Technology in Society 76 (2024) 102457

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Data privacy: From transparency to fairness☆

Chao Wu (吴超)[1]

School of Public Affairs, Zhejiang University, China

ARTICLE INFO

Keywords:
Data privacy
Data economy
Data fairness
Privacy regulation
Distributed modeling

ABSTRACT

In recent years, data privacy has attracted increasing public concerns, especially with public scandals from top Internet companies, emerging over inappropriate data collection, lack of transparency in data usage, and manipulation of users' preferences. This has led to responsive efforts from multiple sectors of society to constrain the violation of individuals' data privacy. Among these efforts, the regulations like GDPR are the most influential, which are believed to be able to change the foundation of the whole data ecosystem. The main concern of current regulation is data transparency, designed to enable individuals to make rational decisions about their data. However, I argue that transparency cannot mitigate a key issue of data privacy, which is the right of receiving benefits from data. This issue is getting severe with the rapid development of the data economy and will become the essential problem of social welfare and fairness in the AI era. I further point out that the key to solving this issue is to migrate the current centralized data utilization paradigm to a new decentralized paradigm. Based on the recent technical advancements, I propose an applicable pathway to implement such a paradigm. And to realize a fair and sustainable data eco-system, joint efforts should be made within this paradigm.

1. Introduction

Although data privacy had become an issue even before the emergence of the Internet [1], it's getting the intensive attention of the public in recent years. With the pervasive usage of new ICTs including high-speed networks, cloud computing, and pervasive data collection [2], we now have a great availability of data which is breeding an emerging data ecosystem. In the meantime, a series of public concerning events like the Cambridge Analytica scandal, in which personal data of up to 87 million Facebook users were collected without their consent [3], has attracted unprecedented discussions about data privacy problem. The concern is not just about the invasive business of Internet giants, but also about the government and other private/public sectors (such as the contact tracing during Covid-19 [4]) (see Table 1).

In response to such a challenge to traditional public values, efforts have been made by different parties. From companies, how to convince their customers that their services and products protect user privacy has become the focus of business and marketing strategy in the past couple of years. Many technical solutions were proposed and deployed [5,6]. Companies also published various AI Ethics guidelines [7] with privacy protection as their primary concern. From the research community, in addition to traditional privacy protection research mainly focusing on access control, authorization, anonymization, and encryption [8], recent research has focused on proposing methodologies to enhance privacy protection in the Internet era, e.g., wireless location protection [9], personalized search [10], video surveillance [11], and blockchain [12]. From governments, attempts have been made from the legal and regulatory perspectives. The EU's GDPR (General Data Protection Regulation) in particular has had an important impact [13]. Many states in the U.S. have also adopted their privacy protection laws, such as the CCPA (California Consumer Privacy Act) in California.

However, despite these efforts, the data privacy issue is still severe. It is caused by the **inherent tension between privacy protection and data utilization**. As AI is becoming a new engine of economic development, such tension is getting even more intense. Data are the fuel for AI modeling, and are gradually viewed as a new type of asset. However, how to protect this special asset is challenging and has not been well solved in the current regulatory and legislative framework [14]. Therefore, to build a trustworthy data privacy environment, we have yet to take a significant step forward. **I argue that the current efforts are mainly about data transparency**, which is difficult to achieve in the real context. More importantly, although transparency helps people understand how their data are being used, **it cannot mitigate the key issue of data privacy, which is the fairness of receiving benefits**

☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
E-mail address: chao.wu@zju.edu.cn.
[1] No. 866, Yuhangtang road, Sandun district, hangzhou city, zhejiang province, 310,058, China (Permanent address)

https://doi.org/10.1016/j.techsoc.2024.102457
Received 17 March 2022; Received in revised form 7 December 2023; Accepted 1 January 2024
Available online 3 January 2024

Table 1
Summarizes the main ideas, and examples of applications of these technologies.

Technology	Main idea	Application examples
Federated learning	Moving modeling to data owners, and aggregating local models to a global model.	Healthcare [70], city management [71], mobile applications [78].
Secure multi-party computation	Jointly computing a function while keeping data from data owners private.	Scientific computations [79], data mining (Bogdanov et al., 2012).
Differential privacy	Sharing aggregated information while withholding data about individuals	Behavior prediction [80], personalized online advertising [81].
Blockchain	A decentralized, distributed, and oftentimes public, digital ledger.	Energy [82], food supply [83].

from data. For example, big tech companies such as Facebook and Google reap huge financial benefits by collecting user data to provide customized ads, but users often don't know about it or can't get a corresponding return from it.

I define **data fairness as the state in which various parties, especially including individuals, receive fair revenue (monetary value) from the data that belong to them**. Please note that the discussion of fairness is limited in its economic aspect, and excludes the other ethical dimensions like bias and accountability. **This kind of data fairness issue will become the essential problem of social welfare in the AI era.** Such fairness has been noticed before (such as in MIT, 2021). And in this work, we provide a more comprehensive analysis of it along with a technological solution for redressing the balance of the data economy. In this work, I will discuss the key logic of data fairness that is ignored in previous privacy protection. I hereby propose a decentralized paradigm and a pathway of potential technical and management solutions to achieve data fairness.

The main contributions of this paper are:

- It reviews the data privacy issue in the AI context and makes a comprehensive discussion about the impact of regulatory effort as well as its limitation.
- It re-identifies privacy issues from a data fairness perspective, which is about the right of sharing the revenue from data.
- It further points out the key issue of data fairness is the centralized data utilization paradigm and proposes a decentralized paradigm.

Please note that I constrain the discussion within the scope of personal data privacy. I also define two roles in the data Eco-system: **data owner** (in many cases, it's also the creator of data) and **data user**. Data owner can be an individual (or any organization) from which data are generated and sensed. A data user is an entity, typically a company, that utilizes data from a data owner for various purposes, such as processing, modeling, aggregating, and selling of data. Under this definition, data users include the data processor and data broker.

This paper is organized with the following structure. Section 2 reviews the increasingly severe issue of data privacy. Section 3 discusses the regulatory efforts and explains why the focus on transparency is not sufficient to ensure effective privacy protection. Section 4 analyzes the properties of data as a new type of asset. In Section 5, the concept of 'Data Fairness' is examined, along with the challenges associated with achieving it within a centralized paradigm. Therefore, a new decentralized paradigm is proposed in Section 6. A further discussion and conclusion are then made in Section 7.

2. Why is the data privacy issue getting worse?

The Internet pushed the data privacy issue under the spotlights, especially when **the Internet was transforming from service-driven to data-driven**. Traditionally, the service and platform providers on the Internet cared more about the information, features, and services provided to their users, in which data collection was more like a by-product. But in recent years, data have become the central focus and more business models based on data were created. This data-driven paradigm drove several factors that intensified the data privacy issue:

1. Data can be more easily collected to form Big Data [15]. Along with over 20 billion sensors [16], people are generating about 2.5 quintillion bytes of data on daily basis [17]. The privacy problem is not just about the great volume: Firstly, these data are more sensitive because many sensors are intimate to humans; Secondly, these data and associated online services are now ubiquitous, collecting data from all aspects of our personal life [18]; Dataveillance, which is the systematic monitoring of people or groups from data [19] is now reinforced with the availability of Big Data (Degli, 2014). Thirdly, these sensitive data are more easily disseminated through high-speed network transmission and low cost of data duplication. A piece of data can be simultaneously harnessed by multiple parties, and it becomes more vulnerable to privacy attacks [20]. Since computing is now distributed, privacy leakages have become more likely [21].
2. Data collection is convergent to the Internet giants. Big companies like Google and Facebook are now able to collect users' data comprehensively, as people move more personal activities and social interaction to their online platforms. On one hand, these companies have high coverage of users, which enables them to understand the overall picture of the population. On the other hand, these companies have multiple channels to collect a single user's data (web searches, emails, location, etc.), which through systematic combination can make the comprehensive profile of the user exposure to the companies (Machanavajjhala, 2006). Google's privacy policy has extended from 600 words in 1999–4000 words in 2019 [22], reflecting the increasing range and amount of what Google collects. Many activities of these big companies are widely perceived as a potential violation of privacy norms. Companies mainly focus on revenue, cost-efficiency, and scalability as their primary goals, which legal compliance with security and privacy has often taken a backseat (Shastri, 2019).
3. Big data economy and AI. With the advancement of our capability of utilizing big data [15], data have demonstrated great value [23–25], and is viewed as the new engine of economic growth. Data's role is further emphasized in AI, especially for deep learning, which consumes a large amount of data for model training. Therefore, extending the applications of AI accelerates the growth of data collection. For example, more sophisticated recommendation systems are based on the monitoring of users' purchasing patterns. And this kind of active collection and usage of data make it vulnerable to privacy violation (Papernot, 2016), such as targeted marketing.

