Machine Learning, AI with Data Privacy

Zhiyi Chen*†

July, 2024

Abstract

With the advent of the big data era, the need for data protection methods to prevent exploitation by illegals is becoming more and more urgent. This paper looks at machine-learned methods such as differential privacy, homomorphic encryption, data desensitization, data obfuscation, and anonymization to protect the data information that people may leave behind when using AI. It also describes the content and differences between two landmark data privacy regulations: the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). The main findings of this analysis underline that the decision between these two regulations is determined by the organization's distinctive qualities and worldwide reach. The GDPR is more protective of people's privacy than the CCPA. To begin, the GDPR establishes stricter data protection rules than the CCPA, such as express consent, data subject rights, and data breach reporting. Second, the GDPR has a larger area of application for enterprises holding personal data that are in the EU, regardless of where the firm is headquartered. The CCPA applies solely to corporations in California; however, compliance is simpler in comparison to the GDPR. To summarize, the purpose of this article is to provide companies with the information and methods they need to make informed decisions on data privacy compliance.

Key words: Machine learning, AI, Data Privacy, GDPR, CCPA, Data desensitization, Anonymize, Differential Privacy.

1 Introduction

There are three major eras in human history: the steam era, the electricity era, and the information era. Since evolution, humans have started to use various tools to improve the quality of their

*Northeastern University (Massachusetts, USA)
†Email: 1825805784@qq.com

lives. Artificial intelligence (AI) is the most important technology in computer science so far. It mimics the intelligence of humans in areas such as learning, problem-solving, decision-making, and natural language understanding (Aldoseri, Al-Khalifa, and Hamouda, 2023). Besides, there are a lot of areas where AI is required. For example, smart security, smart robotics, automatic driving, and so on. The concept of AI is becoming increasingly popular. It seems to allow people to reduce their workloads at work and their chore time at home. The term artificial intelligence originated at the Dartmouth Conference in 1956 and was coined by John McCarthy. McCarthy proposed and summarized the topics in the fields of computers, natural language, neural networks, and other fields into the term artificial intelligence as a new academic study (Moor, 2006). AI has gradually entered our lives. There is no concrete definition of what artificial intelligence is. AI may be in mathematics, software engineering, linguistics, or psychology. AI researchers think they will know what the answer is before they are faced with the problem. They believe that some mathematical or other form must be the best way to express the content of the knowledge people have. So, AI is an exercise in finding the right formalism to represent knowledge (Schank, 1987). It depends on what method the person wants to use.

When people understand a little bit about artificial intelligence, they will also come along with terms like machine learning, deep learning, and neural networks. These words are all jargon, so much so that many people do not recognize their connection to AI. Machine Learning is a method for realizing artificial intelligence. It is a multi-disciplinary subject that involves probability theory, approximation theory, convex analysis, statistics, and more. Machine learning is at the heart of AI. It makes computers smarter, thus making them available in various fields. It is categorized into unsupervised learning, supervised learning, and reinforcement learning. Unsupervised learning allows machine learning to find patterns in networks and code on its own. Supervised learning makes predictions about the future of the real world by collecting past experiences and summarizing them. For example, it can determine exactly what an object is by collecting features. The central idea of reinforcement learning is to allow the AI to learn within its environment. Each action will correspond to its own reward, and the AI learns by analyzing the data and determining what should be done in what situation. Deep learning is a technique for implementing machine learning. Neural networks, on the other hand, are an algorithm for machine learning.

