Selling Personal Information: Data Brokers and the Limits of US Regulation

Denise DiPersio
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
dipersio@ldc.upenn.edu

Abstract

A principal pillar of the US Blueprint for an AI Bill of Rights is data privacy, specifically, that individuals should be protected from abusive practices by data collectors and data aggregators, and that users should have control over how their personal information is collected and used. An area that spotlights the need for such protections is found in the common practices of data brokers who scrape, purchase, process and reassemble personal information in bulk and sell it for a variety of downstream uses. Such activities almost always occur in the absence of users' knowledge or meaningful consent, yet they are legal under US law. This paper examines how data brokers operate, provides some examples of recent US regulatory actions taken against them, summarizes federal efforts to redress data broker practices and concludes that as long as there continues to be no comprehensive federal data protection and privacy scheme, efforts to control such behavior will have only a limited effect. This paper also addresses the limits of informed consent on the use of personal information in language resources and suggests a solution in an holistic approach to data protection and privacy across the data/development life cycle.

Keywords: personal information, privacy, data broker

1. Introduction

A principal pillar of the US Blueprint for an AI Bill of Rights is data privacy, specifically, that individuals should be protected from abusive practices by data collectors and data aggregators, and that users should have control over how their personal information is collected and used. An area that spotlights the need for such protections is found in the common practices of data brokers who scrape, purchase, process and reassemble personal information in bulk and sell it for a variety of downstream uses. Such activities almost always occur in the absence of users' knowledge or meaningful consent, yet they are legal under US law. This paper examines how data brokers operate, provides some examples of recent US regulatory actions taken against them, summarizes federal efforts to redress data broker practices and concludes that as long as there continues to be no comprehensive federal data protection and privacy scheme, efforts to control such behavior will have only a limited effect. This paper also addresses the limits of informed consent around the use of personal information in language resources and suggests a solution in an holistic approach to data protection and privacy across the data/development life cycle.

2. What Is Personal Information?

Acknowledging that there is no legal framework governing the use of 'personal information', the Blueprint for an AI Bill of Rights does not attempt to define the term. It instead focuses on the ways industry and government use individuals' data, particularly in 'senstitve' domains that include health, employment, education, criminal justice and personal finance. Similarly, as shown below, a significant part of the discussion about data brokers refers to their use of 'sensitive' geolocation data.

In US human subjects data collections, researchers refer to 'personally identifiable information' (PII) as something that must be protected. But the Common Rule – the federal regulation governing human subjects research – does not define that term.[1] Some US government agencies have developed their own definitions of PII. The US General Services Administration (GSA), the body responsible for managing federal property and providing contracting options for government agencies, defines PII broadly as 'information that can be used to distnguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual.' [2] This definition also recognizes that PII must be assessed on a case-by-case basis since it is not characterized by a single category or technology. [3]

By contrast, the global standard as expressed in Article 4(1) of the General Data Protection Regulation (GDPR) schemes in the European Union and the United Kingdom expands the notion of 'personal data' to factors including a person's economic, cultural, social and physical identity.

[1] The Common Rule speaks in terms of *private information* and *identifiable private information*, both of which refer to non-public data or behavior. 19 CFR 46.102(e), (4), (5).

[2] GSA Rules of Behavior for Handling Personally Identifiable Information (PII), https://www.gsa.gov/directives-library/gsa-rules-of-behavior-for-handling-personally-identifiable-information-pii-2#.

[3] The author is familiar with a case where an earlier version of the GSA definition was adopted by another agency and applied to a language resource data collection conducted by the Linguistic Data Consortium.

This paper means to refer to personal information in the broadest sense, ranging from information provided by individuals in the course of their normal interactions with web platforms and applications (including social media and mobile phone use), to linguistic data collected under research protocols and accessible in published language resources. This includes personal data and sensitive personal data as described in the Blueprint for an AI Bill of Rights, private information as referenced in the Common Rule, PII such as defined by the GSA and personal data within the meaning of the GDPR. Effort will be made to distringuish among these descriptions below.

