Data Ethics Report

# “Global Data-Privacy Regulations: History, Development, and Implications for Machine Learning Applications.”

## Intro:

On May 22, 2023, the Meta Platforms received the largest GDPR (General Data Protection Regulation) fine in history: € 1.2 billion for transferring European user data to US servers. This violation was not an accident or malicious action, it was simply moving data across the Atlantic to power Facebook’s advertising algorithms and recommendation systems. This routine engineering decision, on which Meta, and any other such company, bases their Machine Learning infrastructure. However, this transatlantic data transfer had become illegal after the 2020 ‘Schrems II’, a ruling by the Court of Justice of the European Union.

The fine represented about 4% of Meta’s annual revenue and Meta was given six months to stop transferring EU user data to the US - a deadline that would require rebuilding fundamental infrastructure that powers their machine learning pipelines. The company ultimately invested billions in European data centers and restructured its entire data architecture to maintain separate training pipelines for different jurisdictions. Cases like these aren’t unusual: Meta had already paid over €900 million in previous GDPR fines. Amazon, Google, and TikTok have collectively paid hundreds of millions more. Yet violations continue, not from negligence, but because they represent a structural conflict between how tech companies monetize user data and how societies now demand that data be protected.

These enforcement actions represent more than regulatory penalties - they mark a global reckoning with how personal data has been collected, analyzed, and monetized over the past two decades. The General Data Protection Regulation (GDPR, 2018), California Consumer Privacy Act (CCPA, 2020), and Personal Information Protection Law (PIPL, 2021) emerged from different jurisdictions with distinct priorities, yet they converge on a common principle: individuals should have meaningful control over their personal information. This report examines the history and development of these three regulatory frameworks, and explores their profound implications for three key stakeholders: developers who build ML systems, users whose data powers them, and companies whose business models depend on them. Understanding this regulatory landscape is essential for anyone working at the intersection of machine learning and society, where technical capability, economic incentive, and ethical responsibility increasingly collide.

**Regulatory landscape**:

The data-privacy regulations we have today are the result of development over several decades. After World War II, the European Convention on Human Rights (ECHR) recognised privacy as a fundamental human right, according to the Autoriteit Persoonsgegevens. Not long after this, European countries like Germany (1977) and France (1978) started to adopt data privacy regulations like the Hesse Data Protection Act (1970). This law introduced the protection of personal data from being misused by government agencies and is considered as the first modern privacy regulation. These efforts in protecting data from being abused by public administrations inspired broader international guidelines including the OECD's Recommendations (1980) and the Council of Europe's Convention 108 (1981) ("Sources of Data Protection Law," n.d.).

This early wave of state regulation eventually culminated in the EU's Data Protection Directive of 1995, which laid the foundation for the *General Data Protection Regulation* (GDPR) which entered into force on 25 May 2018. This law greatly strengthened data protection, especially through the "right to be forgotten" (RTBF), and introduced substantially higher penalties, which can now reach up to 4% of global annual turnover or 20 million EUR, whichever is greater. After the wake of GDPR, new legislation rapidly emerged globally, as for example the *California Consumer Privacy Act* (CCPA) which was approved in June 2018 and took effect in 2020. Similarly, China created the *Personal Information Protection Law* (PIPL) in August 2021 which became effective in November 2021.

Today these are three main global frameworks that dominate data privacy: GDPR, CCPA, and PIPL, each reflecting different legal and political philosophies. According to *eSecurity Planet*, the GDPR represents a rights-based approach centred on strict opt-in consent. In contrast, the CCPA follows a consumer-protection model in which organizations may use personal data unless a user actively opts out (eSecurity Planet, n.d.). The CCPA gives users the right to correct inaccurate personal information and limit the use of sensitive personal information, which may include precise geolocation data and security codes. Meanwhile, China's PIPL takes a different viewpoint, placing emphasis on data sovereignty which requires strict control over sensitive data and government assessments for cross-border transfers (Cooley, 2022). A comparative study published on ResearchGate shows that PIPL establishes the most strict location requirements, whereas CCPA is the least stringent when it comes to international transfers (ResearchGate, n.d). These differences reflect broader political objectives: EU focuses on fundamental rights, California focuses on consumer choice, and China prioritises national security and state oversight.