With all these factors, user privacy is attracting more public concern in recent years (Victor, 2016), especially about the privacy violation scandals. Tech companies have a history of constantly testing legal limits on privacy invasion. Take Facebook as an example, it has faced legal opposition and social protest for its applications like DeepFace (Facial recognition for Tag suggestion), expensive tracking methods [26], and sharing users' data with third parties without the users' consent [27]. More notorious events include the 2012 presidential campaign (in which Facebook was criticized for allowing the Barack Obama presidential campaign to analyze and target select users by providing the campaign with friendship connections of users) and Facebook–Cambridge Analytica scandal. Take Google as another example, Google has been criticized for its actions including the scanning of users' email [28], collecting users' locations [29], fusing users' data across its multiple services [30], etc. Recently, some companies take action for better privacy protection and even claim privacy as the primary goal of their products and services.

In addition to private enterprises, the practice of collecting and analyzing data from the public is increasingly a core part of how modern

government bodies operate [31], and governments are criticized for monitoring citizens based on their online data [32,33]. However, more discussion about governmental data misuse is out of the scope of this paper.

3. Privacy regulation

In response to the privacy issues discussed before, the solution from the government is to introduce the privacy protection regulatory frameworks. In the US, the Consumer Privacy Protection Act (2011, 2015), Commercial Privacy Bill of Rights Act (2011, 2014), Data Broker Accountability and Transparency Act (2014), and California Consumer Privacy Act (CCPA), were all designed to inject transparency into commercial data collection. In the EU, the Data Protection Directive 95/46/EC (DPD95) has been adopted to protect the privacy of personal data collected, defining the privacy of personal data as a fundamental right of all European citizens. In May 2018, the General Data Protection Regulation (GDPR) came into effect to replace the DPD95 and to harmonize the laws across the EU. While the DPD95 was just a directive, the GDPR is a regulation [13]. GDPR puts "data protection by design and default" at the core, and utilizes the consent of the data subject as the central vehicle to regulate the data lifecycle, data subjects' rights (including new rights of data erasure and portability), and legal obligations, to ensure the protection of personal data of EU citizens. Since 2018, GDRP has caused significant impact [34], and became a challenge for Internet companies [35]. While GDPR has been regarded as the trigger for some positive effects [36], there are also some unintended impacts. Some criticized that the formulation of the GDPR was vague and ambiguous [37]. Other issues include the negative post-GDPR effects on EU ventures, relative to their US counterparts [38]. To summarize, the effectiveness of GDPR is still difficult to conclude at the moment [39].

The efforts of public policy and regulatory oversight focus on **transparency.** Such transparency is mainly achieved by the consent mechanism, and is designed as a means to improve the control of individuals on their data, and empower data owners to transact with data users, as equals within data eco-system. It is widely believed that data owners need access to information about data collection and processing to make rational decisions about their data. However, although theoretically, transparency can eliminate the information asymmetry and thus logically protect users' privacy by "giving consumers a seat at the table" [40], but there are many problems in the implementation process :

1. First of all, comprehensive transparency is difficult to implement. Data are not easily traceable once collected. A data processor can apply a series of operations upon data (aggregation, transformation, modeling, etc.) to make it unidentifiable as the original data. Transformed data and the models upon it can then be sold in multi-layered markets. For example, consumers' transaction history can be mined, enriched, and used to train a recommendation model, which extracts the precious while sensitive knowledge from data. Therefore, information asymmetry still exists between the data owner and data user by separation via complex data processing and data markets. It is difficult to control the usage of data in "downstream" contexts.
2. Secondly, although the regulatory efforts have the power to fundamentally change the eco-system of the data economy, it will be a long process with continuous efforts. The technical developments may outpace the regulatory framework. And responsive changes from data users (especially the big tech companies) are sometimes hypocritical and deceptive, to make an illusion of transparency while leaving basic power imbalances intact. In response to the regulatory forces and public concerns, and to maintain customer trust today, many companies demonstrated that data privacy is one of their core values in recent years. However, many actions are to provide users a vague image of the extent of data protection and largely misconstrue actual practices of data collection and utilization. For example, most Internet companies are requested to provide a privacy policy for their services. It is ironic that while these policies have been greatly extended [41], they suffer from bad readability and other complexities which make them unusable [42]. Although it is not a fundamental problem and can be mitigated through regulation, it demonstrates the asymmetric position between the big tech companies and the end-users (even also between the big tech companies and regulations). Therefore, a lot of extra regulatory efforts are needed in the future.
3. Thirdly, these regulations are not fully incompatible with the current data economy environment. Transparency violates many inherent business models in the data economy, including the data broker industry. For example, the capability of extracting incremental data value depends on the algorithms and models belonging to the data broker. If the algorithms and models are forced to become transparent, the data broker then cannot protect its core business value. Zarsky believed current regulations failed to properly address the surge in Big Data practices and made the data ecosystem suboptimal and inefficient [43]. It is not that transparency has not yet been properly configured, it is that the consumer empowerment frame misunderstands key dynamics of commercial surveillance and therefore offers flawed policy solutions. In many people's beliefs, privacy is even seen as a barrier to economic growth [44]. More importantly, **current regulations ignore the citizens' right of getting a fair share of the revenue from data, which will become more crucial in the future data economy.** I will discuss this in the next section.

An evidentiary example of the above issues is the emerging market of behavioral prediction and modification, in which data are used for massive consumer profiling and ad targeting. In this market, data broker (transnational data corporations, like Acxiom, Experian, and ChoicePoint) plays a key but controversial role in data monetization (U.S. Committee on Commerce, Science, and Transportation, 2013). They gather consumer data as raw materials from millions of people, to create various informational goods and provide services, and then distribute them to clients for a wide range of uses. Acxiom alone claims to retain over 3000 pieces of information for nearly every adult consumer in the United States and offers "multi-sourced insight into approximately 700 million consumers worldwide" (Acxiom [45]. These data brokers historically operated under the radar of consumer awareness but got intensive attention and regularization in recent years. Many efforts were brought in to make the industry more transparent but although a certain degree of transparency was achieved, most responses from the industry are superficial (U.S. Committee on Commerce, Science, and Transportation, 2013). As discussed before, the data broker can apply complex data processing and repackaging, to make the real usage of data difficult to track. More importantly, the principle of transparency is incompatible with their core business value. The global data broker industry is estimated to comprise around 5000 companies of various sizes generating $178 billion in revenue in 2020 (Twetman, 2021). Data are their core assets, and the right to get value from data is their core interest. Total transparency will leave no space for this industry and thus is difficult to achieve. A more realistic target should be data fairness.

4. Data as assets

Data privacy entitles an individual the rights to his/her data. **As the right of transparency is difficult to preserve, it's necessary to protect another type of right, which is about fairly receiving monetary value that an individual can derive from his/her data.** To understand why such fairness of data revenue is as important as (if not more important than) transparency, we need to look at data from the economic perspective. In this section, I will check the economic properties of data, and then discuss their impacts on privacy protection.

