The Indian Journal for Research in Law and Management

Open Access Law Journal – Copyright © 2026
Editor-in-Chief – Dr. Muktai Deb Chavan; Publisher – Alden Vas; ISSN: 2583-9896

PRIVACY POLICIES PUT TO TEST: EVALUATING LEGITIMACY OF WEB SCRAPING BY THE BIG TECH

~ Kanupriya Saini & Saumya Sirohi

ABSTRACT

Web scraping refers to the collection of personal data and publicly available data on social media platforms for purposes like training, researching and testing products and services. Web scraping can be both legitimate when working in compliance with data protection safeguards and unethical when such data is breached for misuse of personal identifiable information. The paper analyses the web scraping policies of four major social media platforms, X, Meta, Reddit and LinkedIn to understand the extent of web scraping through third party applications and advertisers, the amount of data scraped, the purposes it is used for and the duration of such retention. The practice of web-scraping is prevalent among these intermediaries to further develop their platforms. These datasets range from private conversations and saved drafts to browser activities. Further, the paper compares the Data Protection Safeguards of four countries, India, California, the European Union and the Peoples' Republic of China to further study the efficiency of such safeguards in regulating the policies of Social Media Intermediaries within their jurisdictions. The paper delves into the differences in application of such policies of the Big Tech and the provisions regarding right to privacy, erasure and right to be forgotten. It also provides an insight into the web scraping practices of the social media intermediaries to train their Artificial Intelligence Platforms through collection and retention of users' data, which can be both personal and identifiable. The methodology is empirical analysis of specific policies in varied jurisdictions. The paper aims to answer whether the Digital Personal Data Protection Act, 2023 is compatible with international frameworks that address data scraping. The legitimacy of these practices varies according to different safeguards, and stricter protocols like the GDPR and CCPA ensure better protection of the user data.

Keywords: ***Web scraping, Data Protection Safeguards, Artificial Intelligence.***

INTRODUCTION

Data scraping is the use of data publicly available in large scale for purposes such as training AI models, developing programmes, research and development purposes or unethical purposes like frauds, misuse and mishandling of such data. Although the term 'data scraping' has not been legally defined, the term has become prominent among privacy analysts, computer scientists and statutory paradigms in the present age. This data is mined using third party interactions on major platforms like social media intermediaries, websites, datasets and so on. Data which is private to an individual is equally public for a social media intermediary and therefore, the collected data ranges from algorithms, surveys, posts and private conversations. The question arises when the limit to such scraping is concerned and when consent becomes a cue for such data scraping. The privacy policies of major social media platforms such as Facebook, Instagram, X, LinkedIn, and Reddit reflect broadly a similar model of data collection, enhancing user experience and limited user control. The platforms collect personal data of users such as their names, e-mail domain names, photos, location, behavioural pattern, and user-generated content to provide personalized services, enhance user experience and improve algorithm. The privacy policies of these platforms are strictly tied to anti- scraping regulations and stringent measure in case of non-compliance. These platforms use paid APIs, CAPTCHAS, Robots. Txt file, Rate Limiting, blocking unknown bots to preserve the publicly available data of users.[1]Yet, these data scraping policies are often breached by third party applications, and the personal data of users is compromised at a cautionary scale. Such cases have surfaced in June 2025 when 4 billion datasets were exposed unethically. These included chats and personal payment details from users in China.[2] Further, in June 2021, 7000 million datasets of LinkedIn users were exposed on dark web.[3] Such happenings led to the establishment of data privacy frameworks with strict monitoring protocols and penalties for the prevention of data breaches related to personal data. However, the statutory safeguards have become stricter and more efficient at the international jurisdictions, they primarily concern the data when it is private, sensitive and non-consensual. Laws such as the GDPR, EU AI Act, Personal Information Protection Law are related to protection of personal data and regulate its

[1] Nanqin Ying, 'The Evolving Landscape of Web Scraping on Social Media Platforms' (*D-Lab*, 1 March 2025https://dlab.berkeley.edu/news/evolving-landscape-web-scraping-social-media-platforms > last accessed 20 April 2026

[2] John Leyden, Dan Swinhoe and Michael Hill, 'The 20 Biggest Data Breaches of the 21st Century' (*CSO Online, 12 June 2025*)<https://www.csoonline.com/article/534628/the-biggest-data-breaches-of-the-21st-century.html>accessed 20 April 2026

[3] Ibid

processing for ethical and unethical use. However, in India, such statutory regulations are either obsolete, awaiting implementation or merely directory in nature. The paper is an empirical analysis of web-scraping and privacy policies of major social media intermediaries operating cross-borders including India. The paper further analyses the efficiency of tech-privacy specific statutes of India and other major jurisdictions, the European Union, California and the Peoples' Republic of China.

RESEARCH PROBLEM

The research delves into the data scraping and privacy policies of major social media platforms that allow the accessibility of personal data of users to third party applications for specific purposes or research and development of their Artificial Intelligence models and often entering into licensing agreements. The paper analyses the implications of such policies and the differences of such implementations in various jurisdictions. Finally, the paper compares the statutory frameworks of four countries to highlight the drawbacks of safeguards in Indian cyber-privacy protection.

RESEARCH OBJECTIVES

The objective of the paper is to effectively carve out the efficiency of various statutory frameworks in preventing data-scraping and the legality of licensing datasets of individuals to third party applications. The research is aimed at understanding how different jurisdictions have stepped beyond traditional measures to protect and preserve personal data of its citizens on platforms that are global, sensitive and highly interlinked with softwares that require scraped data for developing their platform models.

RESEARCH QUESTIONS

1. Whether social media intermediaries prevent or promote data scraping through their web-scraping and privacy policies?
2. Whether data-scraping policies operate differently in accordance with jurisdictional regulations?
3. Whether statutory frameworks in India sufficiently address data scraping in compliance with extra-jurisdictional norms?

RESEARCH METHODOLOGY

The research paper provides an insight into the data-scraping models and privacy policies of major social media intermediaries X, Reddit and Meta to understand their data protection compliances. The paper further discusses a case study of scraping data for training AI models backed by such intermediaries. The research compares the data protection laws of major jurisdictions of the European Union, California, the Peoples' Republic of China and India. The paper analyses the efficiency and familiarity of Indian statutory frameworks with the modern-day data scraping practices of publicly available data.

WEB SCRAPING POLICIES OF THE BIG TECH

Specific observations of three prominent social media intermediaries have been provided as under with major focus on their web scraping policy, the amount of data collected and the duration of retention of such data.

I. X

The first platform collects the content of the posted material, the bookmarked content, the Communities joined, tagged contents, viewing history and other broadcasts. Further, it scrapes the metadata under Direct Messages including of encrypted designs, the content of such encrypted conversations; moreover, when concerning log in credentials and information, the data collected is also associated with browsers that are not used to sign into the concerned social media platform. The platform uses such collected or scraped data to generate personalized advertisements, connecting with third party applications and also to train machine learning models for artificial intelligence models. The information is further shared for research and development purposes including product testing and other improvisations. The platform further uses API embedded technology for sharing the data with third party applications, going so forth as to charge licensing fees for large-scale access to the data shared on their platform. The platform explicitly states that the information about the users may be shared, sold or transferred if in case of a merger, asset change or bankruptcy. The social media platform retains the collected data for different durations for different purposes as required, ranging from 12 and 13 months to the duration of the activity of the account till the removal of the content shared. Further, even after removal or deletion of content from the platform, the copies of the same can be retained on other third-party applications. Moreover, the privacy policy was updated on January 15, 2026.[4]

[4] X, 'X Privacy Policy' (*X Privacy Policy*, 15 January 2026) <https://x.com/en/privacy> accessed 20 April 2026

II. Reddit

Another public platform that we researched upon practices strict mechanisms to abide by privacy policies and statutory safeguards specific to the operating jurisdictions and restricts web-scraping accordingly. It collects data such as username, profile picture, public content like posted material and its related metadata and also non-public content, private messages and private chats with other users, information submitted to surveys and analytics and so on. Also, information is collected from other third-party applications if the platforms are connected or interrelated, such as web pages or ad technologies. However, the platform enters into agreements with third parties like researchers and developers for licensing public content. These agreements help to restrict the amount of data scraped and the use of such data to comply with statutory regulations, like sharing of sexual content or use the deleted content. Further, the data shared and scraping is consensual and restricted to a few moderators only. The concerned platform uses data to manage and abide by the age verification protocols, restrict content according to age protocols, to monitor inconsistent activities and for affiliate partners and its subsidiaries. The concerned platform's policy states that it shall store the collected information for as long as required for specific purposes, only in accordance with the applicable law. The privacy policy of the platform was updated on January 6, 2026.[5]