3. The Data Broker Ecosystem

Anyone participating in the digital world leaves a personal information footprint: from their browser history, website visits and related activities like credit card transactions, to their messaging content and behavior, and beyond. As part of those transactions and interactions, the companies and platforms accessed by users repurpose the information left behind to further monetize it, usually without user knowledge or consent. This can occur in several ways, including through third-party apps containing the data broker's software development kit (SDK), in the broker's own mobile apps, and from information the broker purchased from other brokers and data aggregators. Results range from targeted marketing recommendations to providing the information in bulk to third party data brokers who distribute, combine, process and resell it downstream in various forms. The most pernicious of these are data sets that contain geolocation data which either in its original form, or when manipulated and combined with other data, reveals personal information about a host of habits, including entertainment choices, travel history, and visits to sensitive locations (hospitals, reproductive clinics, places of worship), the latter of which can result in threatening behavior toward identified individuals.

3.1 The Current Regulatory Landscape

In the United States, it is legal to buy data through data brokers. And because there is no general US data protection and privacy law, there is no federal mechanism to scrutinize the privacy implications in data broker transactions. Nevertheless, one can identify some key areas where those transactions violate general privacy principles.

The US regulations governing human subjects research provide that informed consent must be obtained prior to using personally identifiable information. However, interacting with web platforms and related applications is not typically considered to constitute human subjects research. Thus, most platforms and apps are not required to, and do not, provide for meaningful consent in the first instance, nor do they disclose that they have the right to sell user data to unidentified third parties for unknown purposes. Even if they did, those terms are typically buried in a long document to which the user clicks consent. Research shows that the majority of users will not read these documents.[4] Similarly, data broker claims that users have "opted-in" to the ecosystem because they share their information on an app fails as well because as indicated above, users have not been meaningfully informed about the downstream uses of their data. Finally, even if a single data set does not disclose individual information, that information can often be easily reidentified when it is combined with other data (Gebhart & Richman, 2023). Research has shown that reidentification is possible from only a few data points (Sweeney 2000; de Montjoye, et al., 2015).

Buying and selling personal information under the EU GDPR and the UK GDPR is covered under the rules for processing personal data. This means that there must be a legal basis to process personal data, that the data can be used only for the purpose for which it was collected, that the purpose is disclosed, that consent is obtained, and that consent can be withdrawn at any time.[5] These rules apply to data brokers even though they may not have collected the personal data originally if the use by the data broker is different from the use for which consent was obtained.

To the extent that US data platforms and data brokers collect, purchase or sell information from EU citizens that would otherwise be subject to GDPR requirements, it seems clear that their practices do not meet GDPR personal data processing standards.

3.2 Problematic Practices and Efforts to Redress Them

Data brokers generally promote themselves as agents of information that operate for good. They boast that their resources can boost business marketing campaign effectiveness, assist academic researchers searching for equitable solutions to a multitude of problems and help the government manage public crises. And they assert that they accomplish those goals while properly protecting individual privacy rights.

For example, **Veraset** claims billons of data points and location date from over 150 countries 'trusted' by more than 100 data scientists.[6] **Cubebiq** touts its

[4] Cameron Dell, The Sad Truth of the FTC's 'Historic' Privacy Win, https://www.wired.com/story/ftc-xmode-outlogic-location-data-settlement/ (citing research that a person needs around 76 working days to review the privacy policies they interact with in one year).

[5] Ivan Lyaskivskij, GDPR requirements to selling of personal data', https://legalitgroup.com/en/gdpr-requirements-to-selling-of-personal-data-ccpa-vs-gdpr-on-insurance-and-trade/; see also Information Commissioner's Office, What common issues might come up in practice?, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/the-right-to-be-informed/what-common-issues-might-come-up-in-practice/ (construing UK GDPR).

[6] Veraset, https://www.veraset.com/about/data-industry.

work with Oxford University and the US Center for Disease Control to provide population 'mobility insights' during the COVID-19 pandemic. [7] **Kochava** promises that anything can be measured: 'Any Channel, Any Device, Any Audience'. [8]

However, as indicated above, the principal way data brokers attempt to defuse privacy-related criticisms is to invoke the notion of a user 'opt-in'. This practice has been the focus of recent disclosures and US regulatory actions against data brokers.