Data transform the economy from two aspects, one is about the enhancement of existing business, and the other is about creating new business:

1. For the first aspect, data are the key inputs to extract added value from the existing commercialization model [46]. In a data-enhanced commercialization model, individuals are no longer mere consumers of goods, information, and services, but the producers of valuable data (e.g., location traits, comments, photos, and purchasing histories), which can then be used to improve the service and product provided, like a targeted advertisement, financial offerings, product recommendation, etc. Since not all value derived from data is primarily monetary, its impact on the current economy is under-estimated. But this model is so successful, that as an intrinsic component of life, data have now become a primary target of commercialization strategies, and it has transformed marketplaces, altering how firms and consumers interact [47].
2. For the second aspect, data are increasingly regarded as a treasured resource or a new form of capital and asset [31] to enable a new business model, especially in the AI era. Data are the fuel for AI is the foundation for technology-driven innovation, spanning industry sectors from health [48], advertising [49], to e-commerce [50], transportation [51], etc. For example, $14.4 trillion of new value are associated with the 'Internet of Everything' [52]. With the advancement of machine learning technology, which is data-driven, a lot of novel and intelligent ways of doing business and governance will be enabled. Therefore, compared with traditional production factors like workforces and capital, data will play an even more crucial economic role in coming years (or even becomes a currency for citizens to pay for their communication services and security [53].

Data has a variety of attributes, both commodity attributes and asset attributes, the two are not contradictory, but in this article I only discuss the attributes of data as assets. For example, data generated by users on social media platforms (such as posts, comments, likes, etc.) can be considered as commodities because they can be sold by the platform to advertisers for precision marketing [54]. At the same time, this data can generate revenue for the platform. By analyzing user data, the platform can better understand the needs and behaviors of its users, thereby improving its services, attracting more users, and increasing advertising revenue [55]. Therefore, data can be considered as a special asset, which is hard to understand straightforwardly within the conventional economic framework and business logic. The special characteristics of data include:

- **Feature 1**: Consuming data doesn't exhaust it. And since it's not exhaustible, it can be shared by multiple data users, rather than being exclusively utilized.
- **Feature 2**: The marginal cost of reproducing data is essentially zero.
- **Feature 3**: The marginal utility of data is unstable. Data are heterogeneous; each piece of data can have its unique value. Besides, data do not have a fixed value irrespective of what it is used for. The value of data is also augmented in a non-linear manner once aggregated with the external data, which renders the accurate a priori valuation impossible. And since data are typically "consumed" by the models to provide services and products, the expected revenue from data and how we can benefit from external data are difficult to predict before modeling.

These important characteristics of data bring large challenges to build an eco-system for data economy with privacy protection:

1. **Data ownership:** Features 1 and 2 make the ownership of data difficult to protect. Although Recital 7 of GDPR states that "Natural persons should have control of the data", which indicates data ownership should remain with the data subject, once data leaves the original owner, it is difficult to trace its consequential circulation and usage, making the actual ownership hard to protect. It gets more complicated when personal data are fused with other data, not only from the original data owners, such as data from private companies, government agencies, sensor networks, etc. Besides, a significant portion of Internet data are computationally generated and therefore have no "real" empirical source.
2. **Data pricing**: Features 2 and 3 make it challenging to price and trade data. Data value depends on its context and how it is used. Since the value of data can hardly be understood before modeling, the buyer will be reluctant to offer a high price. And because the data can be replicated with no cost, once it is traded, its ownership is not transferred to the new owner, but just enlarges its coverage. The topics regarding data pricing and trading are essential to building a sustainable data economy and are still being delineated or theorized.

Current regulations cannot solve the issues regarding blurred ownership protection. And the data pricing issue reveals the fundamental characteristic of "privacy asymmetry", which is caused by the **capability gap of utilizing the data** (or the capability of "**data commodification**"). The crucial issue is that the data owners like individual users are excluded from the business model based on data assets. Data are taken with little regard for compensation [31]. Although some believe that regulations like GDPR will mandate a complete business model shift (from a data ownership model to a data leasing model) for companies [13], it is not that case: even if companies can protect their users' data privacy from arbitrary and malicious usage, they have not established an approach to measure the contribution of users' data and give the revenue back to the users. Companies might defend themselves like this: Our assets are primarily insights extracted from data, not just the data itself. This weakens the idea of financially compensating users. However, we must realize that without data, there is no data insight. Data is the raw material for generating insights, and without high-quality data, any insight or prediction can be wrong [55]. In addition, the benefits generated by the data themselves are distributed to companies like Google and Facebook, and financial compensation for users does not ignore the contributions of large companies.

Therefore, the real problem is not about transparency, but the power imbalance of utilizing data, which caused data inequality trends (MIT, 2021). Because those companies that have the data and analytics capabilities may reap greater benefits, while those who do not have those resources may be marginalized [56]. And the root of the power imbalance is the commodification of personal information (U.S. Committee on Commerce, Science, and Transportation, 2013). **The absence of structural reciprocities between the data users and data owners** caused incursions actions into legally and socially undefended territory [57]. That is why we need to place data commodification and its fairness at the center of data privacy analysis.

More broadly speaking, the impact of data privacy and the power imbalance of utilizing data is not just in economics. It is also related to the public core values of the society. In the absence of privacy protection, the technology giants, all of whom are heavily investing in and profiting from AI, will dominate not only the public discourse but also the future of the public core values and democratic institutions [58]. This is why voter manipulation by Cambridge Analytica in the 2016 presidential election attracted so much attention. There is a connection from data commodification to a structural understanding of capitalist imperatives. A key insight of commodification theory is that individuals are compelled to act "within a social field whose terms of engagement are primarily set by capital" [59]. Many researchers identified this trend of ubiquitous data utilization and monetarization as a new paradigm called "datafication" [60], and believed it is becoming a leading principle, not just for economics (there are similarities between financialisation and datafication), but also for an opportunity to create a revolutionary **political-economic relationship** between data and

capitalism. A distributed and largely uncontested new expression of power forms, as a ubiquitous networked institutional regime, which a profoundly anti-democratic threat. This new political-economic relationship is referred to by a variety of labels (such as 'surveillance capitalism', Zuboff, 2019), which indicates the enlargement of the social terrain of a new type of capital accumulation. Please note that my views combine a lot of political economy perspectives and are not limited to Zuboff's. From a political economy perspective, the digital economy has not truly surpassed capitalism. In fact, it is merely a new manifestation of capitalism, conforming to the characteristics of capitalist relations of production. Therefore, even in the context of the digital economy, we can still comprehend and predict its development by analyzing the nature and dynamics of the economic system. This requires us to understand more deeply the new form of asset - data, and its role in economic activities.

After understanding the asset properties of the data, it is important to find a counter-acting force, a force from the crowd with the capability of utilizing their own data, to balance such power imbalance in this political-economic relationship. While substantial risks to rights and liberties are emerging, at the same time vast opportunities to create value, promote welfare, and enhance various social objectives are unfolding.

5. Data fairness and centralized paradigm

Data economy is a vast, interconnected set of technologies, companies, marketers, and billions of humans. As a valuable asset to monetize in the data economy, data will take an increasing proportion of revenue created by the whole society [61]. **When the data economy takes a majority share of the economic system, how to fairly share its revenue will become the central issue of social welfare**. We should move beyond current efforts on data privacy, and take it as a significant issue about social wealth distribution and fairness in the data economy. Therefore, I propose to treat **data fairness** as the new target for privacy protection. I define **fairness as a state in which various parties get fair revenue from the data belonging to them**. "Fair revenue" means when a data product (e.g., a machine learning model) creates profit, the contributions of the input elements (including data from individual users) can be fairly estimated, and the profit can be distributed to these elements' owners.