III. Meta

Another contemporary social media platform practices an updated privacy policy according to the age regulations. It collects data such as the family device IDs, camera access and photos, network connected to the device and whether the app is working in the foreground in addition to mouse clicks. Further it collects the information about the apps and games on the device and also the usage of third-party services whether in person or through online networks. This data is scraped through business tools and integrations that collect data upon visiting their applications. However, the concerned platform mandates that such data is rightfully collected and shared. The platform uses data similar to other platforms like the posted material, accessible pictures, frequency of use. However, it considers some data as protected such as sexual orientation and political views in accordance with jurisdictional

[5]Reddit, 'Reddit Privacy Policy' (*Reddit Privacy Policy,* 6 January 2026) < https://www.reddit.com/en-us/policies/privacy-policy> accessed 20 April 2026

rules. The platform uses the information to test new features and products, conduct research and surveys, personalize advertisements and detect suspicious activities. The data used with advertisers is limited to empirical as in the frequency of engagement with their advertisements, yet, keeps identifiable information from these advertisers. The information shared with external researchers is privacy protected with goals of advancing scholarship and innovation. The AI integrations include the data shared with AI platforms by the users and not more than the same. For effectively training generative AI models, however, the platform collects data which is publicly available online and any other licensed information from other partners, which may include private information**.** A non- user's data is also collected for such training purposes through any visual presence in posted material or content of a platform's user. The information used includes chats with AI bots and any deleted information will be retained but not used for future training practices. It also includes an anti-scraping team however its effectiveness is restricted to the content outside its privacy policies. The privacy policy was last updated on 16th December, 2025.[6]

Other intermediaries either have similar policies or lack such policies which makes user data on them more vulnerable to being scraped and misused. However, these intermediaries through user agreements or cookies often get a broad unfettered global license to use, sell, store, process, and share such data within their ecosystem to train their large language models or AI bots such as Chat GPT, Gemini, Grok and with third party partners for monetization and commercial usage of the data while users of these platforms retain a nominal control over their publicly available data. This creates an imbalance between external platform compliance and internal violation of regulations. These big social media companies follow an opt out model which allows the users to withdraw their consent to any future data being shared through privacy settings or account deletion. However, this model is limited in nature because it operates prospectively, that is, the opt out model does not erase or undo the already stored data in the database of these social media giants. In contrast, GDPR follows opt in method which is globally considered to be stricter than opt out method followed by India. The opt in method makes it compulsory for the organizations to clearly and unambiguously specify to what extent the consent is being taken of the users. The GDPR model allows the users to withdraw their consent at any time by using their right to be forgotten as elaborated under art. 17 of the GDPR.

[6] Meta, ' Meta Privacy Policy' (*Meta Privacy Policy, 16 December 2025*) <https://www.facebook.com/privacy/policy?subpage=1.subpage.4-InformationFromPartnersVendors> accessed 20 April 2026

This makes the GDPR model flexible yet compliant in nature. India's Digital Personal Data Protection Act, 2023 attempts to address this gap by recognizing rights such as the right to erasure under sec 12 of the act, allowing the users to request deletion of their personal data, though its restricted enforcement and scope still evolving. Thus, while current platform policies emphasize user control in theory, the limitations of opt out mechanism highlights the need for stronger, GDPR like regulations to preserve data autonomy of users without overriding their right to privacy.

WEB SCRAPING FOR TRAINING AI PLATFORMS: A CASE STUDY

AI development at a large scale in India requires Open Data to train the AI platforms operative in India by companies such as Meta AI, OpenAI, Microsoft, Haptik, Tata Elxsi and more[7]. The datasets used for training, evaluating and testing AI can be structured, unstructured, targeted or large volumes of scraped data.[8] These large-scale automated web scraping have often infringed the Copyrighted data, which is violative of and governed under the WIPO and TRIPS Agreements[9] and particularly the Copyright Act, 1957 in India[10]. The unstructured and randomly collected data poses threats to personal data that is shared publicly and is accessible on social media platforms such as LinkedIn. This data is breached to create datasets required for training AI systems by Tech giants like Microsoft without validating consent[11]. Such open datasets, once created, continue to utilize the scraped data for training AI models despite the deletion or removal of content from publicly accessible platforms.[12] For the purposes of training AI platforms, India has established IndiaAI Dataset Platform for Open Data Access, an archive with non-personal datasets[13] similar to the consent token-based data sharing system

[7] Mayank Shukla, 'List of Best AI Companies in India-2026' (*KSOLVES,* 14 January 2026) <https://www.ksolves.com/blog/artificial-intelligence/best-companies-in-india> accessed 20 April 2026

[8] Lee Tiedrich, Karine Perset, Sara Fialho Esposito and Audrey Plonk, 'Intellectual Property Issues in AI Trained on Scraped Data' (2025) 33 OECD <https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/02/intellectual-property-issues-in-artificial-intelligence-trained-on-scraped-data_a07f010b/d5241a23-en.pdf> accessed 20 April 2026

[9] ibid

[10] Copyright Act 1957

[11] Amit Kapoor and Mohammad Saad, 'The illusion of privacy in India's AI boom' (*The Economic Times,* 8 December, 2025) <https://economictimes.indiatimes.com/news/company/corporate-trends/the-illusion-of-privacy-in-indias-ai-boom/articleshow/125842310.cms?from=mdr> last accessed 20 April 2026

[12] Abdulkareem Alsudais, ' Training Vision AI Models with Public Data: Privacy and Availability Concerns' (2025) 56 Communications of the Association for Information Systems <https://www.researchgate.net/publication/390453030_Training_Vision_AI_Models_with_Public_Data_Privacy_and_Availability_Concerns> accessed 20 April 2026

[13] Press Information Bureau, 'India's AI Revolution' (*Press Information Bureau Government of India*, 6 March 2025) <https://www.pib.gov.in/PressReleasePage.aspx?PRID=2108810&reg=3&lang=2 >last accessed 20 April 2026

developed by Data Empowerment and Protection Architecture (DEPA)[14]. India's AI Governance Guidelines acknowledge transparency failures in the use of personal data for developing AI platforms, in its risk-based assessment and state the Information Technology Act, 2000 and The Digital Personal Data Protection Act, 2023 as the governing statutes for AI Systems[15].

In India, personal data is protected through the strict consent-based mechanism for sharing data with Data Fiduciaries under the Digital Personal Data Protection Act, 2023. The Act enables the Data Principal or the person to whom the Digital Data is related to, to update, modify and erase the personal data which was previously permitted to be used by Data Fiduciaries, under Section 12[16]. The provisions regarding the mandatory removal of personal data being accessed by Data Fiduciaries after a certain time period is provided under Rule 8 of the Rules of 2025[17] that render the use of scraped data for training models associated with Artificial Intelligence, unlawful after a certain duration. However, a lacuna is evident regarding the data that is shared publicly since the use of such data is exempted from the purview of the Act of 2023 and further the scraping and utilisation of personal data for research and archiving purposes remains unchecked under the Digital Personal Data Protection Act, 2023 and its corresponding Rules of 2025. This pertains to the question of legitimacy of scraping data for teaching and priming the AI platforms particularly the personal data not consented to be used for research purposes and the personal data digitally shared on publicly accessible platforms in India.

STATUTORY SAFEGUARDS AND INTERVENTIONS

India

The Digital Personal Data Protection Act, 2023 aims to safeguard the digital personal data of the citizens, referred to as the Data Principal as one to whom the data belongs. This data pertains to the personal information ranging from texts and images to video-taped data. The 2023 legislation mandates a mechanism for seeking and validating consent of the Data Principal for

[14] Press Information Bureau, ' Indian AI Governance Guidelines' (*Press Information Bureau Government of India*,)<https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf > last accessed 20 April 2026

[15] Press Information Bureau, ' Indian AI Governance Guidelines' (*Press Information Bureau Government of India*,)<https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf> last accessed 20 April 2026

[16] Digital Personal Data Protection Act 2023, s 12

[17] Ministry of Electronics and Information Technology, 'Digital Personal and Data Protection Rules, 2025' (*Ministry of Electronics and Information Technology Government of India* 14 November 2025) <https://www.meity.gov.in/documents/act-and-policies/digital-personal-data-protection-rules-2025-gDOxUjMtQWa?pageTitle=Digital-Personal-Data-Protection-Rules-2025> last accessed 20 April 2026

processing personal data shared with the Data Fiduciaries under Section 4[18]. This guides and checks the privacy policies of tech companies like Google, Microsoft, Meta and others that have a wide base of consumers in India. In contradiction, the 2023 Act explicitly exempts a significant volume of data in Sections 3 and 17 respectively. This excluded data refers to the information shared freely on publicly accessible platforms in addition to the personal data for which, the consent mechanism is absent regarding "Research, Archiving or Statistical" purposes.