SafeGraph (2022). It was discovered that this firm purchased data from the Life360 app – designed to connect family location information – that included location data for US Planned Parenthood clinics -- which it in turn offered for sale. The company removed its family planning center data in response to protests. (Gebhart & Richman, 2023). This disclosure also resulted in a 2023 class action lawsuit against Life360 claiming that users' location data was sold without permission.[9]

Kochava (2024). The US Federal Trade Commission (FTC), an agency that regulates unfair trade practices, filed a lawsuit against this firm in 2022 for selling geolocation data from mobile devices tracing individual movements to and from sensitive locations. The court dismissed the complaint, but allowed the FTC to amend it with specific examples of consumer harm. In February 2024, the court denied a motion to dismiss the amended complaint which means that the case will proceed. The FTC seeks an injunction to stop Kochava from selling sensitive data without user consent. [10]

X-Mode/Outlogic (2024). The FTC settled its complaint against this firm in January 2024 by entering into a consent order under which, among other things, Outlogic will be prohibited from sharing or selling any sensitive location data; it must also destroy all non-deidentified, sensitive location data previously collected. The company must establish clear and simple user procedures for withdrawing consent, for obtaining the identity of organizations who bought their data, and for removing their data from the company database and recipients' databases. Finally, no recipients of Outlogic's data sets must be able to associate the data with locations relating to LGBTQ+ services, locations of political or social demonstrations or protests, or the location of a specific individual. [11]

3.3 Connections to Research, Law Enforcement and Government

In addition to their commercial customers, data brokers sell their data sets and related resources (tools, APIs) to academic institutions, law enforcement organizations and government agencies. These transactions support the corporate message that data broker services benefit society. But is that the case?

Broker data sets are typically described in journal publications about academic research as data that is 'anonymized' or 'privacy-compliant', which is not always true (Gebhart & Richman, 2023). Those descriptions are perpetuated in open research data sharing mechanisms where safe use is assumed. It also raises the question, posed by Gebhard & Richman, whether this makes researchers 'accomplices' to the practices of data brokers (Ibid., 2023).

In 2023, the US Office of the Director of National Intelligence issued a report showing that agencies including the Federal Bureau of Investigation, the Internal Revenue Service, and the Department of Homeland Security, among others, purchased databases from data brokers, thus avoiding the need to obtain a warrant, a court order, or a subpoena (Ayoub & Goitein, 2024). The US Constitution's Fourth Amendment requires the government to obtain a warrant to access material in which individuals have a reasonable expectation of privacy. Some agencies claim that the Fourth Amendment does not apply to data sold to the government. (Ibid.). This practice seems likely to continue for the time being. Pending legislation would ban government purchases of communications data only. In addition, the Blueprint for an AI Bill of Rights exempts from its coverage government agencies engaged in national security and law enforcement activities.

Data brokers also sell personal information to customers outside the United States. No law regulates or prevents those transactions, notwithstanding the risk that such data can be used against US interests.

3.4 What Americans Think About How Their Personal Information Is Collected and Shared

Americans are becoming increasingly concerned about the privacy of their personal information. In a 2019 study by the Pew Research Center, a

[7] Cubeiq's Data for Good Program: Where We've Been and Where We're Going, https://www.cuebiq.com/resource-center/resources/cuebiqs-data-for-good-program-where-weve-been/.

[8] Kochava, https://www.kochava.com/.

[9] Jon Keegan, Life360 Sued for Selling Location Data, The Markup, https://themarkup.org/privacy/2023/06/01/life360-sued-for-selling-location-data.

[10] Ashley Belanger, Data broker allegedly selling de-anonymized info to face FTC lawsuit after all, https://arstechnica.com/tech-policy/2024/02/data-broker-selling-de-anonymized-info-to-face-ftc-lawsuit-after-all/.

[11] Federal Trade Commission, FTC Order Prohibits Data Broker X-Mode Social and Outlogic from Selling Sensitive Location Data, https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-order-prohibits-data-broker-x-mode-social-outlogic-selling-sensitive-location-data.

nonpartisan organization that conducts opinion polling and other research, over 80% of respondents indicated that they did not have control over data collected about them by companies or the US government and that the risks associated with company-collected data outweighed the benefits. (Auxier & Rainie, 2019). Most did not understand how companies (59%) or the government (78%) used data collected from them.[12] Respondents generally preferred more government regulation but were resigned to the idea that their online activity is being tracked and their personal data collected. (Ibid.).