I argue that the **key regarding data fairness is the centralized paradigm of data utilization** (i.e., analysis and modeling for mining value from data), which also is the root issue of data privacy. Below are the reasons:

- **Data ownership**: The current paradigm of data analysis and modeling suffers a high risk of data ownership violation, due to its centralized architecture. Data are extracted and **collected** to a centralized infrastructure. Once the data have been collected, it is hard to protect its ownership and the right to get incentives. And with information asymmetry, given today's methods of accessing and utilizing data, individuals usually cannot directly retrace what happens to their personal information.
- **Long value-added chain and undefined incentive**: In a centralized paradigm, data need to go through a long value-added chain to become a data product or to enable a service. The existence of a centralized paradigm drives a series of actions including data extraction, collection, processing, redistribution, etc. This value-added chain can be very complicated in real cases, with multiple platforms, brokers, and processors applying different manipulations on data. Through this value-added chain, the real value of original data and the contributions of data owners are diminished or ignored. And due to the difficulty of data pricing, users cannot get a fair incentive.

The formation of the current prevalence of the centralized paradigm is natural. The original intention of the Internet is to build an open, free and interconnected space. However, there was always a trend of centralization of the Internet services, due to the cost efficiency, the platform effect, and the technical constraints. I also need to point out that the monopoly position of big tech companies and the centralized paradigm has enhanced each other during the last decade, exacerbating the discussed data unfairness. On one hand, building a centralized platform to support large-scale applications requires high construction and maintenance costs, which is only affordable by big tech companies, and thus enhances their monopoly position, by excluding those small companies. On the other hand, the monopoly position of big tech companies promotes the adoption of the centralized paradigm, because those companies are willing to aggregate or augment different platforms, services, and data together, to deliver more comprehensive products and extract more insights from data.

Therefore, we need to realize that the development of the digital economy is a dynamic process, and the early free digital economy does not contradict the current trend of centralization, they reflect the different stages of development and market dynamics of the digital economy. The resulting Internet nowadays is centralized, with most traffic and profits concentrating on a small group of big tech companies. This does not mean that the free digital economy is dead, but that it is evolving in new forms and conditions. At the same time, it also means that we need to constantly adjust and change to meet new challenges and opportunities.

Although the centralized paradigm has its advantages like cost efficiency, the profit is mainly occupied by big companies, with little incentive for individual users except providing some free Internet services [62]. described this situation as "data colonialism", in which individuals are dispossessed of the data commoditization, while companies become a new means of capitalist with "accumulation by dispossession". Therefore, I argue the real danger in the AI era is not the so-called "the evil AI", but the asymmetric power relationship and the resulting enlargement of the wealth gap, i.e., the gap between the data users (companies along with their entrepreneurs and employees) and massive data owners. Without intervention, current unfairness will be the norm [1]. But we should understand data commodification as a dynamic and contested process. As we are still in the early days of large-scale commoditization of personal data, it is now a crucial time to establish the necessary new norms for acquiring, sharing, and reselling data to ensure data fairness. It is important to propose an alternative approach to the current data economy paradigm, making a counter-force to the current trend. Our answer to this issue will shape the character of information civilization in the future, just as the logic of industrial capitalism and its successors shaped the character of industrial civilization over the last two centuries. There have been some ongoing attempts already, including the "Data Dividend" project[2] and EU's B2C data sharing for the public interest.

6. Achieving data fairness through decentralization

The fundamental solution is to shift from a centralized paradigm to a decentralized paradigm, and this is mainly achieved through the decentralized paradigm. Of course, I must admit that there is also decentralization in the centralized model. There may be an argument that increasing user control and autonomy over data is more important than centralizing or decentralizing data, in other words, there is no need to spend a lot of effort to establish a decentralized paradigm. What I want to explain is that even if there is a certain degree of decentralization in the centralized model, users may not be able to exercise such power, and it is difficult to truly achieve the decentralization in the existing model. It is in the context of the mismatch of capabilities and powers in reality that we propose decentralized paradigm as a

[2] Data Dividend Project: https://www.datadividendproject.com/.

temporary alternative. At the same time, in a decentralized model, the use of data can be recorded through technologies such as Blockchain, which allows users to understand and control their data better [63]. But in the future, increasing public control and autonomy over data will remain the direction of our regulatory efforts.

Data are novel resources, decentralized paradigm can calculate its value and identify its participants. In this part, I discuss the basic logic of this new paradigm and its technical background. The comparison of centralized and decentralized paradigms is illustrated in Fig. 1.

The proposed basic logic of the decentralized paradigm of data utilization is as follows:

1. The data remain where it is collected, therefore its ownership always remains at its owner and will be never transferred to other parties.
2. To utilize the distributed data, modeling is dispatched to participants with data. Therefore, the modeling is decentralized.
3. Each participant builds its local and personalized model. These models can then be federated to build a global model. The federation can be either collaborative (transferring knowledge to other participants), or adversarial (achieving a balance between conflicting goals).
4. The contributions of participants are then estimated and used as the foundation for data pricing. Please note that this kind of pricing is about data usage, instead of data ownership.

Importantly, the proposed logic should be able to "translate" legal requirements into technical solutions (World Wide Web Consortium, 1995). Mapping legal obligations into software functionality are non-trivial [64]. The mismatch and disconnection between the legal and the technological mindsets [65] would make it difficult in practice. Compared with the current approach of "privacy by policy", which leaves the responsibility to comply with the regulation in the hands of legal staff, this paradigm should be possible to build on applicable technical solutions, not just from the legal or management perspective. Besides, it is always stated that it's a trade-off between privacy protection and data utility [66]. However, with the advance of technology, especially the recent development of distributed AI modeling, it is possible to achieve privacy protection and data utilization at the same time. Below I introduce the key technologies we can rely on to build the decentralized paradigm and to create the counter-acting power to confront Surveillance Capitalism:

- **Federated learning**: Distributed modeling can be viewed as a kind of Privacy Enhancing Technology (PET). Historically, the main principles of PETs are data minimization and identity protection by anonymization. Federated learning adopted a distinctive way. To build models based on data sets that are distributed across multiple devices while preventing data leakage, in federated learning, clients collaboratively train a shared model under the orchestration of a