With respect to Artificial Intelligence, web crawling or mining personal data for training can be guised as research or archiving purposes, however, the practice conflicts gravely with the regulations regarding the use of such personal data particularly with respect to withdrawal of consent from data sharing and the right to erasure of personal data of the concerned Data Principal under Section 12[19]. The DPDP Act, 2023, mandates strict compliance with explicit, clear and unambiguous methods of giving consent to the processing of personal data and its withdrawal under Section 6[20].

However**,** when the digital data is publicly shared willingly or unwillingly by the principal, it can be accessed and used by Companies for training Artificial Intelligence. This practice reflects two conflicts, firstly, the absence of consent mechanism of such digital personal data publicly shared[21] leaves a gap in exercising the right of erasure of such data. Artificial Intelligence models incorporate text and data mining and build upon them. Such practice leaves little space for previously accessed personal data to be removed from the datasets fed to these models and this space is further reduced in the absence of an established protocol for consent withdrawal regarding publicly shared data[22]. Secondly, despite the presence of consent validating mechanism for processing data for strictly informed purpose, the 'Research, Archiving and Statistical Use' provision provides a license to the Fiduciaries to exploit this personal data for purposes other than consented to, provided that the processing is done in accordance with such standards as may be prescribed[23]. The concerned standards are listed

[18] Digital Personal Data Protection Act 2023, s 4
[19] Digital Personal Data Protection Act 2023, s 12
[20] Digital Personal Data Protection Act 2023, s 6
[21] Digital Personal Data Protection Act 2023, s 3
[22] Abdulkareem Alsudais, ' Training Vision AI Models with Public Data: Privacy and Availability Concerns' (2025) 56 Communications of the Association for Information Systems <https://www.researchgate.net/publication/390453030_Training_Vision_AI_Models_with_Public_Data_Privacy_and_Availability_Concerns> accessed 20 April 2026
[23] Digital Personal Data Protection Act 2023, s 17

under the Second Schedule of the Digital Personal Data Protection Rules, 2025. These standards include the processing of personal data to be practised in a lawful manner, for a prescribed duration so long as the Fiduciaries exercise reasonable safeguards for the security of such personal data[24]. The Third Schedule of Digital Personal Data Protection Rules, 2025 further mandate the Data Fiduciaries such as e-commerce entity and social media intermediary that have a registered user base of at least two crores in India to cease the processing of personal data after three years of such initiation[25].These provisions indicate the presence of a consent committed data sharing protocol for the purpose of data processing. However, these compliances create a distance between the letter and spirit when Data Fiduciaries blatantly exempt from disclosing the purpose and utilisation of the processed data, time period for storing such data and further, the information regarding sharing of processed data with third-party platforms.

Harsh measures and tedious procedure of withdrawing consent from Data Fiduciaries indirectly restrict the right to erasure of data previously processed. Meta's guidelines and information about datasets in training GPT-4[26] are becoming increasingly opaque despite its policy of utilizing data from Instagram to train its AI models[27]. Meta provides the opt-out option from data processing to countries like the EU and the UK only, where AI specific statutes have been enforced[28]. The May 2025 deadline of consent withdrawal from Meta reflects a pattern of restricting enforcement of the right of erasure of data. Post the deadline, any objection would only cease further processing, however, would not remove the previously processed data from its datasets[29]. These measures hinder, restrict and violate the statutory provisions including the right of erasure of personal data[30] and the 'Right to be Forgotten', flowing from the Doctrine

[24] https://www.meity.gov.in/documents/act-and-policies/digital-personal-data-protection-rules-2025-gDOxUjMtQWa?pageTitle=Digital-Personal-Data-Protection-Rules-2025
[25] https://www.meity.gov.in/documents/act-and-policies/digital-personal-data-protection-rules-2025-gDOxUjMtQWa?pageTitle=Digital-Personal-Data-Protection-Rules-2025
[26] Lauren Leffer, ' Your Personal Information is Probably Being Used to Train Generative AI Models' (*Scientific American* 19 October 2023) https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/ last accessed on 20 April 2026
[27] Karthika Rajmohan, Shravani Nag Lanka, ' The Opt out conundrum: Analysis of Web Scrapping to train AI models' (*Internet Freedom Foundation* 19 December 2023) https://internetfreedom.in/analysis-of-web-scrapping/ last accessed on 20 April 2026
[28] Karthika Rajmohan, Shravani Nag Lanka, ' The Opt out conundrum: Analysis of Web Scrapping to train AI models' (*Internet Freedom Foundation* 19 December 2023) https://internetfreedom.in/analysis-of-web-scrapping/ last accessed on 20 April 2026
[29] Information and Data Protection Commissioner, ' Meta to Use Facebook and Instagram Personal Data for AI training- GDPR rights and considerations (*Information and Data Protection Commissioner,* 22 April 2025) < https://idpc.org.mt/news-latest/meta-to-use-facebook-and-instagram-personal-data-for-ai-training-gdpr-rights-and-considerations/> last accessed 20 April 2026
[30] Digital Personal Data Protection Act 2023, s 3

of Privacy enshrined under Article 21[31] of the Constitution and interpretations of Justice K.S. Puttaswamy (Retd.) v. Union of India (2017)[32] and Prathamesh Nagesh Patil v. State of Maharashtra (2024)[33], which constitutes the very basis of the Digital Personal Data Protection Act, 2023. Although the Digital Personal Data Protection Act is not the sole governor of personal data in India, the Information Technology Act, 2000 is a crucial legislation that states proper guidelines related to data being accessed online or on any internet platform using any computer source. Section 43A of the Act of 2000 mandates reasonable practices to secure the personal data that classifies as sensitive[34]. However, IT Act does not cover the personal data and its regulation as specifically as the DPDP Act is formulated to do. The dilemma that the Indian Technological and Innovative Diaspora faces is that the IT Act does not protect and legislate the Digital Personal Data of an individual further than the consent framework and the DPDP Act does not enforce its two crucially conflicting provisions, Sections 3 and 17 before 14th May, 2027 that are concerned with publicly available data[35].

When the legitimacy of web scraping for training AI models is concerned, India's AI Governance Guidelines specify that the AI models' handling shall be done in strict compliance with laws, not limited to the IT Act and the DPDP Act, with measures applicable to protect privacy and publishing mandatory transparency reports informing the posed threats to the Indian society[36]. The guidelines further specify the Data Protection Board as a regulator to study the risks of AI and the need[37] for any amendments in the existing legislations[38]. The Data Protection Board of India has been established under the enforcement of particular provisions of the DPDP Act on 14th November 2025. However, the core compliance protocols are still awaited to be enforced. The scope of the Digital Personal Data Protection Act, 2023 is the