At the same time, privacy has become a big issue in this era of machine-learning-based artificial

intelligence. The concept of the right to privacy was first introduced in 1890 by American private law scholars Brandeis and Warren in an article in the Harvard Law Review (Glancy, 1979). People are beginning to awaken to privacy concerns. When artificial intelligence and machine learning began, concerns about their impact on society and potential misuse increased. In response, governments and organizations have begun to regulate AI to ensure that AI technology is used responsibly and ethically. Issues like privacy, transparency, and fairness are now at the forefront of the AI debate. Privacy is also no longer a traditional privacy issue after being associated with artificial intelligence. And machine learning has both advantages and disadvantages for people. For example, when technology is used in healthcare or to protect the environment, people can sense or predict the patient's condition more accurately and faster through the technology. Thus, doctors can make can prevent it earlier. For example, when a large amount of data is used by lawless elements, people's bank card information and personal information may be leaked. Because of the development of modern technology. People are organizing and saving their data in a very different way than in the past. And this data will be stored on the Internet forever. People do not know if the companies or organizations they register with are quietly collecting information about their users or consumers and selling it in the future. Consumers can also sometimes feel that they are being monitored by software or websites. By tracking a user's online activity, for example, marketers can provide targeted advertising and content. For example, individual software can be personalized to constantly adapt to user preferences.

In this paper, I will discuss what artificial intelligence and machine learning are. In addition, I describe in detail the importance of mainstream data protection methods such as differential privacy, homomorphic encryption, data desensitization, data obfuscation, anonymization, and the GDPR and CCPA laws in terms of today's information protection methods (technical and legal). And conclude by describing where current technology needs to be developed even further.

2 Data Privacy Techniques in Machine Learning

Big data contains significant business value, especially when it includes users' private information. Various industries are analyzing and organizing big data. For example, a Web Crawler is a kind of program or script that automatically crawls internet information according to certain rules.

Operators and enterprises explore the potential business value to maximize their profits. Therefore, as people enjoy the convenience brought by the era of big data, they will inevitably suffer from privacy leakage. People have started to take an interest in developing techniques to protect privacy. There are lots of approaches to protect privacy. In general, they can be categorized into privacy protection techniques based on restricted release and privacy protection techniques based on data encryption. For example, data desensitization, anonymization, differential privacy, and homomorphic encryption. Today, many program areas benefit from data sharing. However, many times, data sharing is undesirable. This is especially true in the healthcare industry, where private information about a patient's condition is unacceptable, regardless of the channel through which it is accessed.

2.1 Data Desensitization

Data desensitization is widely used in various industries. It is also a technique that is now commonly used in industry to handle private and sensitive data. It was first proposed by a statistician in the late 1970s (Dalenius, 1986). He argued that if private information in a database is protected, it means that no one can access the database and get accurate information about anyone. There are many methods of data desensitization, and its main objective is to reduce the sensitivity of sensitive data through replacement, distortion, and other transformations while retaining certain usability and statistical characteristics. One of the most widely known aspects is the mosaic technique. It protects private information about individuals by blurring or obscuring the private parts. Alternatively, the personal information in the database is encrypted one by one so that prying eyes cannot get the user's personal information accurately. For example, in Chinese hospitals, the last character of the last name and the first name are displayed on the caller ID screen. In this way, no one else can guess the real name of the patient who has just entered the consultation room. These are static desensitization. The more representative ones are development, testing, data analysis, and so on. Another type of desensitization is dynamic desensitization, which is generally used in production environments. This is generally used for direct access to the production environment. It is needed to connect to the production data, such as customer service personnel, through the application to query the user information and so on.

2.2 Anonymize

The second method is to anonymize the data. In the age of big data, many organizations or institutions need to make their research results or data public. For example, academic research and medical data. Before these organizations release data information, they usually need to anonymize the private information. Privacy preserving data publishing (PPDP) provides a set of models, tools, and methods to guard against privacy threats posed by data published by data miners or analysts (Majeed and Lee, 2020). PPDP provides a set of models, tools, and methods to guard against privacy threats posed by data published by data miners or analysts. PPDP has two well-known settings, non-interactive and interactive. Non-interactive is where the data holder simply tinkers with the raw data. They then publish the complete dataset in an anonymized manner. Interactive, on the other hand, is where the data owner provides an interface to the data miners, and they get different results (Majeed and Lee, 2020). Suppose a manager of a stock exchange needs to show his performance to his managers. As shown in the example in Table1, the form has data on the customer's name, gender, age, zip code, SSN, and the product the customer purchased.

