violate the field of anonymization. Take whether recursive identification is allowed as an example. Suppose that for the same information subject, she wears a red fleece jacket, a green short skirt, and has a non-mainstream hairstyle. A often sees her in Chengdu, and B has attended a meeting with her and knows. She majored in law in college. C saw her hurriedly entered a people’s court one morning. On the same day, Ding was knocked down by her bicycle. She told Ding that she was in a hurry to go to the back of the court to contact her and then left. All I know is that the accident car is a red Toyota Camry with a cartoon pattern of Monkey King and the time of the accident; the previous and next streets of the incident section are under video surveillance, and no one knows her name and other information. According to the information held by A, B, C, and D respectively, none of the parties can actually identify the identity of ”her”, but if A, B, C, and D are all employees of a company, when chatting at noon, because ”she” Her appearance is more memorable, and all parties have talked about her, so it can be inferred that this woman, most likely from Chengdu, a legal major, may be a lawyer, who owns a red Toyota car and knocked Ding down one day After leaving the accident scene, and then going to a people’s court for a court session, because the appearance of the vehicle is rather special, assuming that Ding later locked the accident license plate number based on the accident time and the monitoring around the accident location, he completed the identification of ”her” personal information. identify. In the whole process, the information that D finally masters is identifiable, then, is the information mastered by A, B, and C recognizable? In fact, as far as a single person is concerned, the identification cannot be completed according to the information he has, and he has no motivation to seek more information. It can be considered that the information held by a single person has no possibility of identification. But the objective result is that Ding completed the final identification because of the aggregation of the basic information mastered by all parties. It seems that the information mastered by a single person has the possibility of identification. If recursion of identifiability is allowed, all the information used in its identification process will be included in the list of personal information; on the contrary, if it is not allowed, only the information that finally identifies a specific natural person will be recognized as personal information.1111See Yang Nan: ”Reflection and Restriction on the Expansion of Personal Information ”Identifiability”, Journal of Dalian University of Technology (Social Science Edition), 2021, No. 2, p. 101.At this time, phrases like ”personal information protection laws do not regulate anonymized information” will become meaningless because there are no applicable scenarios.

Moreover, as mentioned above, ”identifiable” judgment is particularly dependent on the judgment scene, and ”anonymization”, as a relative concept, also has similar characteristics. A person’s voice is indistinguishable to a stranger if it is not special enough, but for a good friend, the identification of the information subject can be completed only by opening his mouth, and even if the voice is distorted, a good friend can Tone and idioms can easily correspond to specific individuals; at this time, for strangers, voices belong to anonymous information and the subject of unidentifiable information, while for good friends, voices are identifiable personal information. Objectively, the scenarios are rich and diverse, and it is impossible for the law to make adaptive distinctions for various scenarios. Moreover, if each case needs to be judged by strictly considering various factors, then its results are unpredictable and legal. Stability and predictability will be difficult to achieve. Therefore, it must be generalized in order to guide behavior, but there will be a problem that individual justice is difficult to achieve.

3. It is difficult to balance the interests of both parties

Under the Internet, on the one hand, while everyone enjoys the convenience brought by technological services, personal information is also processed all the time, just like taking off one's clothes. ”Algorithmic black box” and ”forced consent” have made information subjects more and more concerned about the illegal processing of personal information. Information subjects need a law to fully, comprehensively and effectively protect their personal information rights and interests. On the other hand, personal information processors enjoy the dividends brought by the collection, analysis, and application of personal information. They can't wait to know themselves better than the information subject in order to maximize their own interests, so they ignore the objection of the information subject and use illegal arrests. Handle personal information as much as possible by taking it, stopping the service without ”consent”, inducing deception, etc. Therefore, in order to balance the interests of both parties, the Personal Information Protection Law defines the scope of personal information as “identified” and “identifiable”, and demarcates the use of “unrecoverable” or “unrecoverable” for anonymized information from personal information. . Even so, as mentioned earlier, the location of this dividing line is difficult to delineate. Usually, the formulation of laws should not only respond to the needs of the public, but also consider the practicality and feasibility of legislation. The law on the balance of interests must be based on the current situation and adapt to the social and economic development of the country. If these factors cannot be properly balanced to properly balance the conflict of interests, in addition to not being able to guide the behavior of both parties, it will easily lead to new disputes.