[31] SCO Team, ' The right to life and personal liberty under Article 21: A timeline ' (Supreme Court Observer, 26 June 2025)<https://www.scobserver.in/journal/the-right-to-life-and-personal-liberty-under-article-21-a-timeline/> last accessed 20 April 2026
[32] SCO Team, 'Fundamental Right to Privacy' (Supreme Court Observer) <https://www.scobserver.in/cases/puttaswamy-v-union-of-india-fundamental-right-to-privacy-case-background/> last accessed 20 April 2026
[33] Anushka, ' Legal Recognition of the Right to be Forgotten in India- A case study of Prathamesh Nagesh Patil v. State of Maharashtra(2024)' (*Lawfullegal,* 7 June 2025)<https://lawfullegal.in/legal-recognition-of-the-right-to-be-forgotten-in-india-a-case-study-of-prathamesh-nagesh-patil-v-state-of-maharashtra-2024/#:~:text=The%20Bombay%20High%20Court's%20ruling,to%20privacy%20and%20personal%20dignity.> last accessed 20 April 2026
[34] Information Technology Act, 2000
[35] DPDPAedu, 'Background and Purpose of DPDPA' (*DPDPA Implementation Consultation*) <https://dpdpaedu.org/docs/Overview/Background%20and%20Purpose%20of%20DPDPA/> last accessed 20 April 2026
[36] India AI Governance Guidelines, 2025
[37] ibid
[38] ibid

security and protection of digital privacy of the person, however, ambiguous definitions, exclusion of publicly accessible data and the absence of addressing Artificial Intelligence deploying systems and the emerging risks creates a gap that renders lawful, the scraping of web and personal data publicly shared for training AI platforms, so long as, it qualifies for research purposes. The Rules of 2025 explicitly provide that data processing shall be done in a lawful manner with reasonable security safeguards, however, the 'reasonable safeguards' await an assertive interpretation, in absence of which web scraping remains largely unchecked. As the sole guide for AI innovations in India, AI Governance Guidelines reiterate the safeguarding of privacy of data, 0discourage web scraping and provide consent-based datasets for companies to train AI platforms**.** Efforts to protect data have been successful in countries with staunch AI legislations classifying high risk systems, even against data scraping for purposes did not consent to[39]. In order to safeguard privacy in an AI driven world, Indian regulatory framework needs to advance from a guideline-based model to compliance-based model for the AI sector through clear definitions of 'research purpose', expanded scope for personal data, stricter opt-out protocols, and materialise grievance redressal forums for the breach caused by web scraping and data mining for training AI platforms.

The European Union

The General Data Protection Regulation is a comprehensive framework passed in 2016 and enforceable since May 25, 2018. The General Data Protection Regulation provides the strictest and an exhaustive privacy regulation for the processing of personal data in the European Union. The Statute flows from the European Convention on Human Rights, 1950 that enshrines the right to privacy. The definition clause under Article 4 of the Regulations extensively defines and covers the subject matter, that is, personal data related to identifiable natural person or data subject, restriction of processing, pseudonymisation when personal data is not attributable to an identified natural person or data subject, and free, specific, informed and unambiguous consent for data processing. GDPR covers biometric data, genetic data, data concerning health among other classifications[40]. Data protection principles are elaborated under Article 5 which mandates that processing is lawful, fair, for a legitimate interest, limited to required data, limits on duration of retention of such data, further, it mandates the demonstration of GDPR compliances and not merely verbatim statements by the Social Media Enterprises.[41] Article 6

[39] Historic AI Team, 'High-Risk AI Systems Under the EU AI Act' (*EU AI Act,* 1 August 2024)<https://www.euaiact.com/blog/high-risk-ai-systems-under-the-eu-ai-act> last accessed 20 April 2026
[40] United Kingdom General Data Protection Regulation 2016, art 4
[41] United Kingdom General Data Protection Regulation 2016, art 5

enumerates the conditions that legalize processing of personal data only to the extent of fulfilment of any of the prescribed principles. These principles include consent for specific purposes, necessary or non-negotiable processing, legal compliances, protection of vital interests of data subjects and those pursued by third parties.[42]

Conditions for consent are exhaustive and elaborate, which highlights the feature of GDPR in its precision and accuracy to protect data subjects. The conditions under Article 7 imposes liability on data controller or the social media intermediary to demonstrate and prove that the consent was provided by the user through a request which distinguishes clearly, in an intelligible manner and understandable language the need for consent from other matters. Within the Article, is enumerated the right to withdrawal of consent at any time as opted by the data subject[43] and further the right to erasure under Article 17[44]. Data protection by design and by default reflects the robust mechanism, rule of liability upon controllers and mandatory protection to data subjects under Article 25 of the GDPR. Thereby, it mandates the data controller to implement appropriate technical measures by default at the time of processing such data. Further, the Article grants higher protection to children's data owing to the lack of substantial understanding of the consequences and it recognizes, acknowledges and protects the different needs at different stages of development[45]. Apart from special categories of personal data under Article 9, GDPR provides for a Data Protection Officer in three cases, as a public authority performing judicial functions, organization monitoring data on a large scale and special data categories as listed under Articles 9 and 10. The functions of these Officers include informing about the Regulation's compliances to controllers, monitor such compliances, conduct impact assessments under Article 35[46]. The penalties include a fine of up to 8.7 million pounds sterling and 2% of the preceding financial year's annual turnover under Article 83[47].

EU AI Act is the leading AI framework in the world for regulations regarding web scraping for training AI platforms through data available on Social Media Intermediaries. The Act was formally published in the Official Journal of the European Union on 12 July, 2024[48]. EU AI Act classifies the risks associated with Artificial Intelligence into 4 degrees of harm,

[42] United Kingdom General Data Protection Regulation 2016, art 6
[43] United Kingdom General Data Protection Regulation 2016, art 7
[44] United Kingdom General Data Protection Regulation 2016, art 17
[45] United Kingdom General Data Protection Regulation 2016, art 25
[46] United Kingdom General Data Protection Regulation 2016, art 35
[47] United Kingdom General Data Protection Regulation 2016, art 83
[48] https://artificialintelligenceact.eu/developments/

unacceptable risk, high, limited and minimal risk. Among these, web scraping for expansion of facial databases is classified as unacceptable risk, in addition to 7 other counterparts. Article 10 regulates training and tests of data sets in accordance with data governance with particular guidelines on the collection processes and maintenance of the intended use and no more of the personal data of the individuals[49]. The Act prohibits the transfer and use of special categories of data and mandates a particular retention period for the same. The prohibited AI practices under Article 5 includes deploying deceptive techniques and exploitation of any vulnerability or using AI systems for evaluation and classification based on predicted personal and personality characteristics[50]. This ensures that web scraping is kept under check and adequate measures are taken when Artificial Intelligence in deployed for purposes of web scraping and otherwise. Together, the EU AI Act and the GDPR provide a strict mechanism that protect user data from not only web scraping but misuse by Artificial Intelligence.

Peoples' Republic of China

China has an amalgamation of multilayered framework for protection of their citizen's data through the Personal Information Protection Law (PIPL), Cybersecurity Law (CSL), Data Security Law (DSL), and the Anti- Unfair Competition Law (AUSL). The primary legislation that regulates the processing of personal information of individuals is Personal Information Protection Law (hereinafter referred as 'PIPL') enacted on 20th August 2021 and came into effect on 1st November 2021. [51] This law is the heart and soul of protection of personal data of the Republic of China. The PIPL is divided into 8 chapters. The legislation has various key prominent features i.e. broad definition of personal information, extra territorial nature, stringent requirements, informed consent and harsh punitive measures.[52] Article 4 of the law defines personal information and its usage, as any data recorded electronically or in other forms that relates to identified or identifiable persons, excluding data that has been anonymized. The processing of personal information includes a broad spectrum of operation such as to store, use, transfer, publish, processing and erasure of personal information. The processing of personal information as defined under Article 5, 6, and 7 should be done for fair use and in permissible limits as set by the law. Processing, selling or transmission of information could

[49] Artificial Intelligence Act 2024, art 10
[50] Artificial Intelligence Act 2024, art 5
[51] DLA Piper, 'Data Protection Laws in China' (*DLA Piper* 13 February 2026) <https://www.dlapiperdataprotection.com/index.html?c=CN> last accessed on 20 April 2026
[52] Hawksford, ' Understanding China's Personal Information Protection Law (PIPL)' (*Hawksford* 5 November 2025) <https://www.hawksford.com/insights-and-guides/china-pipl-compliance-guide> last accessed on 20 April 2026

only happen through obtaining informed consent of the people, which is enumerated under article 13, 14, and 15 of the law. The Articles obligate data processors to obtain a prior consent of the individuals which must be voluntary, free, and informed. The law gives flexibility to individuals to withdraw their consent in their convenient manner. The PIPL places individuals at the core of the regulatory framework by granting them substantial control over their personal information as elaborated under Article 44, 45, and 46 of the law to know about, restrict, refuse, modify, obtain copies of and most importantly right to erasure of their data. The act places obligation on the processors as explained under Article 51 and 52 that while processing any personal information they shall run an impact assessment which shall include type of personal information being processed, impact on individual's liberty, and prospective breach of privacy along with which the data processors shall appoint a Data Protection Officer to supervise the processing of personal data and ensuring the implementation of data processing safeguards. In case of non-compliance of the provisions, it imposes severe penalties as given under Article 66 which are imposing fines up to RMB 50 million or 5% of annual turnover, along with suspension of business or revocation of license, and in exceptional circumstances a criminal liability might also be imposed.[53] The Chinese framework is efficient for regulating personal data processing within and outside the borders. It has strong enforcement mechanisms including heavy penalties, regulatory oversight and accountability.