Original data form					
Name	Gender	SSN	Age	Zip-Code	Product
Alice	Female	123-45-678	22	01111	Stock
Alex	Male	234-56-789	22	01111	Fund
Benjamin	Male	345-67-890	22	01111	Fund
Candy	Female	568-89-110	22	21222	Bond
David	Male	898-98-234	22	23322	Stock

Table1: Original data form

To protect the privacy of the customers, the manager replaces the key information (name, SSN) with * and shows it to his manager.

Anonymize data form					
Name	Gender	SSN	Age	Zip-Code	Product
***	Female	***	22	01111	Stock
***	Male	***	22	01111	Fund
***	Male	***	22	01111	Fund
***	Female	***	22	21222	Bond
***	Male	***	22	23322	Stock

Table 2: Anonymize data form

But if the leader wants to find a particular customer, they can make a guess by looking at it in conjunction with the admission form. As soon as the two tables are merged, it is easy to spot the original data (Table 3). This does not serve to protect private information.

Admission form				
Name	Gender	City	Age	Zip-Code
Alice	Female	LA	22	01111
Alex	Male	LA	22	01111
Benjamin	Male	BOS	22	01111
Candy	Female	NYC	22	21222
David	Male	BOS	22	23322

Table 3: Admission form

So, people will use k-anonymity to circumvent this vulnerability. This concept was introduced by Sweeney (Sweeney, 2002). The k-anonymity model hides private information by categorizing different values. It prevents the privacy attributes of any one record in the database from corresponding one-to-one. The principle of k-anonymity is to categorize Name, Gender, City, Age, and Zip-code into quasi-identifiers (QI_T and then have k matches). When k-anonymity is performed, people are less likely to learn private information by combining different forms.

2.3 Differential Privacy

Differential privacy prevents people from inferring private data from newly obtained data. For example, in Table 2, one knows that it is already known that two people purchased stocks and funds,

and one person purchased bonds. Then, when the data of a new person is added to Table 2, it will be easy for people to know which product the new person invested in. Differential privacy considers an algorithm that analyzes a data set and calculates its statistics, such as the mean, variance, etc., of the data. With this algorithm, it is difficult to change the behavior of individual data when it is added to or removed from the dataset (Differential Privacy, n.d.).

2.4 Homomorphic Encryption

Homomorphic encryption is a type of encryption in the field of cryptography. Homomorphic encryption is considered more secure than traditional encryption. This is because traditional encryption schemes cannot encrypt data without first decrypting it. Users must compromise their privacy when using cloud services such as file storage, sharing, and collaboration. Homomorphic encryption provides a perfect solution to this problem. It preserves the functional characteristics and format of the original encrypted data when users use third-party software or websites (Acar et al., 2018). This ensures that when people use untrustworthy third-party cloud platforms, homomorphic encryption stops private data leakage due to third-party theft.

2.5 Data Obfuscation

Sometimes, differential protection and homomorphic encryption may not fully secure private information. For example, merely a clever choice of plain text or secret information is very much open to selective text attacks on encryption techniques like privacy homomorphism. This is the time when data obfuscation techniques come into play in the dataset. The main advantage of this technique over data encryption is the ability to distribute different amounts of obfuscated data depending on the end user's needs, thus providing multiple levels of data protection (Bakken et al., 2004).