A 2023 Pew study focused on data privacy revealed that Americans had grown more pessimistic about how their personal information is used. Over 70% had growing concerns over how the government uses the personal data it collects, and they do not trust companies to use their data responsibly. (Faverio, 2023). Even when they make the right decisions to protect their personal information, most believed that their actions do not make a difference in the way companies or social media executives protect their privacy. (Ibid.). As in 2019, the majority of respondents support more government regulation of how personal data is used. (Ibid.; Auxier & Rainie, 2019).

4. Recent Regulatory Developments

4.1 Executive Branch Actions

In the gap left by the lack of a comprehensive law addressing threats to data protection and personal privacy in the digital space, as well as international pressure, the US Executive Branch has taken steps to set down principles and rules designed to address the threat to individuals from the growing scope of AI-powered technologies and systems.

In 2022, the White House Office of Science and Technology Policy issued a Blueprint for an AI Bill of Rights based on five principles: safe and effective systems, algorithmic discrimination procedures, data privacy, notice and explanation, and human alternatives, consideration, and fallback. The data privacy pillar acknowledges the abuses in the way personal data is collected and used by stipulating that data should be collected and used for a particular, stated purpose and context, that consent to collect and use that data should be obtained and the conditions of consent respected, and that any data used in 'sensitive domains' should be subject to further review and potential restraint. Despite the violations of legal process committed by various government agencies using personal data from brokers, the AI Bill of Rights exempts from its coverage government agencies involved in law enforcement and national security.

Building on the AI Bill of Rights, President Biden issued an Executive Order in 2023 on safe, secure, and trustworthy artificial intelligence. The order urges Congress to pass data privacy legislation and identifies actions to enhance privacy protection, such as developing technologies for that purpose, reviewing data collection practices, and establishing federal guidelines. The Order requires technology companies to share with the government the safety testing results of their AI models, a move that has been criticized as stifling innovation and raising the specter of government misuse of such information. Critics also claim that the order usurps legislative authority in the way it outlines a broad, multi-agency effort without prior enabling legislation.

The Executive Branch took a step toward addressing data broker transactions in 2024 in an Executive Order that curtails data brokers' ability to sell sensitive information to non-US customers in, or vendors selling data in, 'countries of concern' (China, Russia, North Korea, Iran, Cuba and Venezuela).[13] This will be accomplished by regulations developed by the US Department of Justice. The order was meant to address in part the disclosures about US government data purchases although it does not prevent the government from purchasing or using such data, nor does it stop data broker sales to non-covered countries.

4.2 Relevant Pending Legislation

Among the many pending legislative bills relating to data protection, privacy, and artificial intelligence, among other things, there are two initiatives with some relationship to data broker activity. The first attempts to prohibit the US government from purchasing communications-related data from brokers. The second is designed to protect consumer privacy by broadly defining personal information. To date Congress has taken no significant action on either.

The **Fourth Amendment is Not for Sale Act (H.R. 4639)** was originally introduced in 2021 and reintroduced in 2023. A response in part to the US Supreme Court's 2018 decision in *Carpenter v. United States*[14] which held that a warrant is needed to obtain an individual's cell phone data, it bars the US government from purchasing communications information, including location data, from third parties that collect or process that information as well as any

[12] 77% of the study respondents had heard 'at least a little bit' about ad targeting (Auxier & Rainie, 2019). The study did not specify data brokers as among "company" data collectors. Based on the information presented in this paper, it can be argued that most respondents would have been unaware or only vaguely aware of the existence of data brokers and the potential downstream uses of their data beyond ad targeting.

[13] The White House, Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern, https://www.whitehouse.gov/briefing-room/presidential-actions/2024/02/28/executive-order-on-preventing-access-to-americans-bulk-sensitive-personal-data-and-united-states-government-related-data-by-countries-of-concern/ .

[14] *Carpenter v. United States*, 585 U.S. ___, 138 S.Ct. 2206 (2018).

such information collected by deceptive means through unauthorized access to a device or online account.

The **American Data Privacy and Protection Act (H.R. 8152)** was introduced in 2022 and is meant to provide comprehensive protection for consumer privacy. Personal information is broadly defined to include anything that identifies, is linkable or 'reasonably' linkable' to an individual. Additional protections are extended to sensitive information. Entities covered under the bill do not include the US government, however.