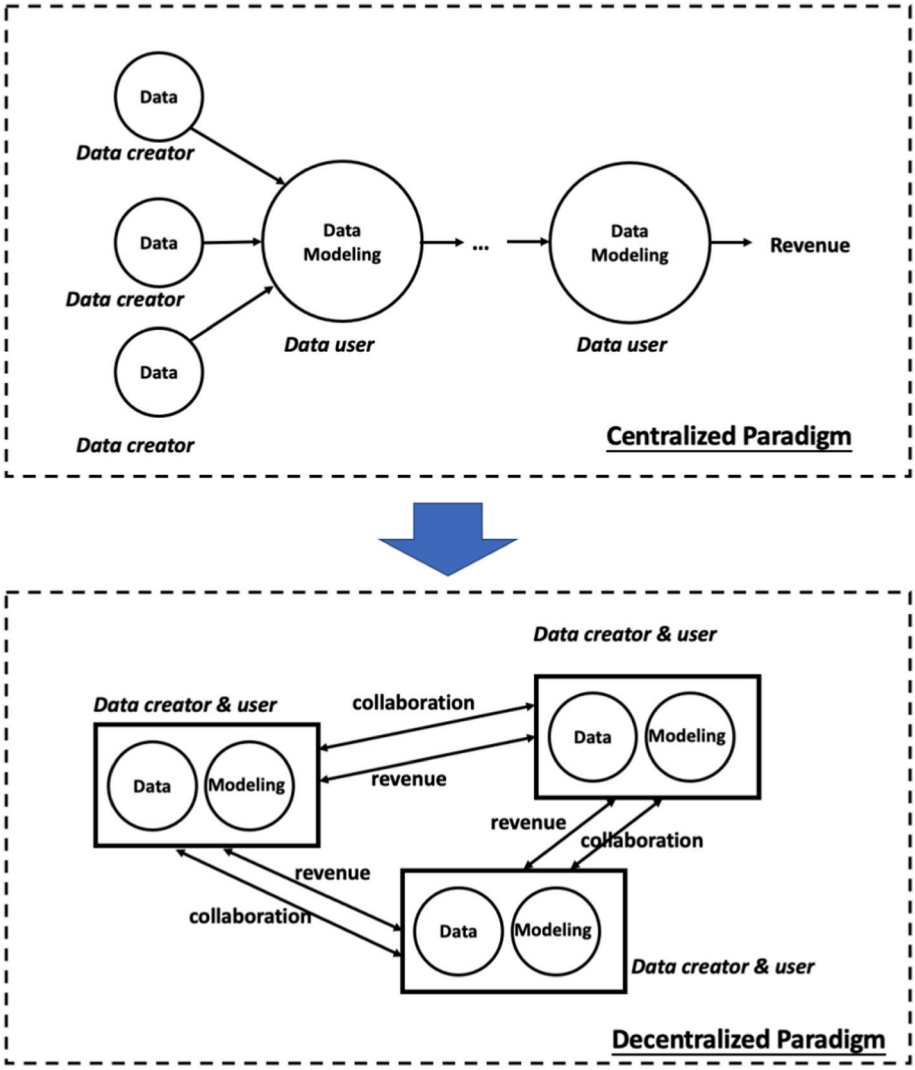

Fig. 1. From centralized to decentralized paradigm.

central server, while keeping the training data decentralized [67]. Since the data are utilized locally (at individuals' devices or within an organization) without sharing with external parties, its ownership and privacy are protected. Due to its advantages in privacy protection, federated learning has got intensive research attention in the past couple of years, and has accumulated many technical solutions, in supporting edge devices [68], reducing communication costs (Sattler Wiedemann et al., 2019), providing personalized services [69], etc. And many real-world applications of federated learning have emerged in healthcare [70], city management [71], etc., to balance the data utilization and data privacy.

- **Secure Multi-Party Computation (MPC)**: Federated learning needs to work in conjunction with security mechanisms to enhance its security level. And in addition to machine learning modeling, there are other types of data analysis and data operations that also need a secure and privacy-preserving approach. The main available options are Secure Multi-Party Computation (also known as privacy-preserving computation) and differential privacy. MPC enables participants to jointly compute a function with their data while keeping data private [72].
- **Differential Privacy (DP):** A related method is differential privacy [73], which is a family of algorithms that can share aggregated information about a statistical database while withholding information about individual participants.
- **Blockchain**: Blockchain [74] provides a robust and transparent protocol to build the network infrastructure in the proposed distributed paradigm. Although as a public distributed ledger, Blockchain has the issues of privacy and anonymity, many recent efforts have been made to overcome these issues (by integrating technologies like MPC [75] and DP as complementarities), making Blockchain a secure tamper-proof ledger to record digital interactions [76]. These interactions not only contain operations on the data but can also record the process of data modeling and analysis, by adopting smart contracts [77]. This ledger is important for ensuring the creditability of the system.

Based on these technologies and their integration, it is possible to build a new decentralized architecture for data utilization. Many attempts have been made with such architecture (e.g., Warnat-Herresthal et al., 2021), trying to **achieve the principle of "Privacy by Design" by "Model-to-Data"**. Data now cannot be used for unrecognized purpose (just like that in Cambridge Analytica's voter manipulation), but can be utilized by specified distributed model, which needs to be agreed between data user and data owner before modeling. In this way, we embed privacy into the entire engineering process, with these technologies effectively involved in the loop [84]. And thus, privacy protection can "come before the fact, not after" [85]. Based on this mode of "model-to-data", we can further aggregate distributed models together. With models personalized according to their local situation, these models can be aggregated to build a more effective global model in many different ways [86].

After understanding the technology, it is necessary to discuss the relationship between P2P and decentralized models. P2P is a computer network architecture that allows multiple computers to connect directly to each other and share files, resources, and services without relying on a centralized server. In a P2P network, each computer is both a client and a server, uploading and downloading data simultaneously. P2P and the decentralized model are not contradictory. In fact, we can use P2P as an architecture at the bottom layer to achieve the decentralization of data and control. Then, we can overlay methods such as federated learning on top of it to achieve data privacy protection and shared learning. This combination enables data management and learning models that are both decentralized and privacy-preserving [87].

Next, to achieve data fairness, we need to find a way for fair data pricing, as the foundation for a fair and sustainable data economy. I argue that it's more convenient to do data pricing with distributed modeling and data usage. Below is the process:

1. Once an aggregated model is used to create utility, its economic benefit is confirmed.
2. The contributions of participants of distributed modeling are then estimated. A popular method is Shapley Values (SV), which takes an average of marginal contributions of all possible coalitions. SV provides a solution to solve the payoff distribution problem for a grand coalition, characterized by a collection of desirable properties (balance, symmetry, zero elements, and additivity [88]). However, SV suffers the challenges like computationally intractability and unfairness in the FL setting. Therefore, many works have proposed to approximate its calculation (Jia et al., 2019) and improve its fairness [89], and it has been utilized in computing the value of data for a user in recent years (Ghorbani and Zou., 2020), according to its data contribution during the modeling.
3. The economic benefit is then divided and paid to the participants according to their SVs.

SVs can cause a computational catastrophe. Therefore, approximations like Monte Carlo-based methods were studied. Alternative to SVs, there are some other data valuation methods like influence functions [90] and reinforcement learning [91]. This kind of data valuation is for data usage, which aligns with the current privacy legislation since the data ownership doesn't trade and is protected. Besides, the valuation is dynamic, in which different modeling tasks lead to different prices. Therefore, it complies with the attribution of unstable data utility. Such data valuation provides a tool for the fair allocation of data benefits for social welfare [92]. A new data eco-system can then potentially be built based on this data utilization architecture. Data owners can protect their privacy and data assets, and get fair revenue. The data eco-system also provides the foundation to build crowd intelligence, in which participants can collaborate through distributed modeling and model aggregation, without previous data collection and transfer problems, and thus is more sustainable.

Fig. 2 summarizes the framework of the whole paradigm.

I need to point out that, to fully build the paradigm and the data eco-system upon it, there are still many open questions to be answered:

1. Non-IID and personalized modeling: While the early efforts in federated learning research focused on overcoming the statistical challenges and improving security. Researchers realized that unbalanced data distribution caused a big challenge for distributed modeling [93], which largely reduces the model accuracy. How to build an accurate while a personalized model with non-IID data distribution is the key real-world application.
2. Model collaboration and aggregation: The essential requirement to establish so-called "crowd intelligence" is how different models can transfer their knowledge to each other, or aggregate to a more effective model when needed.
3. There are still some important technical issues to be solved for data pricing, such as the data combination requirement and computational complexity of calculating SV.

I am also concerned that decentralized data management and benefit distribution models have been tried many times in the past few years, but many attempts have failed to achieve the desired success. For example, Datum is a decentralized data storage network that allows users to sell their data to advertisers and market research firms. However, Datum closed in 2019, mainly due to its business model not being able to attract enough users and data buyers. Many users may be reluctant to spend time and effort managing their data, especially if the benefits they derive from it are relatively small [94]. DataWallet is a similar platform that allows users to control and monetize their data. However, DataWallet closed in 2020 due to its inability to attract enough users and data buyers, as well as its data quality issues.