2. An investigation of anonymity judgment criteria from the perspective of comparative law

(1) Anonymization of personal information in Europe and the United States

a. European Union

The preamble to the GDPR states that data protection principles do not apply to anonymized information; anonymized information is information that is not related to an identified or identifiable natural person, or that is not related to anonymized information that is presented in a way that is not or no longer identifiable to the subject of the information. information. Specifically, in the process of anonymizing personal information, the direct identifier and indirect identifier in the personal information must be removed at the same time, so that they cannot be re-identified, so that the information subject can lose control over the personal information. The GDPR also provides for “pseudonymization,” which refers to the processing of personal information in a way that no longer attributes personal information to a particular information subject without the use of additional information. This additional information should be kept separately, and technical and administrative measures should be used to ensure that the corresponding personal information is not obtained by natural persons who have been identified or who may be identified.1212See EUROPEAN GENERAL DATA PROTECTION REGULATION. Article 4.It can be seen that pseudonymization information merely removes the direct identifier. Judging from the risk of re-identification, the re-identification risk of pseudonymization is higher than that of anonymized information. In particular, the direct identifier and the indirect identifier are a relative concept, and a certain feature information is not always in the same type, and it will be different according to the different scenarios when judging. Scenario 1: In most cases, the name refers to a specific information subject, but if the name is a common name, it is impossible to identify many subjects. For example, use ”Zhou Jie” in the ”Sichuan Public Security Government Affairs Service Network” to query the same name , it shows that there are 2773 people with the same name in Sichuan Province, at this time ”Zhou Jie” is an indirect identifier; but if you search for ”Zhou Jiebai”, it shows that there is only one person in Sichuan Province, and ”Zhou Jiebai” is a direct identifier. Scenario

2: Even if the name is very common, as long as the regional base is small enough, it can be identified. Taking "Zhou Jie" as an example, it is not searched within the scope of Sichuan Province, but only in a certain community or community, which can be almost accurate. At this time, "Zhou Jie" is a direct identifier. Scenario 3: Assuming that "Zhou Jie" is a well-known movie star, when talking about entertainment gossip or movies, if "Zhou Jie" is mentioned, people will immediately recognize it as a movie star "Zhou Jie". At this time, "Zhou Jie" is also direct identifier. In the scenario of "Zhou Jie" as an indirect identifier, if other information is added, it is easy to identify the information subject. Therefore, "pseudonymized information" that only removes indirect identifiers cannot separate the information from the information subject, and does not belong to "anonymized information" in the strict sense. The Article 29 Working Group pointed out in its Opinion No. 05/2014 that it is wrong to treat pseudonymization as a method of anonymization, and pseudonymization can only reduce the possibility of personal information being linked to the true identity of the information subject. sex, and therefore only one type of security measure.1313See Opinion 05/2014 on Anonymisation Techniques, p.18, https://ec.europa.eu/justice/article-29/docume ntation/opinion-recommendation/files/2014/wp216_en.pdf, April 2014.Because big data processing involves new technologies that pose high risks to individual rights and freedoms,1414see Manon Oostveen, translated by Cao Bo, "Data Boundaries-Privacy and Personal Data Protection", Shanghai People's Publishing House, August 2020 1st edition, p. 170.Before the anonymization of personal information, the personal information processor has an obligation to evaluate and prevent it in advance. Objectively, there are various anonymization techniques, such as adding noise, identity permutation, differential privacy, generalization, information aggregation and K-anonymity, L-diversity and T-similarity, etc. The Article 29 working group also elaborates Explains the principles, advantages and disadvantages of various anonymization techniques, as well as common mistakes when using each technique.1515See Opinion 05/2014 on Anonymisation Techniques, p.12, https://ec.europa.eu/justice/article-29/docume ntation/opinion-recommendation/files/2014/wp216_en.pdf, April 2014.Therefore , each anonymity technology has its own suitable application scenarios, taking into account factors such as the nature, scope, content and purpose of the processing, evaluate the different degrees of impact of processing on the rights and freedoms of natural persons, in an effective way Implement personal information protection. At the same time, this assessment obligation is continuous, and the EU requires that anonymous information not be "released and forgotten".1616See Qi Yingcheng: "Review and Alternative Choices of my country's Personal Information Anonymization Rules", Global Law Review, No. 3, 2021, p. 57.Human understanding is limited, and science and technology are constantly developing. The technical or/or organizational measures taken by personal information processors before acquiring, analyzing, and applying data may not meet the needs of protecting personal information in the future, and need to continue to be appropriate. The effectiveness of "anonymization" is evaluated to determine whether measures should be optimized.