5. Language Resources, Personal Information and Privacy

Personal information is a key component of language resources that support machine learning and natural language processing tasks. Handling personal information during the data collection, processing and data sharing phases is subject to various laws, regulations and ethical best practices. The language resource and evaluation community has largely respected the need to obtain informed consent and to protect personally identifiable information in human subjects collections. This is in contrast to the mostly unrestrained behavior exhibited by data brokers. US and European regulations and their limits are briefly reviewed below, followed by a discussion about ways in which the field is adopting a more holistic approach to data protection and privacy that shows promise.

5.1 Limits of Informed Consent

Informed consent is the linchpin for collecting data from humans for research in the United States. This means that a person must be given sufficient information about the study and about how their data or information will be collected, used and shared. If they agree to participate, they signify their consent, typically in writing or electronically. Similarly, the EU and UK GDPR schemes require consent that describes, among other things, the specific purpose and lawful basis for the collection.

The community typically preserves individual privacy under human subjects research regulations by assigning random identifiers to participants which are stored with the research data; participants' personal information (e.g., their name), is stored separately. The data may also be anonymized or otherwise de-identified after collection and before the material is broadly shared. This is the usual standard for published language resources containing data obtained from human subjects.[15]

However, human subjects regulations do not adequately address the normal interactions between humans and the digital world since those are not typically considered to constitute human subjects research (at least under US law). Even when consent is provided for in click-through terms and conditions, users typically cannot easily find it, nor are the terms clearly explained. In other words, 'consenting' under these circumstances does not rise to the level of informed consent.

Activities such as uploading content to public websites (text, audio, image, video) can implicate personal information in a number of ways, either directly, by containing traditional identifiers like name and other contact information, or indirectly, by containing biometric information, for example. Such multimodal data is highly desired for machine learning and natural language processing applications. Most sites containing such information require that users obtain the consent of the individual uploaders to copy, process and/or share such material. But this is a requirement that is honored more in breach than observance.[16]

Sharing language resources that have not been subject to a legal, ethics and/or privacy review and that are not properly documented in that respect can lead to the continued reuse of problematic data and/or the models and systems developed from it. This is not a trivial concern given the vast number of options for data sharing, many of which provide little or no oversight with respect to the resources posted there.

5.2 An Evolving Holistic Approach to Data Protection and Privacy

As society has become increasingly aware of the ways in which individual personal information can be used without their knowledge, the idea of a broad notion of privacy, separate from copyright protection, is emerging. It encompasses all of the types of data collected about individuals in the digital space and the potential ways in which that data can be used, processed and shared, including problematic downstream effects on algorithm development and system performance. The Blueprint for an AI Bill of Rights endorses a comprehensive approach to data protection and privacy along these lines. This approach is also consistent with provisions in the EU and UK GDPR as well as in various national data protection and privacy laws.

Gaining traction in the community is the thought that data protection, ethics and privacy should be considered and re-considered at all stages of the data/development life cycle: from the research plan, through data collection, milestones, testing, and deployment. This is not a new idea. The concept of

[15] Biometric data, such as a person's voice or image, can also be considered an identifier, or the informed consent may limit the way in which that information can be shared. Thus, published resources may include masked speech, blurred faces, or data to which other methods have been applied to protect privacy. The details about such methods and their efficacy are beyond the scope of this paper.

[16] This attitude is bolstered in large part by the prevailing view of US courts that using certain web data for a machine learning use case constitutes a fair use under the exception to US copyright law. (DiPersio, 2018).

'Privacy by Design' originated in the 1970s and has garnered renewed attention since the late 1990s in the US and EU (particularly post-GDPR). (Kamocki & Witt, 2020). Attempts have been made to articulate what a privacy-designed project looks like. One example is an 'ethos life cycle' showing six stages of a data science workflow – problem identification, data discovery, exploratory data analysis, modeling, interpretation and conclusions, and communication. (Boenig-Liptsin, et al., 2022). In another example, potential sources of bias are identified across the cycle; they include historical, representation, measurement, aggregation, learning, evaluation, and deployment biases. (Suresh & Guttag, 2021). A third example focused on personal data describes a software tool that allows individuals to control their data during a study and choose the data they contribute to researchers. (Clos, et al., 2022). In all of these instances, researchers continue to be involved in thinking about data protection and privacy through the entire data collection, development, data sharing and deployment process.