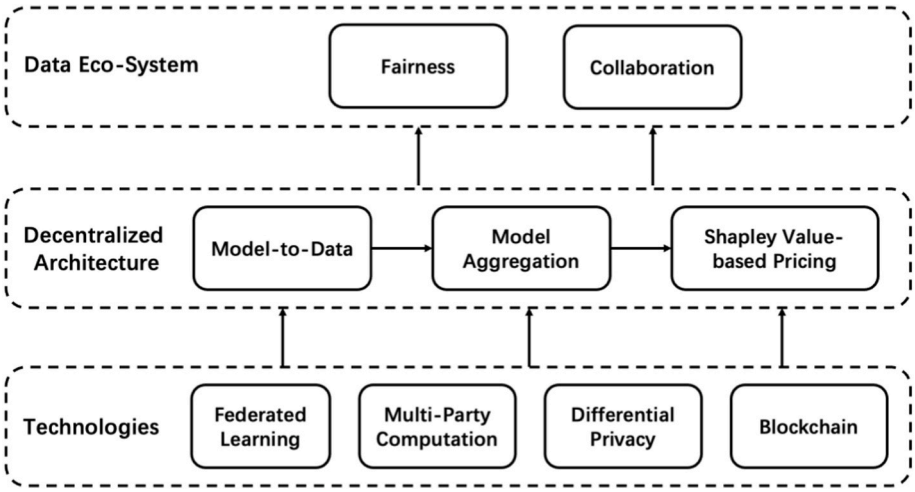

Fig. 2. The framework of the decentralized paradigm.

Decentralized data platforms may not offer the same data quality and accuracy as traditional data brokers [94]. These failure cases and reasons are worth reflecting on and improving.

7. Discussion and conclusion

To summarize, in this work I argue that the real issue of data privacy in the AI era and data economy is its unfair distribution of the revenue extracted from data. How to deal with this issue is the key to social welfare in the future world. Current efforts of transparency regulation cannot solve this issue. It is necessary to establish a new distributed paradigm based on the technologies like federated learning, to find a solution both technically sound and theoretically grounded. However, such a technological solution will have a long way to promote social welfare. Obstacles still exist in distributing the wealth generated by data from individuals. The limited modeling capacity of data owners and the black-box model developed by tech elites cause the actual unfairness in distributed data modeling. And there are counter-forces to this distributed paradigm, including the development of large-scale pre-trained models [95] and model monetization and marketplace. But many developments are supporting the future distributed paradigm, such as personalized model training [86], few-shot learning [96], distributed learning on Blockchain [97], trusted execution environment (TEE) on edge devices [98], decentralized model marketplace [99], etc. With these efforts, it is possible to form a practical solution to support data fairness. And it's essential to link such technical solutions to the runtime mechanism of the real social welfare systems in the future.

It is necessary to mention that under previous technology levels, data was dispersed through decentralization, making the aggregation effect of data difficult to manifest, thus greatly reducing the value of data. But current technology is gradually overcoming this issue, such as the Federated Learning and Secure Multi-party Computation mentioned in this article. These methods can synthesize and utilize data under decentralized conditions to enhance the value of data. Our technology will continue to improve towards this goal in the future.

I also argue that to enable such a paradigm on a large scale, joint efforts from the research communities, governments, companies, and the public are needed, which is "a mammoth task that falls to society as a whole" (MIT, 2021):

Research and technical development: Research and technical development are important to force in the data ecosystem. On one hand, the rapid development of AI promotes data utilization and data economy. On the other hand, such advancements brought great challenges like data privacy. To handle these challenges, separate efforts were made from the different research communities. For example, there is a long history of separation between two research threads. One thread is more about methodology and technical aspects, designing new algorithms, engineering methods, and platforms to achieve better systematic performance; while the other thread cares the data research more from the ethics, management, legal, and other social science aspects. An example of consequence is that in AI ethics research, most efforts are from legal and ethical people. Their research can be hardly integrated into the engineering framework. The ignorance of each other prevents us to build a theoretically grounded and technically sound solution for data privacy protection. As we've seen, recent technical breakthroughs like Blockchain and federated learning provide a good foundation to build a sustainable data ecosystem. To make it benefit society, a lot of joint research needs to be done. And for the problems like data pricing, separated effort from a single discipline is not sufficient. The AI methodology should take the privacy concern during its design. And the research of social science should propose the scientific problems from its perspective, to drive the innovation of methodology and engineering.

Regulation and government: Although, as discussed before, regulations like GDPR still have limitations at the moment, they can bring significant effects if well designed, in promoting public awareness and driving technology development. For the regulations to better play their roles, they need to evolute corresponding to the public requirement and environmental changes. Therefore, I propose that future regulation should take serious consideration of data fairness and appropriately increase the content of the tax law. Data companies, especially multinational tech companies, are often accused of using existing tax code provisions to avoid taxes. The business model and global operation of these companies allow them to set up subsidiaries in countries with low tax rates to reduce the tax burden by shifting profits [100]. In response, some countries, such as France and the United Kingdom, have begun to tax digital services, which complements the traditional corporate income tax system. Digital services tax is a way of taxing the digital services income generated by a company in a certain country or region, rather than taxing the company's profits. We can also learn from this approach. The government also needs to monitor the running effects of the regulations, as the technical development and the market situation can also give reactive counter-acting force to the forming of future regulations. Therefore, it is important to review the real effects of regulations like GDPR, and see how to adapt them to fit the technical development. In addition, other than regulations, the government should also make clear its role in data fairness and the context of data economy, and promote its role in maximizing the overall welfare of the society with a rapidly growing data economy.

Commercial company: The effect of commercial companies in the data ecosystem is subtle. On one hand, due to the regulation and the

consideration of their public images, the companies are active in developing and adopting privacy-preserving products. And through big Internet giants, these products can achieve scalability. However, on the other hand, the inherent business logic of extracting value from data is difficult to be changed for these companies. Therefore, many efforts are superficial, leaving the real issue of data fairness remaining or even deteriorating. That is why although all people are talking about privacy, the revenue from data utilization of those big companies growing constantly. Therefore, it is important to further involve the public in the current practice of data business and push the companies to adapt to the new distributed paradigm.

The public: How do we deal with the privacy issue would have a significant impact on the value of the public and even the value of democracy. The ultimate and decisive force is the public orientation and decision. The awareness and cognition of privacy issues have been enhanced in the past years. However, as discussed, the main focus is now still conventional, with little awareness of the issue of unfairness in the data economy. Therefore, more discussions among different parties should take place, to make the public clearer about the fundamental issues and all available solutions. And the innovation of privacy-preserving business models should be encouraged.

CRediT authorship contribution statement

Chao Wu: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing.

Data availability

Data will be made available on request.