California

The U.S. State assembly of California passed the California Consumer Privacy Act (herein after referred as 'CCPA') in 2018 and came into effect on 1st January 2020 which made California to be the first state of U.S. to have a national legislative act regulating processing of resident's personal data. [54]The CCPA is build and strengthened upon California Privacy Rights Act, 2020 (hereinafter referred as 'CPRA'). A new amendment was made in the act by CPRA which came into effect on 1st January 2023, refers to the inclusion of a new category of personal information called as sensitive personal information, expansion of rights such as opt- out right, and establishment of California Consumer Regulatory Agency. [55]The CCPA established

[53] Personal Information Protection Law 2021 (People's Republic of China)
[54] Salar Atrizadeh, ' The CCPA Impact Privacy and Cybersecurity Guideline' (*ISACA* 9 March 2020) https://www.isaca.org/resources/news-and-trends/industry-news/2020/the-ccpa-impact-privacy-and%20cybersecurity%20guideline?gad_source=1&gad_campaignid=20381363240&gbraid=0AAAAAD_A9K8ktZMSIKgbCKSF4DWt-IL_A&gclid=Cj0KCQjw-pHPBhCdARIsAHXYWP9f7pIFxOh1ePlHSPIukKXj26b5im6eaGK2adeY7XuReL3wLsexio0aAqg7EALw_wc > last accessed on 20 April 2026
[55] Natasha Piirainen, 'CCPA vs. CPRA: What's Different and What's THE Same?' (*Termly* 2 March 2026) < https://termly.io/resources/articles/ccpa-vs-cpra/ > last accessed on 20 April 2026

foundational privacy rights, it had certain limitation i.e. to erase a digital footprint of an individual it required them to contact potentially hundreds of the companies separately. Thus, to address this, California passed the Delete Act, 2023 which led to the creation of the Data Broker Requests and Opt-Out Platform (hereinafter referred as 'DROP'). The DROP enabled the users to submit a single request with registered data brokers to get their unauthorised data deleted from their platforms within a time frame of 45 days.[56] Personal information is defined under section 1798.140 of CCPA which defines any information that is capable of being associated to a particular consumer inclusive of their names, postal addresses, IP address, email name, citizenship status, religious belief and commercial history. The act provides consumers with substantive power to control their data, that is the consumers have the right to know, delete, access, opt out, limit the use of sensitive personal information, and non- discrimination as elaborated under section 1798.100, 1798.105, 1798.110, 1798, 115, 1798.120, and 1798.125. The act puts an obligation on business platforms to implement reasonable security measures to protect unauthorized access of the personal data of consumers, business platforms must update their privacy policy after every 12 months, the platforms that indulge in selling or sharing of information must inform the consumers beforehand and must comply with any request of data erasure, furthermore the data processors are mandated to run an risk assessment before involving in high risk data processing activities including selling, transmission or to train AI bots as enumerated under section 1798.100, 1798.130, 1798.135, and 1798. 185. As defined under section 1798.120 the business platforms under CCPA shall not process any information of someone who is less than 16 years of age. The consent mechanism of the act primarily follows an opt out model, however in the case of minors the act has difference stance i.e. to obtain consent of minors through opt in model which is considered to be stringent and strict. In case of non- compliance with the afore mentioned provisions, it imposes hefty fines up to $2500 per unintentional breach and $7500 per intentional data breach.[57] The California framework is a landmark, well- articulated user centric law giving millions of consumers the meaningful autonomy over their publicly accessible data

OBSERVATIONS

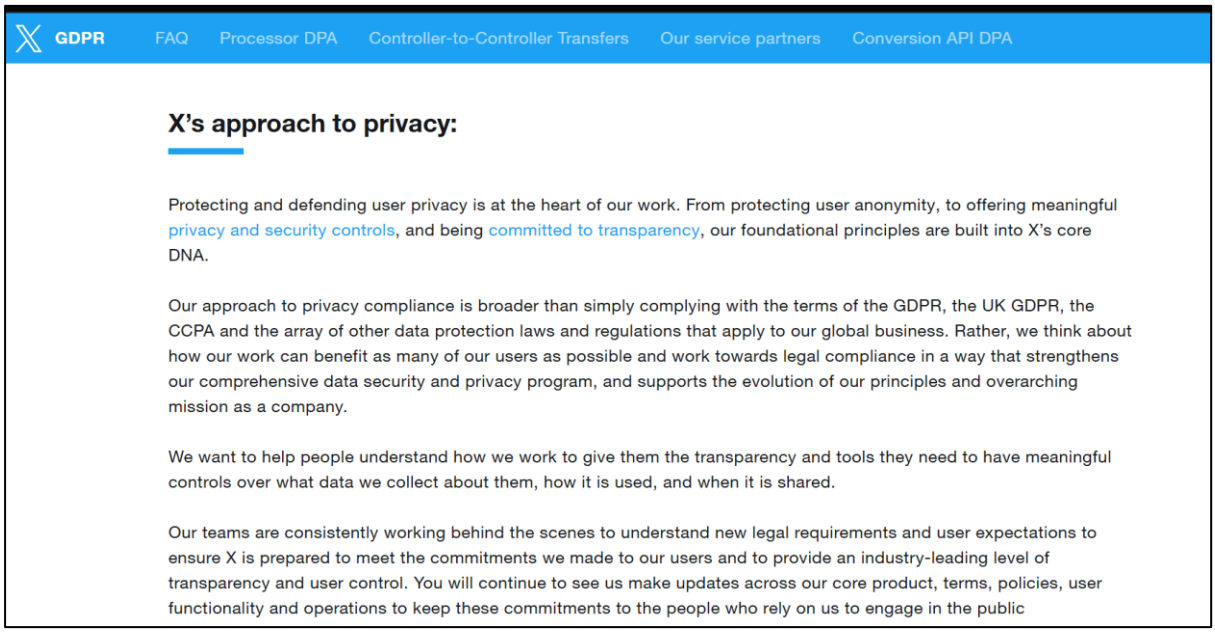
The data protection policies of different jurisdictions work on similar lines yet create varied responsibilities for data intermediaries. Such distinctions promote stricter compliance to

[56] Ian Phippe, ' The Delete Act and DROP: What You Need to Know' (*DataGrail* 17 November 2025) https://www.datagrail.io/blog/regulations/the-delete-act-and-drop-what-you-need-to-know/ > last accessed on 20 April 2026

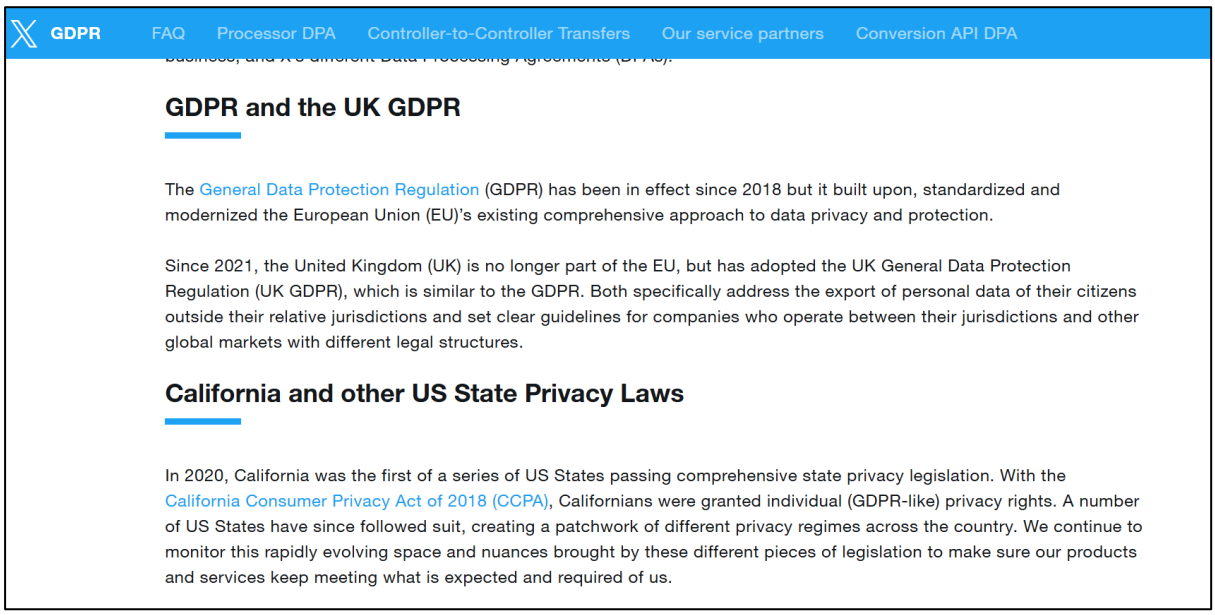
[57] California Consumer Privacy Act 2020 (California)

safeguard user data. Ultimately, the Social Media Intermediaries amend and update their privacy policies to comply with jurisdictional safeguards and prevent penalization. Such updates were observed in four major intermediaries, X, Meta, Reddit and LinkedIn.