**Impact on ML Development:**

The evolution of global data-privacy regulations has fundamentally reshaped how machine learning (ML) systems are designed, deployed, and maintained. Technical challenges arise immediately when laws like GDPR, CCPA, and PIPL grant individuals rights over their

personal data. ML models learn patterns from large datasets containing information about individuals, such as browsing history, purchases, location etc. Laws that give users rights over their data, such as the ***right to deletion***, create significant challenges. Once a user requests their data to be removed, engineers cannot simply delete the raw dataset because the model may have already incorporated that information. To address this, teams must use **machine unlearning** techniques or retrain models entirely, which can be computationally intensive and operationally costly.

Another critical requirement is **data lineage tracking**, which involves recording how each data point flows through the system, from collection, preprocessing, and model training to final predictions. This ensures that requests to access, correct, or delete personal data can be fully implemented. Regulations also increasingly require **explainability**, meaning ML systems must provide interpretable outputs that explain how decisions are made. This is particularly important in high-stakes applications, such as loan approval, healthcare recommendations, or hiring decisions, where users and regulators need clear justification for automated decisions

These regulations also transform the **ML development lifecycle**. **Privacy by design** requires teams to consider privacy from the earliest stages of planning and system architecture. **Data minimization** limits the amount of personal data collected, encouraging the use of aggregated, synthetic, or anonymized datasets. **Purpose limitation** ensures that data cannot be reused for new ML models without explicit consent, limiting the flexibility that developers previously had to reuse large datasets across projects.

Real-world examples illustrate these impacts. After receiving a €1.2 billion GDPR fine in 2023 for transferring European user data to the U.S., Meta restructured its ML infrastructure to separate European and non-European data, implement automated deletion workflows, and redesign recommendation algorithms to comply with privacy rules. Similarly, Google implemented **differential privacy** and on-device computation in products like location history and analytics tools, allowing ML models to learn population-level patterns without exposing individual identities. These examples show that privacy regulations are not merely legal obligations, they are **technical design constraints** that fundamentally influence how modern ML systems operate, while also embedding ethical considerations into engineering practices.

**The Idea of Consent:**

To illustrate the different ideas behind the frameworks it makes sense to look into the aspect of consent in data processing and how the different laws approach it.

Under the GDPR consent is the central tool and must be given before any data processing happens. Consent is defined as follows in Article 4 of the GDPR (2018):

*"any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her."*

The framework also states that there always has to be a clear option to refuse and a separate consent for each data processing operation. This also applies when the same data is used just for a different process. The data subject also has to be clearly informed about the data processing operations that are going to be done with the data as well as have the ability to withdraw the consent. Additionally, the identity of who is using the data must be made known by the company (European Data Protection Board, 2020). This clearly reflects the European focus on a rights-based approach with opt-in consent.

The Chinese PIPL framework has a quite similar model. It also requires individual consent given freely, voluntarily and based on full information. A new consent must be given, when the purpose of data processing changes. Separate consent must also be given before personal information is transferred out of the country and when sensitive personal information is processed (Personal Information Protection Law, 2021; China Briefing, 2023). This illustrates the focus on national data sovereignty.

The CCPA approaches consent in a different way than the other two legal frameworks. Its main focus is on the consumer's right, which the consumer themself has to act on, such as the right to know, correct, delete one's data and opt-out of data selling and sharing. Generally, the consumer does not have to give consent before a business uses his data. This differs for minors. Data of under 16 year olds is not allowed to be sold or shared unless the minor themself gives consent. For minors under 13 the consent has to be given by the consumer's parent or guardian (California Office of the Attorney General, 2024; TermsFeed, 2023). The focus on opt-out rather than opt-in presents the idea of consumer choice and that rights need to be claimed.

**Conclusion and Future Implications:**

The growing pressure by regulatory frameworks such as the GDPR, PIPL and CCPA illustrate the tension between using data as a tool and the need for data protection. Machine learning systems require large amounts of data (often being transferred over national borders) to thrive. They have to be built differently as shown by the restructuring of Meta's ML infrastructure. Another problem for developers which will intensify in the near future is the requirement for explainability, as deep neural networks or LLMs are often too complex to even be understood.

As the importance of data and the international competition keeps growing, regulatory frameworks are going to be even more important than they are right now. This is why as a ML practitioner it is important to understand the laws, in order to stay within them and find creative solutions in how to make ML models work well within those constraints. Balancing technical performance and regulatory responsibility will be important in the development of future ML models.