3 Policy About How to Protect the Privacy

In the age of big data, most organizations or businesses seek access to people's data. They want to use private data to better understand or target individuals. Organizations aim to capture customer needs and customer segments through data analytics, while advocacy organizations want to use data analytics to understand the people who might join them. However, there are individuals with malicious intentions, such as scammers, who seek to exploit private data to various scams. Because there are numerous ways to capture people's private information, some governments have introduced corresponding laws to protect privacy, such as The Data Protection Act 2018. The regulations safeguarding data privacy in practically all EU member states have been replaced with the General Data Protection Regulation (GDPR). In California, USA, the California Consumer Privacy Act (CCPA) was established following the adoption of the GDPR. It is important to note that the GDPR and the CCPA are currently two of the most well-known and extensively used legislations in existence. These two measures have significantly enhanced the privacy of individuals by regulating businesses and imposing restrictions and controls on the handling of sensitive information. The General Data Protection Regulation and the California Consumer Privacy Act protect people's privacy from being exploited by unscrupulous individuals.

3.1 General Data Protection Regulation

The goal of the General Data Protection Regulation (GDPR) was to enhance people's rights over their own privacy information in databases. It applies directly to personal data processing activities related to EU territories or markets (Albrecht, 2016). The GDPR was adopted in April 2016 by the EU Conference and the EU Council. After a transitional period, the GDPR began enforcing regulations in May 2018, and sanctions and fines can be imposed on companies and sectors that fail to meet the standards. Starting from this date, companies violating its provisions may face sanctions of up to 4% of their global annual turnover, or up to 100 million euros in all cases (Albrecht, 2016).

Its main principles include legality, fairness, confidentiality, security, transparency, purpose limitation, storage limitation, data minimization, integrity, and accountability (Zaeem and Barber,

2020, 2021). Due to the detailed and specific nature of these principles, the GDPR is one of the most comprehensive personal data protection laws in existence. The GDPR brings an end to the fragmented digital market and strengthens the enforcement of data protection regulations. For example, regarding cookies, which are pieces of information that identify the user, compliance with GDPR parameters must include a Cookie Accept button. This button allows users to choose which type of cookies to accept. Other requirements include providing information about the purpose for which the data will be used, specifying whether third parties use the user's data, and including a link to the privacy policy or cookie policy page, which must describe the cookies and the conditions for acceptance (Pantelic, Jovic, and Krstovic, 2022).

3.2 California Consumer Privacy Act

Similar to Europe, the United States did not have a well-developed data protection law in its early days. Inspired by the GDPR, California issued the 'California Consumer Privacy Act' (CCPA). This landmark statute represents the first comprehensive data privacy legislation in the United States, addressing a significant gap in data privacy-specific legislation. Like the GDPR, the CCPA is designed to strengthen consumer privacy and data security protections in California and is currently the most stringent consumer data privacy protection legislation in the U.S. It applies to any resident of California, making it an interstate law. Therefore, if a business providing services to California residents meets the CCPA's applicability threshold, it must comply with its privacy provisions when collecting, processing, buying, and selling users' personal information.

Meanwhile, the CCPA, while an interstate law, does not really affect just the state of California. There are two reasons for this. First, it is well known that most high-tech conglomerates are headquartered in Silicon Valley. Silicon Valley companies must comply with the provisions of the CCPA (Baik, 2020). Those companies are basically complying with the protective laws of the CCPA. At the same time, the CCPA is either a benchmark for federal law or an important reference for state lawmaking. So, some small businesses in other states invariably comply with the CCPA. Also using cookies as an example, the parameters to observe when considering whether a cookie banner complies with the CCPA (California Consumer Privacy Act) are as follows:

Use of Cookies - Users should be informed about cookies and whether the site owner is sharing

the information with third parties.

Cookie acceptance buttons - one of the most important differences between the GDPR and the CCPA is that a website can delete some cookies before the user clicks the acceptance button.

No-sell buttons - the CCPA requires companies to provide customers with the option to opt out of the sale of personal information (Pantelic, Jovic, and Krstovic, 2022).