b. America

Compared with the EU GDPR, there is no unified personal information protection law in the United States; however, for personal information issues, the United States has long been positioned to deal with it from the perspective of users by citing and adapting privacy protection. . "The Right to Privacy" published by Warren and Brandies in Harvard Law Review is the pioneering work of privacy research in the United States. Although the article was published more than 100 years ago, the arguments in it are still instructive to this day. The article argues that the latest technological inventions and business practices have evoked further protection of the legal interests of personality, as well as the right to ensure Justice Thomas Cooley's right to solitude. Innumerable mechanical devices have allowed "room whispers to be spread openly on rooftops", thus arguing that the law should recognize the right to privacy in all respects. On this basis, Professor William analyzed the cases in practice, and believed that the infringement of privacy rights involves four different interests: (1) infringing on others' solitude or private affairs; (2) publicly revealing disturbing private facts; (3) public misunderstanding by public disclosure; (4) use of another's name or characteristics for one's own benefit. Later, the fourth interest developed into "publicity rights".1717See Wang Zejian: The Law of Personality Rights, Peking University Press, 2013, 1st edition, pp. 182-184.After Whalen v. Roe (1977),1818See Whalen v. Roe, 429 U.S. 589 (1977).The protection of information privacy in the United States gradually developed.1919Long Weiqiu: "Research on the Construction of New Data Property Rights and Its System", in "Politics and Law Forum", No. 4, 2017.For example, the 1974 Privacy Act regulates the federal government's handling of personal information, the 1986 Electronic Communications Privacy Act prevents the government from accessing private electronic communications without permission, and the 1994 Driver Privacy Protection Act The Act regulates the privacy and disclosure of personal information collected by state motor vehicle departments, and the Children's Online Privacy Protection Act of 1998 focuses on the protection of personal information for children under the age of 13. At the state level, Illinois' Biometric Information Privacy Act and California's Consumer Privacy Protection Act have received much attention, the latter clarifying that personal information does not include de-identified information.2020See CCPA, §1798.140 (K) (3), http://leginfo.legislature.ca.gov/faces/codes_-displayText.xhtml?division=3.&part=4.&lawCo de=CIV&title=1.81.5, June 2018 .In 2021, Virginia became the second state after California to have a comprehensive data privacy law with the enactment of the Consumer Data Protection Act. As an agency directly under the U.S. Department of Commerce, the National Institute of Standards and Technology of the United States issued the "Guidelines for the Protection of Personally Identifiable Information" in 2010. According to the privacy rules of the guideline, only identifiable personal health information is included in the scope of protection, and vice versa. Health information that does not identify an individual or has no reasonable grounds to believe that an individual could be identified - i.e., anonymized information is excluded.2121See Elizabeth A. Brasher, "Addressing the Failure of Anonymization: Guidance from the European Union's General Data Protection Regulation", Columbia Business Law Review, vol.2018, no.1, 2018.Further, the guidance explains from a technical perspective To the so-called anonymization or de-identification is to remove personally identifiable identifiers from personal information; and it is stipulated that as long as the 18 types of identifiers listed in the guide are deleted, and the personal information processor does not know that the information is alone or combined with other information Personal information is considered anonymized when a specific natural person can be identified.2222See Liu Ying and Gu Jiaqi: "Personal Information De-Identification and Its Institutional Construction", Academic Research, No. 12, 2020, pp. 65-66.On the whole, the United States has relatively loose protection of personal information of information subjects, and has a more friendly attitude towards personal information processors, which facilitates the circulation and application of data.

(2) Comparative study on the judgment criteria of anonymity

The descriptions of "unrecognizable" and "unrecoverable" in the anonymization of personal information seem to point to a standard of objectification; but in some cases, it is considered impossible to achieve anonymization technically; even if it is possible, it will be Because personal information loses all characteristics and becomes non-commercially valuable. However, if "can't" and "can't" are regarded as subjective judgments, it seems to be too arbitrary, and "different from person to person" makes anonymization not legally stable. Therefore, if a set of practical judgment rules cannot be established to clarify the boundary between personal information and anonymized information, it is meaningless to talk about information utilization. my country's "Personal Information Protection Law" was born, and the "Network Security Law" and "Personal Information Security Specification" (GB/T 35273-2017) related to the anonymization of personal information also only adopted the same regulations as Article 73 of the "Personal Information Protection Law". Paragraph 4 of Article 4 is similar to the expression "the process of processing personal information that cannot identify a specific natural person and cannot be recovered". It lacks detailed provisions to guide judgment and lacks relevant practical data. Therefore, the following continues to examine relevant foreign standards. Looking at various countries, most of them use a combination of "subjective" and "objective" to solve this problem. The details are as follows:

a. EU: Active infringers

In the preamble of the GDPR, the standard of "active infringer" is established on whether the anonymized information satisfies the de-identification, which is used to judge whether the perpetrator has fulfilled due care obligations. When an attacker attempts to re-identify the data subject from the anonymized data, taking into account objective factors such as the cost at the time, the technical level and technological development at the time of information processing, if the personal information processor can associate When all reasonably possible means are made so that the combination of anonymized data and other data can no longer identify the subject of the information, such anonymization is considered to be compliant and not subject to the GDPR.2323See Zhang Chenyuan: ”Legal Regulation of Data Anonymization”, in ”Journal of Chongqing University of Posts and Telecommunications (Social Science Edition)”, No. 6, 2017, p. 54.means that the GDPR requires personal information processors to pre-judge all possible technical means that may be used by infringers, and extremely tests the personal information processors' cognitive level, legal judgment and technical selection ability when processing personal information; indeed , it is impossible for anyone to take into account all aspects of risks in detail, and cannot have a ”God's perspective” like ”after the fact”. The high requirements of this standard for personal information users to use anonymized information are very difficult to meet.

b. Great Britain: Intentional Intruders

”Anonymization: Information Protection Risk Management for Practice” provides guidance for the judgment of information anonymization standards - using ”intentional intruders” to examine the risk of re-identification of anonymized information.2424See STALLA-BOURDILLON S, KNIGHT A. Anonymous Da- ta V. Personal Data [J]. Wisconsin International Law Journal, 2016(2) : 284-322.Specifically, suppose there is such a willful infringer, who has More methods, knowledge, and skills than the average person, but not at the level of an expert. For example, resources can be obtained through open channels, and information can be collected using investigation methods that ordinary people can take, but they cannot use professional techniques like network hackers, use computing power or computing equipment beyond ordinary people's ability, or use criminals. method to obtain.2525See Zhang Chenyuan: ”Legal Regulation of Data Anonymization”, in ”Journal of Chongqing University of Posts and Telecommunications (Social Science Edition)”, No. 6, 2017, p. 55.Under this premise, if the willful infringer cannot identify the subject of the information based on the anonymized information, the test is passed. Compared with the ”active infringer” of the EU GDPR, the ”intentional intruder” standard is much more ”mild”. It makes the judgment standard objective, and it is not difficult for others. Personal information processors can better predict the legality of their own actions. estimate.

c. The United States: Safe Harbor and Expert Judgment Criteria

The so-called safe harbor means that after a personal information processor makes a legally required deletion action on personal information, if the processed information alone or in combination with other information can be used to identify a natural person, it meets the standard of anonymization of personal information. For example, Paragraph b of Article 164.514 of the ”Guidelines for the Protection of Personally Identifiable Information” stipulates that as long as the name, date of birth, phone number, social security number, license plate number, social security number, fingerprints and other biometric information, ID cards and other documents are deleted from personal information When the personal information processor does not subjectively know or should know that the information can be re-identified, the personal information is considered to have achieved the effect of anonymization.2626See Liu Ying and Gu Jiaqi: ”De-Identification of Personal Information and Its Institutional Construction”, Academic Research, No. 12, 2020, p. 66.The Health Insurance Portability and Accountability Act has similar provisions. Although the ”Children's Online Privacy Protection Act” does not explicitly delete specific identifiers to achieve the standard of anonymity, the types of identifiers in the Act are specified by enumeration, and it can be presumed that the specific identifiers in the Act do not exist in personal information. It should not be subject to the regulation of the Act, in fact, it also establishes a safe harbor for personal information processors.2727See Qi Yingcheng: ”Review and Alternatives to my country's Personal Information Anonymization Rules”, Global Law Review, No. 3, 2021, p. 58.In addition, the ”Health Insurance Portability and Accountability Law” also sets up expert judgment standards, pretending that there is a person with statistical expertise, and using his professional knowledge to re-identify the de-identified information as a judgment. Meet the information anonymization requirements. Of course, the expert here is an abstract person, and it needs to be based on the basis of ordinary people, that is, experts = ordinary people's cognition, ability + statistics; and the British ”deliberate intruders" have more than ordinary people in various professions. More knowledge, cognition and ability should be higher than the "expert" in the American standard. Combined with the safe harbor standard, the requirements for anonymization of personal information in the United States are significantly lower than those in the United Kingdom and the European Union, and the operability is stronger.