6. Future Outlook

As long as there continues to be no comprehensive US data protection and privacy regulatory scheme, the pattern of piecemeal enforcement seen to date will persist. The FTC, an agency with limited powers to regulate unfair trade practices, has carried the principal burden of protecting individual privacy. This is seen most recently in the actions against data brokers Kochava and X-Mode/Outlogic. Many view the consent order against the latter a significant achievement. Yet, some think that the penalty should have been more severe and ultimately will not change data broker behavior.

The steps taken by the Executive to articulate AI's collective harms (and benefits) are encouraging and to a large extent, they reflect the concerns of most Americans as the Pew studies demonstrate. One can surmise that such activity was motivated in part by a desire to appear in step with the rest of the world. Indeed the 2023 Executive Order was issued just one month before the UK AI Safety Summit. Another goal was likely to highlight the US Congress' failure to act. Overall, however, these Executive actions will have a limited effect.

The outlook for Congressional action on the pending bills discussed above is bleak. Political differences have made it difficult to enact even the most non-controversial measures. Those differences will be exacerbated in 2024, an election year. Moreover, the strong technology lobby has consistently opposed regulation despite their public statements acknowledging the need for increased data protection and transparency.

This is not good news for the many Americans that support the enactment of laws protecting their personal data and their privacy. Companies face uncertainties even as they continue to develop AI applications. Some have created internal processes for using data, taking into account existing state and federal laws, best practices, and assumptions about future regulation based on the current discourse. At a time when the EU is moving forward with the AI Act, which will bring needed transparency to the development and use of artificial intelligence, the United States remains a passive observer.

7. Conclusion

This paper examines the role of data brokers in collecting, aggregating and selling personal information, usually without users' knowledge and consent and attempts to demonstrate the need for effective measures regulating data broker behavior. The recent Executive Branch actions and orders around AI and data brokers' non-US transactions are encouraging, but their effect is limited. This paper also addresses the limits of informed consent on the use of personal information in language resources and suggests a solution in an holistic approach to data protection and privacy across the data/development lifecycle.

8. Ethics Statement

This paper describes ethical issues implicated by the practices of data brokers and data aggregators as well as general ethical issues around data protection and privacy. It does not present any output, formula or suggestion that can be implemented in an unethical manner.

9. Bibliographical References

Auxier, B. and Rainie, L. (2019). Key takeaways on Americans' views about privacy, surveillance and data-sharing, https://www.pewresearch.org/short-reads/2019/11/15/key-takeaways-on-americans-views-about-privacy-surveillance-and-data-sharing/, accessed 1 March 2024.

Ayoub, E., and Goitein, E. (2024). Closing the Data Broker Loophole, https://www.brennancenter.org/our-work/research-reports/closing-data-broker-loophole, accessed 28 February 2024.

Belanger, A. (2024). Data broker allegedly selling de-anonymized info to face FTC lawsuit after all, https://arstechnica.com/tech-policy/2024/02/data-broker-selling-de-anonymized-info-to-face-ftc-lawsuit-after-all/, accessed 28 February 2024.

Boenig-Liptsin, M., Tanweer, A. and Edmundosn, A. (2022) Data Science Ethos Lifecycle: Interplay of Ethical Thinking and Data Science Practice. *Journal of Statistics and Data Science Education*, 30(3) : 228-240.

Bousquette, I. (2024). AI Is Moving Faster Than Attempts to Regulate It. Here's How Companies Are Coping., https://www.wsj.com/articles/ai-is-moving-faster-than-attempts-to-regulate-it-heres-how-companies-are-coping-7cfd7104 accessed 28 March 2024.

Carpenter v. United States, 585 U.S. ___, 138 S.Ct. 2206 (2018).

Clos, J., McClaughlin, E., Barnard, P., Nichele, E., Knight, D., McAuley, D. and Adolphs, S. (2022).

PriPA : A Tool for Privacy-Preserving Analytics of Linguistic Data. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 73–78, Marseille, France. European Language Resources Association.

Cubeiq. (2024). Cubeiq's Data for Good Program: Where We've Been and Where We're Going, https://www.cuebiq.com/resource-center/resources/cuebiqs-data-for-good-program-where-weve-been/, accessed 1 March 2024.