3.3 Difference Between GDPR and CCPA

In previous parts, I detailed the CCPA and GDPR models, which are both extensively used data privacy and protection frameworks. In this section, I will compare the two models and discuss their distinct strengths, flaws, and implications for organizations and individuals. My goal in examining the key differences and similarities between the CCPA and the GDPR is to gain a comprehensive understanding of these key regulations and assist organizations in making informed data privacy compliance decisions, ultimately ensuring the protection of an individual's personal information in the evolving digital landscape.

3.3.1 Regulatory Scopes and Obligations

According to the CCPA, it applies to businesses that conduct business in California for profit or financial advantage and whose business involves the acquisition and/or processing of personal information that fits one or more of the conditions listed below (Torre, 2018).

1. Annual gross revenue exceeds $25,000,000.
2. Purchases, sells, or transfers personal information from over 50,000 consumers, devices, or households.
3. Generates more than half of its annual revenue from the sale of users' personal information.

The GDPR is different in that anyone who has a place of business in the EU, or a member state needs to comply with the GDPR laws (Voigt and Von dem Bussche, 2017).

Compared to the GDPR, the CCPA does not apply to nonprofit organizations or California state and local government entities. The regulatory scope of the GDPR covers virtually any organization or business that processes the personal data of EU citizens. In a nutshell, the CCPA applied to companies that use customer information to generate revenue, while the GDPR applied to all companies that collect customer information.

3.3.2 Rights of Data Subjects

California consumers are given the ability to access their personal data, request its deletion, and choose not to have their data sold as part of the CCPA policy (Torre, 2018).

The GDPR policy grants data subject additional rights, such as the opportunity to transfer their data, the right to rectification, and the right to be forgotten (Voigt and Von dem Bussche, 2017).

3.3.3 Civil penalties

CCPA: Imposes fines of up to $2,500 for inadvertent violations and $7,500 for willful ones (Torre, 2018).

GDPR: Applying fines of up to EUR 20,000,000.00 or up to 4% of worldwide yearly revenue, whichever is higher, imposes harsher penalties (Voigt and Von dem Bussche, 2017).

3.4 Pros of the CCPA and GDPR

Customers are granted numerous rights under the CCPA regulation, including the power to request that a website or organization remove their data and access to their own personal data. It also compels businesses or groups to reveal the methods by which they gather, utilize, and handle personal information. All of this can assist customers in taking greater control over their personal data and in making wiser decisions in various scenarios.

In addition to giving people in the EU rights like the right to data correction, the right to object, and the right to revoke consent, among others, the GDPR is currently the most comprehensive data privacy law. Simultaneously, it stipulates steep sanctions that will force businesses or organizations to prioritize data protection.

3.5 Cons of the CCPA and GDPR

The CCPA's narrow range is by far its most noticeable drawback. It does not restrict businesses or groups operating in other US states; it only applies to the state of California. The GDPR's drawback is that small firms find it more difficult and expensive to comply with.

4 Discussion

In response to growing concerns about the security and confidentiality of personal data, privacy protection has undergone significant legal and technological changes. Traditional privacy protection technologies often fail to meet the privacy protection needs of users, especially in an era when data is outsourced to third-party companies. Many third-party entities separate ownership and control of software, raising concerns about data management and ownership, resulting in individuals having limited control over their private information.

In recent years, machine learning technologies have played an important role in addressing these challenges by providing a variety of methods to protect private information. These methods range from differential privacy, data obfuscation, and data desensitization to homomorphic encryption and anonymization. Each technique serves a specific purpose and targets different scenarios to ensure data privacy while allowing for data sharing and processing. In addition, the introduction of regulatory frameworks such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) provide a legal basis for enhancing the privacy and security of personal information.