(3) Evaluation of existing anonymization standards

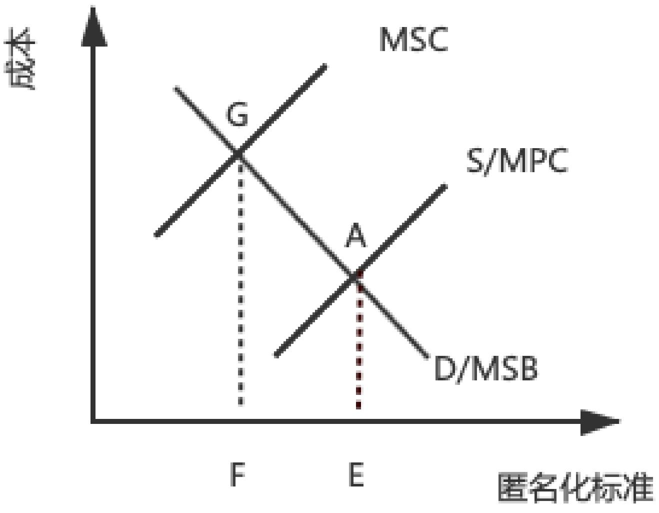
The EU's adoption of the active infringer standard has undoubtedly put forward an anonymization standard that is almost difficult to pass the test for personal information processors. any third party identification. Since then, it seems that personal information can be better protected, but the value of the information has also been greatly reduced. No longer process personal information; or, because of the huge benefits that personal information can bring, according to Marx, as long as the benefits are large enough, no matter how high the cost and the great risk of breaking the law, someone can take the risk. But no matter what the result is, it is not what the legislators want; unlike in the past, "cars, horses, and mail are slow", we are now in a digital environment, and the collection and processing of personal information is a necessary prerequisite for the provision of digital services. Information subjects have also developed serious path dependence on these digital services. Therefore, under EU standards, it is difficult for personal information processors to comply with them, so everyone has become a lawbreaker. In order to prevent the consequences of illegal harm from being discovered, personal information processors dare not to share information, let alone promote the flow of information. In the database, if the information is not leaked, it is not necessary to spend the effort to strictly anonymize personal information, not to mention that a large amount of personal information is enough to give itself a huge market advantage. Gradually, illegal activities are hidden. deeper.

In contrast, British and American standards give personal information processors leeway for legal processing. Specifically, from the standpoint of personal information processors, American standards are more relaxed than British standards. In addition to the relaxation, there are doubts about the risk of information re-identification. When anonymization is no longer absolute, whether it adopts the "intentional intruder standard" or the "safe haven and expert judgment standard", the information is likely to be re-identified, and the rights of the information subject will not be absolutely guaranteed, even if the United States is anonymizing The information circulation compliance requires personal information processors to prohibit the counterparty from re-identifying the data in the contract, but as long as the information is in circulation, it is possible to be identified, and the contractual binding capacity is still questioned. If other third parties conduct re-identification How to deal with it? The Personal Information Protection Law is a law of balance of interests. Some people are happy and others are sad. The requirement for anonymization of personal information processors has been lowered, and the information subject will say that it is not conducive to the protection of personal information. The anonymity standard is too high, and personal information processors complain again. It is difficult to be strong, and it is really nerve-racking which standard is suitable. It turns out that no matter what standard is proposed, it cannot satisfy either the information subject or the personal information processor. But is full satisfaction of both parties a legal pursuit?

3. Value Objectives of Personal Information Protection Law

The values of law mainly include fairness, justice, safety, freedom, efficiency, etc. They are pluralistic and closely related to the diversity of people's needs and social relations.2828See Zhang Wenxian, editor-in-chief: Jurisprudence (5th edition), Higher Education Press, 2018 edition, pp. 313-314.The general legislation is the result of the value selection and rank of each law. There are many application scenarios of personal information, and the value orientations involved are also quite rich. Re-integrate and analyze the anonymized information to identify that personal information violates the security of citizens' personal information, and the improper use of personal information is more likely to affect social, economic order and even national security; "killing" transactions based on differences in personal information affect consumption fairness; Information cocoon rooms and concerns about information exposure constrain the personal freedom of information subjects; absolute protection of personal information is not conducive to reducing transaction costs and affecting transaction efficiency. Personal information protection laws need to take into account multiple values such as security, fairness, freedom, and efficiency. In legislation, how to understand the relationship between various values is particularly important, which can be determined in the following ways: First, through the method of value rank or value order, a certain value occupies a dominant position, if it fails to provide such value If it is fully guaranteed, it is not allowed to achieve other values by compromising this value; the second is to use the methods of value conversion and value balance, so that different values can be compared with each other, and an institutional arrangement that maximizes the overall value is selected; the third is to use multiple objectives. Planning and other means, first set certain optimization standards, and seek the optimal mechanism design that can be achieved between the value judgment systems of all parties.2929See Su Yu: ”On the Value Objective and Mechanism Design of Algorithmic Regulation”, in Dialectics of Nature Communications, No. 10, 2019, pp. 8-15.Examining the legislative purpose of the Personal Information Protection Law, we can get a glimpse of its value goals. Article of the EU GDPR stipulates that the regulation stipulates the protection of personal information of natural persons and the free flow of personal information, focusing on protecting the basic rights and freedoms of natural persons, especially the rights of natural persons to personal information; the UK Data Protection Act will ”maintain the Credibility”, “promoting future trade development” and “ensuring security” are the goals; Article first of China’s Personal Information Protection Law stipulates that in order to protect the rights and interests of personal information, it regulates personal information processing activities and promotes the rational use of personal information. In general, the values of ”safety” and ”efficiency” are more prominent, and there is no obvious rank difference or order distinction between the two. It seems to be comparing the two with each other and seeking a balance. The following will try to find a suitable equilibrium point with the analysis method of usage economics.