Dell, C. (2024). The Sad Truth of the FTC's 'Historic' Privacy Win, https://www.wired.com/story/ftc-xmode-outlogic-location-data-settlement/, accessed 29 February 2024.

deMontjoye, Y., Radaelli, L., Singh, V.K. and Pentland, A. R. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347 (6221): 536-539.

Department of Health and Human Services. (2019). Protection of Human Subjects. 45 CFR Part 46.

DiPersio, D. (2018). A US Perspective on Selected Legal and Ethical Issues Affecting the Development of Language Resources and Related Technology. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18), W21, Legal Issues and Ethics*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

DiPersio, D. (2022). Data Protection, Privacy and US Regulation. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 9-16, Marseille, France. European Language Resources Association.

EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Faverio, M. (2023). Key findings about Americans and data privacy, https://www.pewresearch.org/short-reads/2023/10/18/key-findings-about-americans-and-data-privacy/, accessed 21 March 2024.

Federal Trade Commission. (2024). FTC Order Prohibits Data Broker X-Mode Social and Outlogic from Selling Sensitive Location Data, https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-order-prohibits-data-broker-x-mode-social-outlogic-selling-sensitive-location-data, accessed 28 February 2024.

Gebhart, G. and Richman, J. (2023). Science Shouldn't Give Data Brokers Cover for Stealing Your Privacy, *Scientific American*, https://www.scientificamerican.com/article/science-shouldnt-give-data-brokers-cover-for-stealing-your-privacy/ accessed 28 February 2024.

General Data Protection Regulation ((EU) 2016/679) (EU GDPR) as it forms part of the law of England and Wales, Scotland and Northern Ireland by virtue of section 3 of the European Union (Withdrawal) Act 2018 and as amended by Schedule 1 to the Data Protection, Privacy and Electronic Communications (Amendments etc) (EU Exit) Regulations 2019 (SI 2019/419).

General Services Administration. (2019). GSA Rules of Behavior for Handling Personally Identifiable Information (PII), https://www.gsa.gov/directives-library/gsa-rules-of-behavior-for-handling-personally-identifiable-information-pii-2# accessed 29 March 2024.

Information Commissioner's Office, What common issues might come up in practice?, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/the-right-to-be-informed/what-common-issues-might-come-up-in-practice/ accessed 28 February 2024.

Kamocki, P. and Witt, A. (2020). Privacy by Design and Language Resources. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3423–3427, Marseille, France. European Language Resources Association.

Keegan, J. (2023). Life360 Sued for Selling Location Data, https://themarkup.org/privacy/2023/06/01/life360-sued-for-selling-location-data, accessed 1 March 2024.

Kochava. (2024). Empowering Marketers and Publishers, https://www.kochava.com/, accessed 1 March 2024.

Lyaskivskij, I. (2024). GDPR requirements to selling of personal data, https://legalitgroup.com/en/gdpr-requirements-to-selling-of-personal-data-ccpa-vs-gdpr-on-insurance-and-trade/, accessed 28 February 2024.

Suresh, H. and Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of EEAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization (EEAMO '21).* ADM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3465416.3483305.

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Pittsburgh, Pennsylvania: Carnegie Mellon University, Data Privacy Working Paper3.

The White House. (2022). Blueprint for an AI Bill of Rights, https://www.whitehouse.gov/ostp/ai-bill-of-rights/, accessed 9 February 2024.

The White House. (2024). Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern, https://www.whitehouse.gov/briefing-room/presidential-actions/2024/02/28/executive-order-on-preventing-access-to-americans-bulk-sensitive-personal-data-and-united-states-government-related-data-by-countries-of-concern/ accessed 2 March 2024.

Thomson Reuters. (2023). How President Biden's executive order on AI impacts the legal sector, https://legal.thomsonreuters.com/blog/how-president-bidens-executive-order-on-ai-impacts-the-legal-sector/, accessed 20 February 2024.
Timelex. (2020). What Should One Keep In Mind When Selling, Purchasing, Or Licensing Personal Data, https://www.timelex.eu/en/blog/what-keep-in-mind-selling-purchasing-licensing-personal-data accessed 28 February 2024.
Veraset. (2024). Your trusted partner for location data, https://www.veraset.com/about/data-industry, accessed 1 March 2024.