At X, its approach to privacy is to move beyond particular compliances with user-centric privacy program. X recognizes and monitors safeguards of the GDPR and the CCPA and evolves its policies accordingly, as mentioned in annexures 1.1 and 1.2.



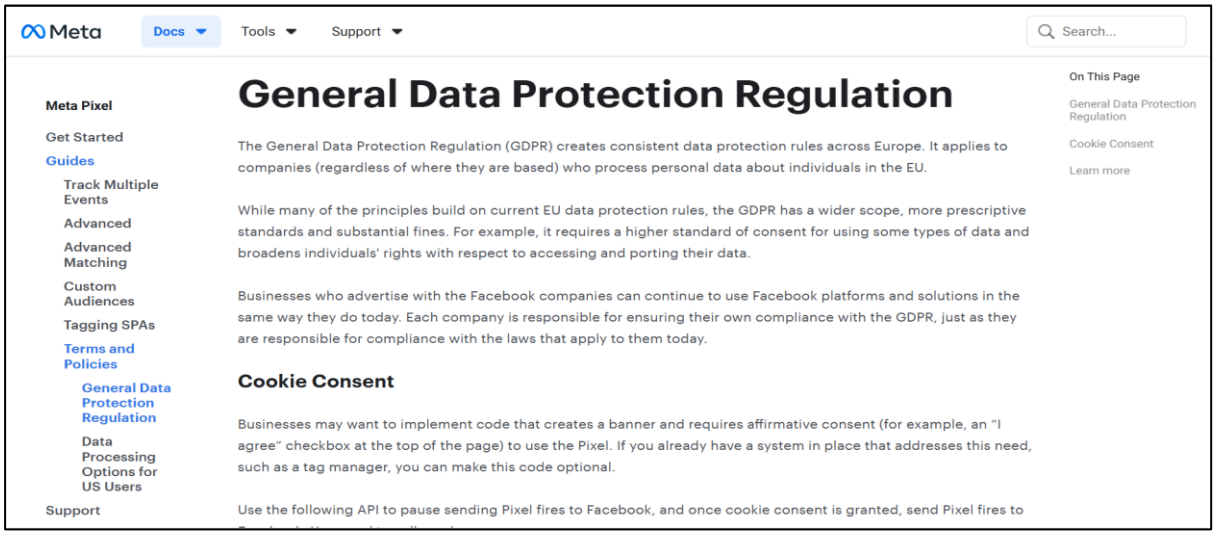
Annexure 1.1- X's approach to privacy[58]



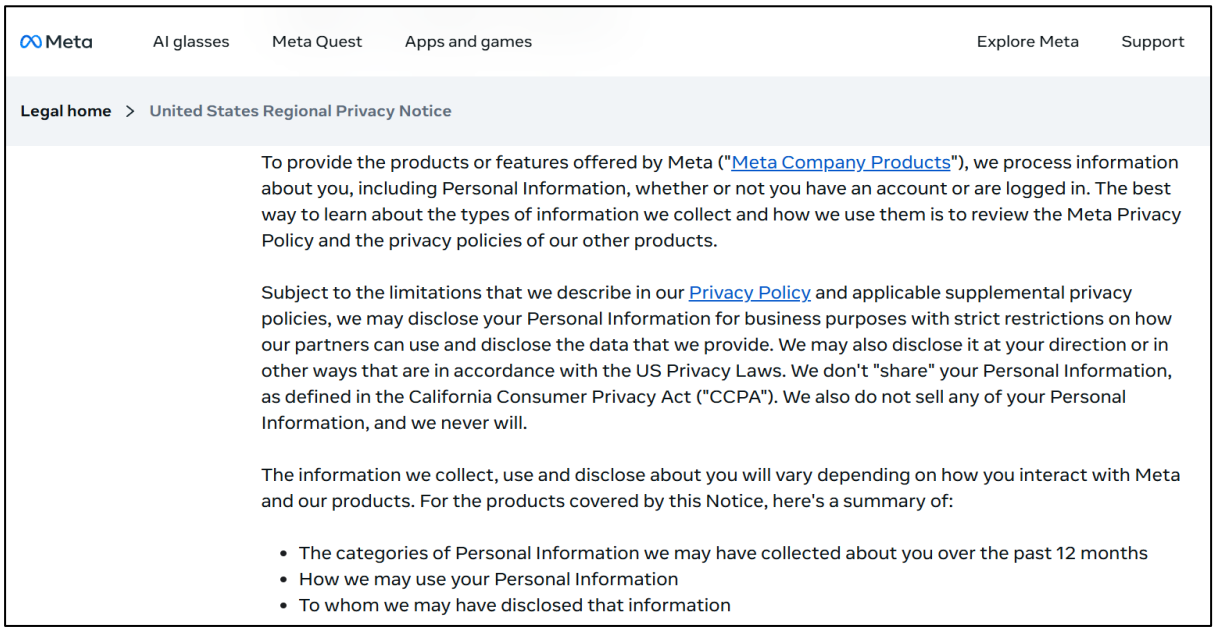
[58] 'Welcome to X's Privacy Hub' (*X GDPR*) <https://gdpr.x.com/en.html> last accessed 20 April 2026

Annexure 1.2- X's compliance with GDPR and US State privacy laws[59]

Secondly, Meta requires advertisers to specifically comply with GDPR directions and further require seeking and validating affirmative consent as under Article 7 of GDPR (Annexure 2.1). Further, Meta's policy explicitly declares its compliance with State specific laws like CCPA, evident from its staunch policy of not sharing personal information (Annexure 2.2). Further, United States Regional Privacy Notice for United States' residents is a right based document empowering its user base to exercise the US Privacy Laws (Annexure 2.3).

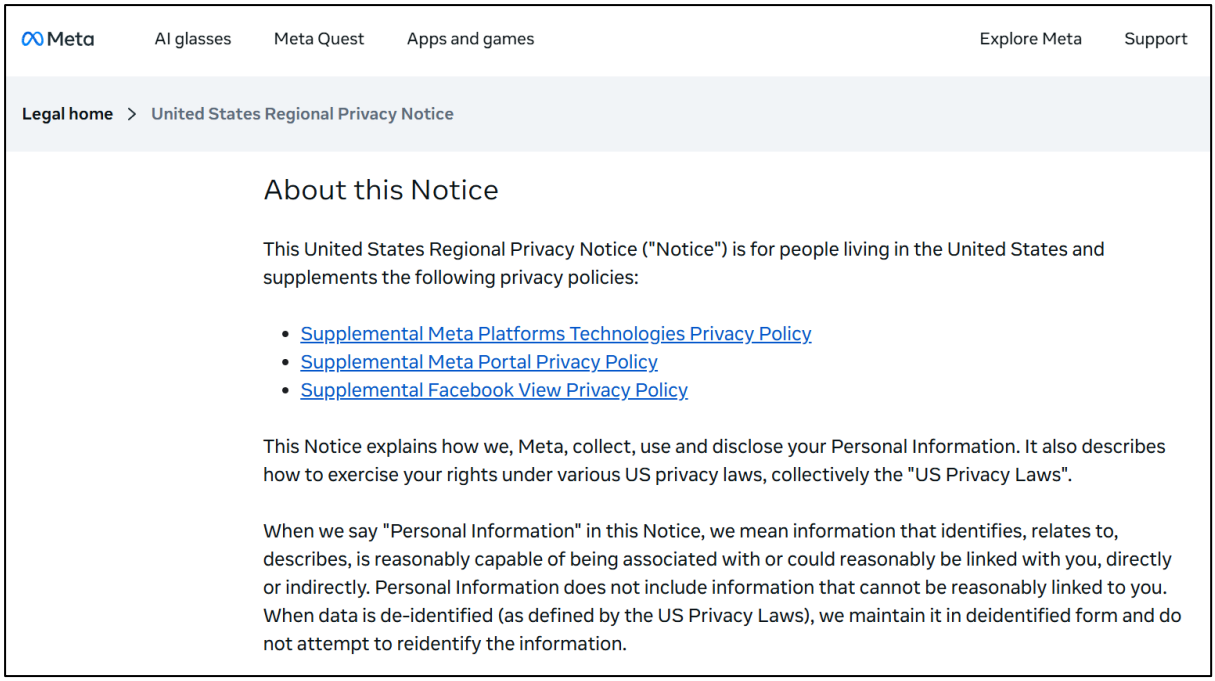


Annexure 2.1- Meta's compliances in accordance with GDPR[60]