References

[1] D. Bouk, The history and political economy of personal data over the last two Centuries in three acts, Osiris 32 (1) (2017) 85–106.
[2] I.A.T. Hashem, et al., The rise of "big data" on cloud computing: review and open research issues, Inf. Syst. 47 (2015) 98–115.
[3] Olivia Solon, Facebook says Cambridge Analytica may have gained 37m more users' data, Guardian (April 4, 2018). Retrieved April 6, 2018, https://www.theguardian.com/technology/2018/apr/04/facebook-cambridge-analytica-user-data-latest-more-than-thought.
[4] A.R. Brough, K.D. Martin, Consumer privacy during (and after) the COVID-19 pandemic, J. Publ. Pol. Market. 40 (1) (2021) 108–110.
[5] Brian Chen, Apple Announces New Privacy Features (2020). Available at: https://www.nytimes.com/2020/06/23/technology/apple-announces-new-privacy-features.html.
[6] G.A. Fowler, Facebook Will Now Show You Exactly How it Stalks You - Even when You're Not Using Facebook, 2020. Washington Post.
[7] T. Hagendorff, The ethics of AI ethics: an evaluation of guidelines, Minds Mach. 30 (1) (2020) 99–120.
[8] B.C. Fung, K. Wang, R. Chen, et al., Privacy-preserving data publishing: a survey of recent developments, ACM Comput. Surv. 42 (4) (2010) 1–53.
[9] B. Schilit, J. Hong, M. Gruteser, Wireless location privacy protection, Computer 36 (12) (2003) 135–137.
[10] X. Shen, B. Tan, C. Zhai, Privacy protection in personalized search. ACM SIGIR Forum, ACM, New York, NY, USA, 2007, pp. 4–17.
[11] Q. Feng, D. He, S. Zeadally, et al., A survey on privacy protection in blockchain system, J. Netw. Comput. Appl. 126 (2019) 45–58.
[12] F. Dufaux, T. Ebrahimi, Scrambling for privacy protection in video surveillance systems, IEEE Trans. Circ. Syst. Video Technol. 18 (8) (2008) 1168–1174.
[13] K.A. Houser, W.G. Voss, GDPR: the end of Google and Facebook or a new paradigm in data privacy, Rich. JL & Tech. 25 (2018) 1.
[14] B. Schneier, Data and Goliath: the Hidden Battles to Collect Your Data and Control Your World, W.W. Norton & Company, New York, NY, 2016.
[15] M. Chen, S. Mao, Y. Liu, Big data: a survey, Mobile Network. Appl. 19 (2) (2014) 171–209.
[16] M. Hung, Leading the IoT. Gartner Report, Available at: https://www.gartner.com/imagesrv/books/iot/iotEbook_digital.pdf, 2020.
[17] B. Marr, Available at: How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read (2018) https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/.
[18] S. Ackerman, S. Thielman, US intelligence chief: We might use the Internet of Things to spy on you. The Guardian, 9 February (2016). Available at: www.theguardian.com/technology/2016/feb/09/internet-of-things-smarthome-devices-government-surveillance-james-clapper.
[19] R. Clarke, Information technology and dataveillance, Commun. ACM 31 (5) (1988) 498–512.
[20] K. Thomas, C. Grier, D.M. Nicol, unfriendly: multi-party privacy risks in social networks. International Symposium on Privacy Enhancing Technologies Symposium, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 236–252.
[21] A. Mehmood, et al., Protection of big data privacy, IEEE Access 4 (2016) 1821–1834.
[22] C. Warzel, A. Ngu, Google's 4,000-Word Privacy Policy Is a Secret History of the Internet. The New York Times (2019). Available at: https://www.nytimes.com/interactive/2019/07/10/opinion/google-privacy-policy.html.
[23] J. Luo, M. Wu, D. Gopukumar, et al., Big data application in biomedical research and health care: a literature review, Biomed. Inf. Insights 8 (BII) (2016) S31559.
[24] J. Liang, J. Yang, Y. Wu, et al., Big Data Application in Education: Dropout Prediction in Edx MOOCs. 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), IEEE, 2016, pp. 440–443.
[25] R. Arghandeh, Y. Zhou, Big Data Application in Power Systems, Elsevier, 2017.
[26] R. Nieva, Here's how Facebook collects your data when you're logged out (2018). Available at: https://www.cnet.com/news/heres-how-facebook-collects-your-data-when-youre-logged-out/.
[27] G.J.X. Dance, et al., As Facebook Raised a Privacy Wall, It Carved an Opening for Tech Giants. The New York Times (2018). Available at: https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html.
[28] B. Herold, Google under fire for data-mining student email messages – education week, Education Week (2014). Available at: https://www.edweek.org/policy-politics/google-under-fire-for-data-mining-student-email-messages/2014/03.
[29] D.J. Leith, S. Farrell, Contact Tracing App Privacy: what Data Is Shared by Europe's GAEN Contact Tracing Apps, 2020. Testing Apps for COVID-19 Tracing (TACT).
[30] CNIL, Google privacy policy: WP29 proposes a compliance package, Commission Nationale de L'informatique et Des Libertés (2014). Available at: http://www.cnil.fr/english/news-and-events/news/article/google-privacy-policy-wp29-proposes-a-compliance-package/.
[31] J. Sadowski, When Data Is Capital: Datafication, Accumulation, and Extraction, Big Data & Society, 2019, https://doi.org/10.1177/2053951718820549.
[32] R. Clarke, Dataveillance by Governments: the Technique of Computer Matching, Information Technology & People, 1994.
[33] B. Aho, R. Duffield, Beyond surveillance capitalism: privacy, regulation and big data in Europe and China, Econ. Soc. 49 (2) (2020) 187–212.
[34] H. Li, L. Yu, W. He, The impact of GDPR on global technology development, J. Global Inf. Technol. Manag. 22 (2019) 1–6.
[35] A.A. Forni, et al., Organizations Are Unprepared for the 2018 European Data Protection Regulation, Gartner, 2017. May 2017.
[36] M. Nouwens, et al., Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, 2020, pp. 1–13. Honolulu, HI, USA, April 2020.
[37] V. Ayala-Rivera, L. Pasquale, The grace period has ended: an approach to operationalize GDPR requirements, in: 2018 IEEE 26th International Requirements Engineering Conference (RE). Banff, AB, IEEE, Canada, 2018, pp. 136–146, 0-24 Aug 2018.
[38] J. Jia, G.Z. Jin, L. Wagman, The Short-Run Effects of GDPR on Technology Venture Investment, National Bureau of Economic Research, 2018, https://doi.org/10.3386/w25248.
[39] J. Sørensen, S. Kosta, Before and after GDPR: the changes in third party presence at public and private European websites, in: The World Wide Web Conference, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1590–1600. New York.
[40] M. Crain, The limits of transparency: data brokers and commodification, New Media Soc. 20 (1) (2018) 88–104.
[41] A.C. Madrigal, The Atlantic, 1 march. Reading the Privacy Policies You Encounter in a Year Would Take 76 Work Days, 2012. Available at: http://www.theatlantic.com/technology/archive/2012/03/reading-the-privacy-policies-you-encounter-in-a-year-would-take-76-work-days/253851.
[42] W.B. Tesfay, et al., PrivacyGuide: towards implementation of the EU GDPR on Internet privacy policy evaluation, in: Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics. Tempe, Association for Computing Machinery, New York, 2018, pp. 15–21. AZ, USA.
[43] T.Z. Zarsky, Incompatible: the GDPR in the age of big data, Seton Hall Law Rev. 47 (2016) 995.
[44] E. Morozov, What happens when policy is made by corporations? Your privacy is seen as a barrier to economic growth, Guardian (2015), 11 July. https://www.theguardian.com/commentisfree/2015/jul/12/. ttip-your-data-privacy-is-a-barrier-to-economic-growth
[45] Acxiom Corporation, Annual report 2014, 28 May, Little Rock, Arkansas (2014). Available at: http://www.acxiom.com/about-acxiom/investor-information/reports/.
[46] R. Aitken, 'All data is credit data': constituting the unbanked, Compet. Change 21 (4) (2017) 274–300.
[47] E. Brynjolfsson, B. Kahin (Eds.), Understanding the Digital Economy: Data, Tools, and Research, MIT press, 2002.
[48] A.L. Beam, I.S. Kohane, Big data and machine learning in health care, JAMA 319 (13) (2018) 1317–1318.