There is an adjacent relationship between the protected personal information and the anonymized information used by the personal information processor. The boundary between the two is ambiguous, or there is overlap. In the process of exercising their rights, it may bring negativity to each other. Influence, which is called negative externality in economics. When making negative external behavior decisions, because the damage caused by the behavior is not within the scope of the external actor’s consideration, the actor will only decide the behavior level according to his own marginal cost and marginal benefit. Therefore, the behavior level will exceed the social maximum. Therefore, Pareto optimality cannot be achieved.3030See Wei Jian and Zhou Linbin: ”Law and Economics”, Renmin University of China Press, 2017 edition, p. 81.



(Figure 1)

As shown in Figure 1, when the security or anonymization standard of personal information processors is higher when processing personal information, the cost is also higher, and the benefits are lower and lower, the demand curve (D) or the marginal social benefit curve (MSB)) extends to the lower right, and the supply curve (S) or the marginal private cost (MPC) extends to the upper right. At the same time, it is difficult for personal information processors to absolutely avoid damage to the personal information rights and interests of information subjects when processing personal information, resulting in social The overall marginal cost curve (MSC) is higher than the MPC. MPC and MSB intersect at point A, when supply and demand are equal or marginal private cost and marginal social benefit are equal, and the corresponding anonymity standard is E. According to the principle of economics, when MSC=MSB and the two lines intersect at G, the Pareto optimality can be achieved, that is, when the anonymization standard is G, the efficiency of personal information utilization can be maximized. Obviously, if E is used as the anonymization standard at this time, due to the existence of negative externalities, the use of information by personal information processors is inefficient. In order to solve the problem of negative externality, it is the basic idea to internalize it. Common forms of internalization include voluntary negotiation, corrective taxes, liability rules, conduct controls, etc.

anonymization standard	high	medium	low
personal information processor	10	80	110
information owner	100	80	10
total social benefit	110	160	120

(table 1)

Taking Table 1 as an example, when the anonymization standard is high, it is assumed that the income of the personal information processor is 10, the income of the information subject is 100, and the total social income at this time is 110; when the anonymization standard is medium, it is assumed that the personal information processor The income is 80, the income of the information subject is 80, and the total social income at this time is 160; when the anonymization standard is low, assuming that the income of the personal information processor is 110, the income of the information subject is 10, and the total social income at this time is 120. When the anonymization standard is medium, the total social benefit is the largest. In order to reduce the negative externalities when the anonymization standard is high and low, because the personal information processor increases the income by 70, and the information subject reduces the income by 20, when the personal information processor Allocating more than 20 and less than 70 incomes to the information subject, both of which can increase income and achieve Pareto optimization. This optimization method is at least reflected in the consumption process of the information subject. For example, the personal information processor pays the cost to provide the information subject with free online chat rooms, online shopping platforms, self-media platforms, etc. Although the information subject does not directly increase the income, but Reduced costs. It is reflected in the Personal Information Protection Law that the information subject can request the personal information processor to compensate for losses and transfer part of the personal information processor's income to the information subject, thereby reducing negative externalities. At the same time, the government takes regulatory actions, requiring personal information processors to complete the effective prevention of personal information anonymization norms, and to punish, supervise and correct personal information processors that do not meet the behavioral standards.

Therefore, security and efficiency do not have the best rank in personal information protection, and only in the balance of the two can an institutional arrangement that maximizes the overall value be selected. This also determines that the protection and utilization of personal information cannot be neglected. Pursuing both at the same time is the correct choice to achieve the value goal of the personal information protection law. Therefore, this paper believes that the EU standard anonymization standard is too high, and security is the first value of the value, without considering the maximization of the total social benefit, which is not conducive to promoting information flow; while the British and American anonymization standards are in the middle position, and the overall social total benefit Higher, which is conducive to the protection of personal information and the promotion of information utilization, but which of the two is better needs further consideration.