[59] ibid
[60] 'General Data Protection Regulation' (*Meta*) https://developers.facebook.com/docs/meta-pixel/implementation/gdpr/> last accessed 20 April 2026

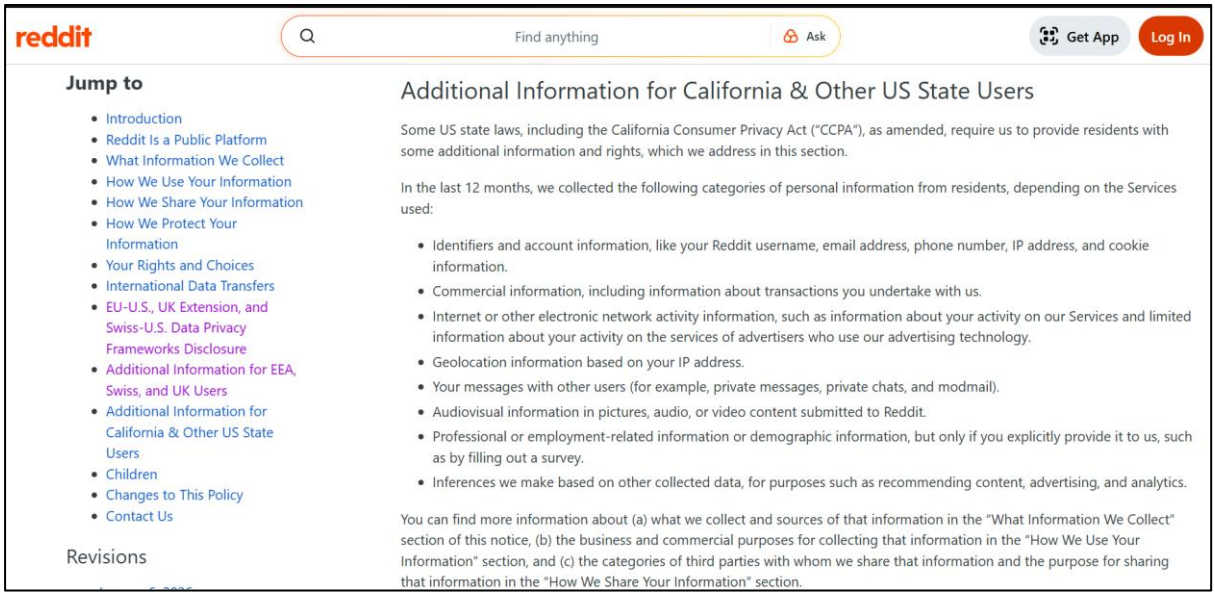
Annexure 2.2- Meta's guidelines in relation to the CCPA[61]



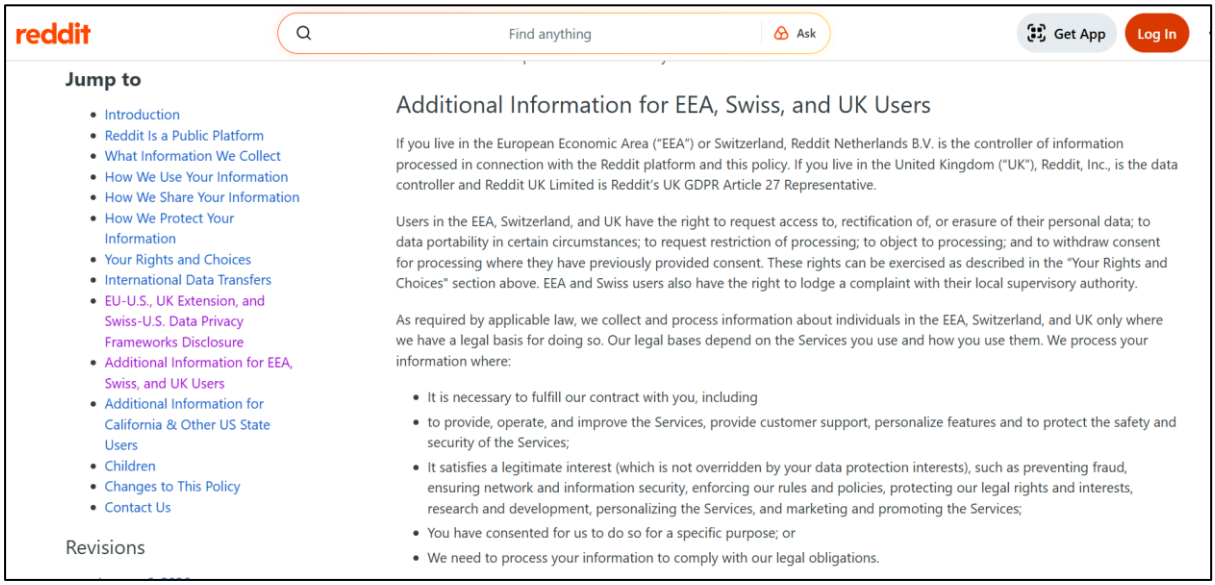
Annexure 2.3- Meta's policy regarding the US Privacy Laws[62]

Moving further, Reddit does not sell or share personal data and provides an opt out model in accordance with jurisdictional safeguards. It explicitly follows requirements of the California Consumer Privacy Act and provides users with the necessary information. It has provided a list of information collected to ensure a transparent mechanism of data scraping and data retention, which further aligns with the compliance of local laws and data protection guidelines (Annexure 3.1). Reddit also informs the users of its GDPR representative for operating in the European Economic Area (Annexure 3.2).

[61] ' United States Regional Privacy Notice' (*Meta,* 1 January 2025) https://www.meta.com/in/legal/united-states/?srsltid=AfmBOookwbyQoHGTK7XBWZca_rsiryKL5_xfgtkTCpZFFJzXe3wS6KH2> last accessed 20 April 2026
[62] ibid



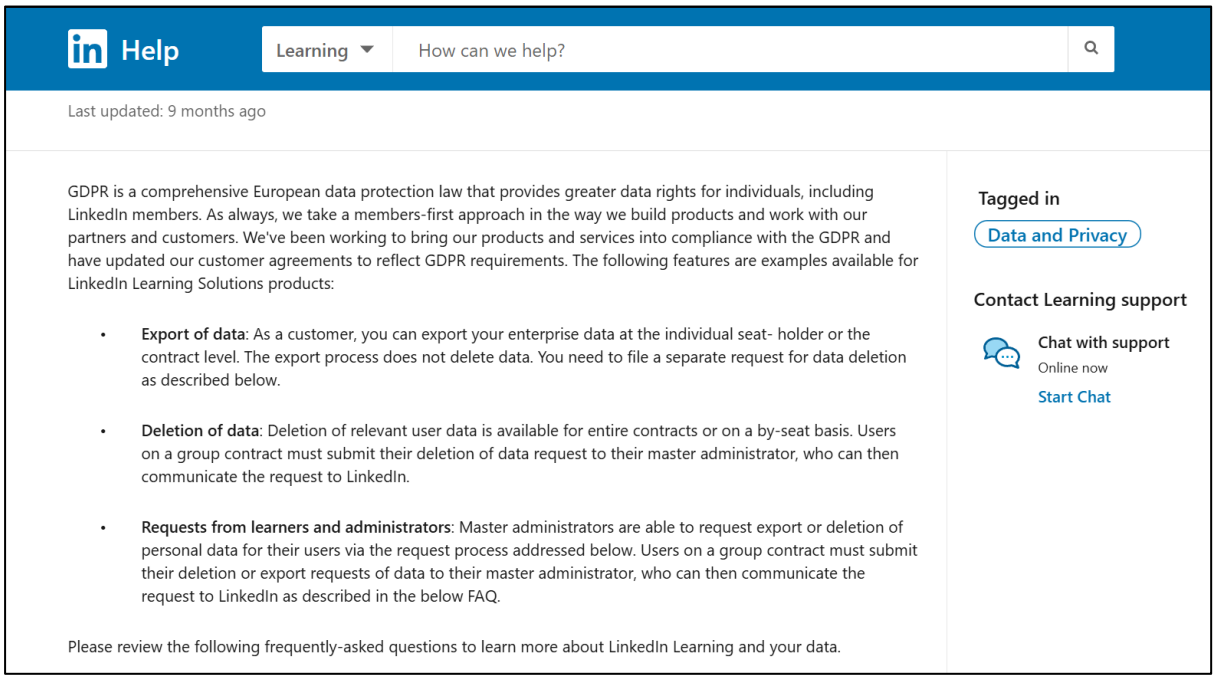
Annexure 3.1- Reddit's information for California[63]