[49] C. Perlich, et al., Machine learning for targeted display advertising: transfer learning in action, Mach. Learn. 95 (1) (2014) 103–127.
[50] M.T. Ballestar, P. Grau-Carles, J. Sainz, Predicting customer quality in e-commerce social networks: a machine learning approach, Review of Managerial Science 13 (3) (2019) 589–603.
[51] F. Zantalis, et al., A review of machine learning and IoT in smart transportation, Future Internet 11 (4) (2019) 94.
[52] J. Bradley, et al., Embracing the internet of everything to capture your share of $14.4 trillion, White Paper (2014). Available at: http://www.cisco.com/web/about/ac79/docs/innov/IoE_Economy.pdf.
[53] V. Mayer-Schönberger, T. Ramge, Reinventing Capitalism in the Age of Big Data, Basic Books, 2018.
[54] J. Van Dijck, The Culture of Connectivity: A Critical History of Social Media, Oxford University Press, 2013.
[55] F. Provost, T. Fawcett, Data science and its relationship to big data and data-driven decision making, Big Data 1 (1) (2013) 51–59.
[56] V. Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, Martin's Press, St, 2018.
[57] S. Zuboff, The Age of Surveillance Capitalism: the Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019, Profile books, 2019.
[58] A.K. Rai, Machine learning at the patent office: lessons for patents and administrative law, Iowa Law Rev. 104 (2019) 2617.
[59] V. Mosco, The Political Economy of Communication, 2nd, SAGE, 2009.
[60] J. Van Dijck, Datafication, dataism and dataveillance: big data between scientific paradigm and ideology, Surveill. Soc. 12 (2) (2014) 197–208.
[61] M. Poon, Corporate capitalism and the growing power of big data: a review essay, Sci. Technol. Hum. Val. 41 (6) (2016) 1088–1108.
[62] J. Thatcher, D. O'Sullivan, D. Mahmoudi, Data colonialism through accumulation by dispossession: new metaphors for daily data, Environ. Plann. Soc. Space 34 (6) (2016) 990–1006.
[63] S. Manski, B. Manski, No gods, No masters, No coders? The future of sovereignty in a blockchain world, Law Critiq. 29 (2) (2018) 151–162.
[64] M. Colesky, J.H. Hoepman, C. Hillen, Acritical analysis of privacy design strategies, in: Security and Privacy Workshops (SPW), IEEE, San Jose, CA, USA, 2016, pp. 22–26. May 2016.
[65] M. Birnhack, E. Toch, I. Hadar, Privacy mindset, technological mindset, Jurimetrics 55 (1) (2014) 55–114.
[66] L. Xu, et al., Privacy or utility in data collection? A contract theoretic approach, IEEE Journal of Selected Topics in Signal Processing 9 (7) (2015) 1256–1269.
[67] J. Konečný, H.B. McMahan, F.X. Yu, et al., Federated Learning: Strategies for Improving Communication Efficiency, 2016 arXiv preprint arXiv:1610.05492.
[68] S. Wang, T. Tuor, T. Salonidis, K.K. Leung, C. Makaya, T. He, K. Chan, Adaptive federated learning in resource constrained edge computing systems, IEEE J. Sel. Area. Commun. 37 (6) (2019) 1205–1221.
[69] V. Kulkarni, M. Kulkarni, A. Pant, Survey of personalization techniques for federated learning, in: 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), IEEE, 2020, pp. 794–797.
[70] J. Xu, et al., Federated learning for healthcare informatics, Journal of Healthcare Informatics Research 5 (1) (2021) 1–19.
[71] J.C. Jiang, et al., Federated learning in smart city sensing: challenges and opportunities, Sensors 20 (21) (2020) 6230.
[72] O. Goldreich, Secure multi-party computation. Manuscript. Preliminary version 78 (1998).
[73] C. Dwork, Differential Privacy: A Survey of Results. International Conference on Theory and Applications of Models of Computation, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 1–19.
[74] Z. Zheng, et al., Blockchain challenges and opportunities: a survey, Int. J. Web Grid Serv. 14 (4) (2018) 352–375.
[75] H. Zhong, Y. Sang, Y. Zhang, Z. Xi, Secure multi-party computation on blockchain: an overview, in: International Symposium on Parallel Architectures, Algorithms and Programming, Springer, Singapore, 2019, pp. 452–460.
[76] C. Wirth, M. Kolain, Privacy by blockchain design: a blockchain-enabled GDPR-compliant approach for handling personal data, in: Proceedings of 1st ERCIM Blockchain Workshop 2018, European Society for Socially Embedded Technologies (EUSSET), Amsterdam, Netherlands, 2018, pp. 8–9. May 2018 2(6).
[77] S. Wang, et al., An overview of smart contract: architecture, applications, and future trends, in: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2018, pp. 108–133.
[78] A. Hard, et al., Federated Learning for Mobile Keyboard Prediction, 2018 arXiv preprint arXiv:1811.03604.
[79] W. Du, M.J. Atallah, Privacy-Preserving Cooperative Scientific Computations, csfw. Citeseer, 2001, p. 273.
[80] N.H. Phan, et al., Differential privacy preservation for deep auto-encoders: an application of human behavior prediction, Proc. AAAI Conf. Artif. Intell. 30 (1) (2016).
[81] Y. Lindell, E. Omri, A practical application of differential privacy to personalized online advertising, IACR Cryptol. ePrint Arch 2011 (2011) 152.
[82] T. Hukkinen, et al., A Blockchain Application in Energy (No. 71), ETLA Report, 2017.
[83] D. Tse, et al., Blockchain application in food supply information security, in: 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) 1357-1361, IEEE, 2017.
[84] Y.S. Martin, A. Kung, Methods and tools for GDPR compliance through privacy and data protection engineering, in: 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, London, UK, 2018, pp. 108–111, 23-27 April 2018.
[85] A. Cavoukian, Privacy by design: the 7 foundational principles, in: Information and Privacy Commissioner of Ontario, 2011. Canada.
[86] Y. Deng, M.M. Kamani, M. Mahdavi, Adaptive Personalized Federated Learning, 2020 arXiv preprint arXiv:2003.13461.
[87] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: concept and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2) (2019) 1–19.
[88] L.S. Shapley, A value for n-person games, Contributions to the Theory of Games 2 (28) (1953) 307–317.
[89] Z. Fan, H. Fang, Z. Zhou, J. Pei, M.P. Friedlander, C. Liu, Y. Zhang, Improving Fairness for Data Valuation in Federated Learning, 2021 arXiv preprint arXiv:2109.09046.
[90] A. Richardson, A. Filos-Ratsikas, B. Faltings, Rewarding High-Quality Data via Influence Functions, 2019 arXiv preprint arXiv:1908.11598.
[91] J. Yoon, S. Arik, T. Pfister, Data valuation using reinforcement learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 10842–10851.
[92] A. Acquisti, The Economics of Personal Data and the Economics of Privacy, 2010.
[93] X. Li, et al., On the Convergence of Fedavg on Non-iid Data, 2019 arXiv preprint arXiv:1907.02189.
[94] A. Lambrecht, C. Tucker, Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads, Manag. Sci. 65 (7) (2019) 2966–2981.
[95] L. Floridi, M. Chiriatti, GPT-3: its nature, scope, limits, and consequences, Minds Mach. 30 (4) (2020) 681–694.
[96] S. Ravi, H. Larochelle, Optimization as a Model for Few-Shot Learning, 2016.
[97] S. Zhou, H. Huang, W. Chen, P. Zhou, Z. Zheng, S. Guo, Pirate: a blockchain-based secure framework of distributed machine learning in 5g networks, IEEE Network 34 (6) (2020) 84–91.
[98] G. Ayoade, V. Karande, L. Khan, K. Hamlen, Decentralized IoT data management using blockchain and trusted execution environment, in: 2018 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, 2018, pp. 15–22.
[99] J. Weng, J. Weng, C. Cai, H. Huang, C. Wang, Golden grain: building a secure and decentralized model marketplace for MLaaS. IEEE Transactions on Dependable and Secure Computing, 2021.
[100] G. Zucman, Taxing across borders: tracking personal wealth and corporate profits, J. Econ. Perspect. 28 (4) (2014) 121–148.

Dr. Chao Wu is a Professor in School of Public Affairs, Zhejiang University, and he is also the Director of Computational Social Science Research Center of Zhejiang University and Honorary Research Fellow in Department of Computing, Imperial College London. He got his PhD in Computer Science from Zhejiang University and Imperial College London. His main research topics are about big data analysis, federated learning and computational social science.