4. Selection of anonymization standards

When talking about the anonymization of personal information, the inner world of the personal information processor is difficult for the information subject or judicial adjudicator to directly know. Through the above investigation, it is found that whether it is the European Union, the United Kingdom or the United States, when judging whether personal information is anonymized or not, both A "rational person" or "standard person" is introduced to evaluate the anonymization behavior of personal information processors, so as to resolve the conflict of interest between information subjects and personal information processors. Usually, the application of rational person standard includes three stages: rational person construction, scene reconstruction, and conclusion through cognitive schema. Among them, the construction of rational people is centered on the realization of rational people's knowledge structure and ability level.11See Ye Jinqiang: "Construction of the Standard of Rational Person in Private Law", in Legal Research, No. 1, 2015, p. 101.This is related to the innate and acquired qualities of rational people.22See Chen Xuan: "The Normative Essence and Judgment Criteria of Duty of Care", in Legal Research, No. 1, 2019, p. 139.At the same time, it should also be considered whether this rational person is abstract or concrete. The difference is that the abstract rational person has a clear standard or definition, while the concrete rational person's standard will vary according to the scene, and there is no uniform standard. The problem that arises from this is that abstract rational people may let go of people with ability and intelligence, and concrete rational people will make the law less predictable. Obviously, in the field of personal information processing, the financial, intellectual and human resources of large companies are higher than those of small companies; if the standard of rational people is higher, in terms of anonymization processing ability, it may be difficult for small companies to achieve anonymity despite their efforts. If the market structure of large and small companies is 1:9, large companies can easily pass the test of rational people, and small companies are not so lucky, and they will be placed in Damocles for failing to pass the test. Under the sword, to make matters worse, the information subject becomes the biggest victim. However, when the market structure of large and small companies is 9:1, under the same standard of rational people, large companies can still pass the test, small companies are still in difficulties and are gradually eliminated from the game, and the remaining large companies will meet with each other. Competing in the use of personal information will gradually break the rules and re-identify the processed information to seek greater benefits, and the original standards will no longer apply. Similarly, assuming that the standard of rational people is low, most companies in the market will pass the test regardless of the size of the company. What happens next will naturally be a competition for information ”re-identification”. At this time, we found that the original rational person standard should not be static, it should at least be dynamically adjusted with different market structures. But should it be absolutely individual? Large companies have high standards and small companies have low standards. In addition to the fact that the company cannot generate expectations, this solution ignores the protection of the rights and interests of the information subject. For example, when a company is small to a certain extent, can it be allowed to use almost no anonymity? What about the use of personal information? Obviously not. It would be great if there was such a rational person standard, most companies can pass the test, and there is no incentive for re-identification or the technology of re-identification or the cost of illegality exceeds the benefits they can obtain, and there is still room for a small number of companies. The re-identified company raises the standard of rational person and loses its interest drive. Of course, there are still some companies that cannot pass the test because they cannot meet the basic expectations of the information subject for personal information protection, and naturally there is no legitimate basis for information utilization. Take it out of the market. Therefore, this paper advocates that the rational person standard for anonymous judgment of personal information should be determined based on the principle of abstract standards and supplemented by concrete standards.

Comparing the British and American standards, regardless of the standard of rational people, the British standard seems to provide a relatively abstract anonymized judgment standard, but in fact, it needs to take into account ”more methods, knowledge, and skills” than ordinary people. What does it mean? Different scenarios, different parties, and different judges may have different understandings. It cannot give clear instructions to personal information processors, and it is difficult for information subjects to prove whether personal information processors have passed the rational person test. Equivalent to a concrete rational person. The American standard first established an abstract standard - safe haven. When a personal information processor deletes 18 kinds of identifiers in personal information, it is preliminarily presumed that the anonymization of personal information has been completed, and a very clear line is drawn. Personal information processing At the same time, the American standard also sets a ”duty of care” for the anonymization of personal information, when the personal information processor does not subjectively know or should know the information after deleting the 18 identifiers When it can be re-identified, it is considered to have achieved the effect of anonymization; the general and the individual are fully considered. Even a rational person in the ”Health Insurance Portability and Accountability Law” - a statistical expert, because of its relatively certain knowledge structure, cognitive ability, and action ability, it is enough to give clear guidance to information subjects or personal information processors. . Therefore, American standards are both abstract and concrete in legislative technology, which is worth learning from. After the legislative technology is determined, the level of the standard, that is, the number of types of identifiers to be deleted, should also be considered.