Annexure 3.2- Reddit's information for UK[64]

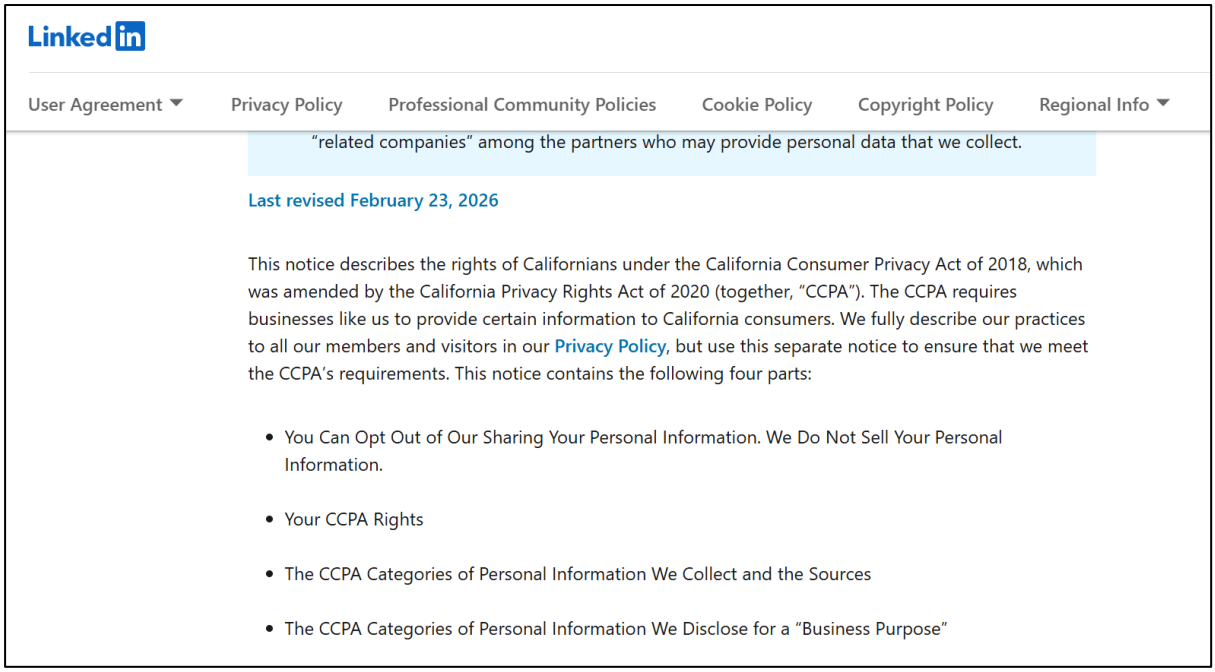
[63] 'Reddit Privacy Policy' (*Reddit)* https://www.reddit.com/en-us/policies/privacy-policy#policy-h2-10> last accessed 20 April 2026
[64] ibid

LinkedIn's policy as updated 9 months ago, reflects its efforts to bring its platform in compliance with the GDPR protocols with specific regulations as the data transfer, right to erasure of data and addressing grievances (Annexure 4.1). Further, the policy revised on February 23, 2026 informs to its users, the platform's compliance with the CCPA rules providing elaborate methods of outing out, rights of users and the personal information collected for the purposes (Annexure 4.2).



Annexure 4.1- LinkedIn's FAQ in reference with GDPR[65]

[65] 'Learning FAQ: General Data Protection Regulation (GDPR) and Learning' (*LinkedIn*) <https://www.linkedin.com/help/learning/answer/a700860> last accessed 20 April 2026



Annexure 4.2- LinkedIn's policy vis-à-vis CCPA[66]

Through these policies and compliances of Social Media Intermediaries with State-specific laws, it is inferred that with effective measure of penalisation and official overlook, breaches of data protection can be checked. Recently, major Social Media intermediaries have updated their policies of privacy and web scraping to ensure that any legal liability is not incurred while their data scraping mechanisms remain robust, legitimate and secure. Such obligations have been fulfilled in the EU and California in addition to other states however, India lacks a robust mechanism like that of EU, California and Peoples' Republic of China. The General Data Protection Regulation of the European Union, Personal Information Protection Law (PIPL) of People's Republic of China and the California Consumer Privacy Act (CCPA) of the United States together represent three of the most exhaustive data protection frameworks in the world. PIPL stands out from other laws because of its unique features such as extra territorial applicability, mandatory appointment of Data Protection Officers, stringent impact assessment, and severe penalties including criminal liability, placing individuals consent and autonomy at its core. The key provisions of the GDPR are age specific provisions, liability of controllers with respect to demonstration of consent and compliance, impact assessments and Data Protection Officers. The CCPA, on the other hand provides for a more consumer centric protective measure through its broad definition of personal information, the Delete Act and

[66] ' California Consumer Privacy Act Notice for California Consumers'(*LinkedIn*) https://www.linkedin.com/legal/california-privacy-disclosure> last accessed 20 April 2026

DROP platform for streamlined data deletion, stricter proactive measures for minor's data through opt in consent model, and the establishment of a dedicated California Consumer Regulatory Agency for enforcement.

The Digital Personal Data Protection Act of 2023 of India is currently only partly enforced, restricted in scope, passe in compatibility with international frameworks and not equally complied with by the intermediaries due to absence of strict compliance mechanisms. The web scraping methods and the increasing need for datasets call for major reforms in the Digital Personal Data Protection Act, 2023 with a stricter enforcement of IndiaAI Guidelines to keep a check on scraping of personal data both publicly available and unavailable. Therefore, the above models of the PIPL, GDPR and CCPA provide robust and efficient methods and regulations that the DPDP Act can enumerate for a more effective, exhaustive and efficient data protection safeguards.

CONCLUSION

The research demonstrates that the web scraping practices are prevalent among intermediaries like X, Meta and Reddit. The platforms collect and utilize vast datasets that intercept posted content, encrypted conversations, saved drafts and location and camera access. The same is utilized for development and research purposes with cases of licensing users' personal information to third party advertisers as well. Their policies of data retention intervene the data protection rights and rights to digital autonomy and informational privacy of users. Moreover, the data collected for training AI purposes ranges from non-personal to sensitive and is vulnerable to be permanent stored in such programmes. The study confirms that these policies operate differently in different jurisdictions in accordance with local Data Protection Safeguards with explicit protection and exemptions provided to residents of the European Union and California. Peoples' Republic of China provides another efficient statutory safeguard under the Personal Information Protection Law with protocols including impact assessments, criminal liability and extraterritorial nature. Such protocols effectively protect the users' data. On the other hand, India's Digital Personal Data Protection Act, 2023 is limited in scope for the exclusion of publicly available data, part enforceability and restrictions in content removal because of stringent prior consent requirements. Therefore, the paper has highlighted key provisions such as extraterritoriality, appropriate penalties, both punitive and criminal, and added liability of data processors, of GDPR, CCPA and PIPL that can be effectively included and practiced within the DPDP Act, 2023 in addition to the Information Technology Act, 2000.