Despite these significant advances, several key issues remain with the current approach to information protection. From a legal perspective, the CCPA, while a substantial step forward, is still not as broad and robust as the GDPR, which encompasses the entire E.U. The CCPA provides better protections for consumers, while the comprehensive scope of the GDPR sets a higher standard for data protection. In addition, the CCPA's definition of privacy is often abstract and relies on illustrative examples. While this can provide clarity, it also opens the door to malicious exploitation of individuals. At the same time, differences in personal information protection definitions across jurisdictions and the flexibility in interpreting these definitions create challenges for uniform pri-

vacy enforcement. As data models evolve, privacy laws and regulations must continue to adapt to enhance protections. The evolving nature of data models and technologies introduces a degree of variability in interpreting personal data concepts. Laws and regulations need to continually adapt and strengthen privacy protections to keep pace with these developments. This means that reaching a universally accepted understanding of personal data remains a difficult challenge.

From a technical perspective, there have been challenges in the application of privacy-protecting technologies, particularly in encryption. Existing technologies have vulnerabilities as leakage of encryption keys can lead to data leakage. Even without the key, individuals proficient in encryption may find ways to recover the original data, highlighting the need for continuous innovation and improvement in encryption methods.

In summary, while substantial progress has been made in privacy protection through legal frameworks and technological advances, the ever-changing nature of data and the dynamic threat landscape still require continuous improvement. The complexity of personal data, changes in regulatory definitions, and persistent gaps in technology emphasize the need for a comprehensive and evolving approach to protecting privacy in the digital age. Privacy protection is an ongoing process that requires a concerted effort from legal, technological, and regulatory perspectives to keep pace with the evolving landscape of data security and privacy.

References

Acar, Abbas et al. (2018). "A survey on homomorphic encryption schemes: Theory and implementation". In: *ACM Computing Surveys (Csur)* 51.4, pp. 1–35.

Albrecht, Jan Philipp (2016). "How the GDPR will change the world". In: *Eur. Data Prot. L. Rev.* 2, p. 287.

Aldoseri, Abdulaziz, Khalifa N Al-Khalifa, and Abdel Magid Hamouda (2023). "Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges". In: *Applied Sciences* 13.12, p. 7082.

Baik, Jeeyun Sophia (2020). "Data privacy against innovation or against discrimination?: The case of the California Consumer Privacy Act (CCPA)". In: *Telematics and Informatics* 52.

Bakken, David E et al. (2004). “Data obfuscation: Anonymity and desensitization of usable data sets”. In: *IEEE Security & Privacy* 2.6, pp. 34–41.

Dalenius, Tore (1986). “Finding a needle in a haystack or identifying anonymous census records”. In: *Journal of official statistics* 2.3, p. 329.

Glancy, Dorothy J (1979). “Invention of the Right to Privacy, The”. In: *Ariz. L. Rev.* 21, p. 1.

Majeed, Abdul and Sungchang Lee (2020). “Anonymization techniques for privacy preserving data publishing: A comprehensive survey”. In: *IEEE access* 9, pp. 8512–8545.

Moor, James (2006). “The Dartmouth College artificial intelligence conference: The next fifty years”. In: *Ai Magazine* 27.4, pp. 87–87.

Pantelic, Ognjen, Kristina Jovic, and Stefan Krstovic (2022). “Cookies implementation analysis and the impact on user privacy regarding GDPR and CCPA regulations”. In: *Sustainability* 14.9, p. 5015.

Schank, Roger C (1987). “What is AI, anyway?” In: *AI magazine* 8.4, pp. 59–59.

Sweeney, Latanya (2002). “k-anonymity: A model for protecting privacy”. In: *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05, pp. 557–570.

Torre, Lydia de la (2018). “A guide to the california consumer privacy act of 2018”. In: *Available at SSRN 3275571*.

Voigt, Paul and Axel Von dem Bussche (2017). “The eu general data protection regulation (gdpr)”. In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10.3152676, pp. 10–5555.

Zaeem, Razieh Nokhbeh and K Suzanne Barber (2020). “The effect of the GDPR on privacy policies: Recent progress and future promise”. In: *ACM Transactions on Management Information Systems (TMIS)* 12.1, pp. 1–20.