In fact, there are more than 18 identification features of personal information. Even if it is deleted, as long as there is enough information, there is a risk of being re-identified with the support of algorithms and computing power. Information processors at least strictly restrain themselves. After all, information is abundant. If the scope of the processed information combined with the identified information is too expanded, it is really harsh to attribute the re-identification performed by others on this basis to the personal information processor. In a sense, the existence of standards is to leave room for compliance for insufficient anonymization of personal information processors, to leave market space for the use of ”anonymized” information, and to avoid affecting the value of information. For the information subject, there is usually no obvious oppression and tension due to the deletion of a certain number of identifiers. Even if it is accidentally leaked, the threat to the personality and property interests of the information subject is obviously less33see item Ding Yi, Shen Jianping: "Research on the Consent Requirements for Commercial Use of Personal Information—From the Perspective of Personal Information Types", in Beijing Law Journal, No. 5, 2017, p. 35.,Take more relaxed regulatory measures for personal information with lower identification ability, and correspondingly reduce or even exempt information processors from their personal information protection obligations.44See Xing Huiqiang: "The Allocation and Realization Mechanism of Personal Information Property Rights in the Context of Big Data Transactions", in Law Review, No. 6, 2019, p. 100.Therefore, "18 kinds of identifiers" should not actually be a fixed number of types, but should increase with the development of society and the increase of the number of new types of identifiers generated by technological progress, which is closely related to a country or a society. The status quo of development is closely related. Some scholars believe that "in the future, my country's personal information protection legislation should shift from the current integrated regulation model to a differentiated regulation model based on the classification of information identification capabilities, and build hierarchical personal information protection obligations according to the different degrees of identification capabilities of different types of personal information. The rule system forms an alternative to the existing personal information anonymization rules."55Qi Yingcheng: Review and Alternatives to my country's Personal Information Anonymization Rules, Global Law Review, No. 3, 2021, p. 62.The author believes that it is not necessary, social life is diverse, business models are increasingly updated, personal information utilization scenarios are becoming more and more abundant, and the degree of typification corresponds to the accuracy of the cognitive model. The higher the accuracy, the greater the cost. It only makes sense when the cost of improving the accuracy is significantly smaller than the cognitive economic benefit of the accompanying precision improvement; moreover, it is necessary to avoid the "too fine-grained and technical terminology caused by typification, which quickly loses its role as a code of conduct and a referee." Demonstration effect" drawbacks.66See Jiang Ge: "Law as an Algorithm", in Tsinghua Law Journal, No. 1, 2019, pp. 71-72.Anonymization is a complex problem, and if it is supplemented by more complex typing, it may not help to solve the problem better. Faced with a problem that is too complex, one needs to make a trade-off between the speed, quality, and generalizability of the solution.77See Robert Sedgewick, Kevin Wayne, ALGORITHMS, Addison-Wesley, 2011, at 921.At this time, the legal system often has to reduce quality requirements in order to ensure speed and universality, and quality is by comparison the most suitable sacrifice of.88See Jiang Ge: "Law as an Algorithm", in Tsinghua Law Journal, No. 1, 2019, p. 73.The integrated regulation model of deleting identifiers makes it more difficult to re-identify information. Although it cannot completely guarantee the anonymity of processed information, it is not only highly universal, and the vast majority of personal information processors do not have to spend a lot of time, energy, and money. Learning and implementing the anonymization requirements under different types, and enabling the judiciary to determine whether the anonymization standard is met within a limited time, ensuring the speed of resolution.

5. Conclusion

No law can solve everything once and for all, and when discussing standards for anonymity, a distinction needs to be made between technical and legal. It is difficult and feasible to achieve absolute anonymity technically. However, when it is used as a legal standard, it must have certainty and predictability, which is a natural and inevitable relationship; legal researchers and practitioners must recognize that legal The "anonymity myth" is not unbreakable. The European Union, the United Kingdom, the United States and other countries have made many attempts on the issue of anonymization judgment standards. In comparison, the European Union prioritizes the protection of personal information rights, which leads to harsher personal information processors, which is not conducive to information utilization and is difficult to achieve. Maximize social benefits. While protecting the information subject, the Anglo-American standards leave enough space for personal information processors, which is of reference significance. Furthermore, the types of rational person standards in the United Kingdom and the United States are different. The United States adopts a compromise between abstract standards and concrete standards, which overcomes the shortcomings of using abstract standards alone to be unpractical and concrete standards to be unpredictable. It is worth learning from. In judging the standard of rational people, the method of deleting 18 identifiers in the United States basically covers the range of currently recognized identifiers and causes enough difficulties for those who attempt to re-identify information. As an abstract standard, it is reasonable The disadvantage is that it does not adopt an open legislative approach and does not take into account the new types of identifiers that will appear in the future.