

How the Legal Basis for AI Training is Framed in Data Protection Guidelines and Interventions: Comparative Perspectives and the Prospect of Global Convergence

Wenlong Li*, Yueming Zhang[†], Qingqing Zheng[‡], and Aolan Li**

Abstract

- This article investigates how the legal basis for artificial intelligence (AI) training is framed within data protection guidelines and regulatory interventions, drawing on a comparative analysis of approaches taken by authorities across multiple jurisdictions.
- Focusing on the European Union's General Data Protection Regulation and analogous data protection frameworks globally, the study systematically maps guidance, statements, and actions to identify areas of convergence and divergence in the conceptualization and operationalization of lawful grounds for personal data processing—particularly legitimate interest and consent—in the context of AI model development.
- The analysis reveals a trend towards converging on the recognition of legitimate interest as the predominant legal basis for AI training. However, this convergence is largely superficial, as guidelines rarely resolve deeper procedural and substantive ambiguities, and enforcement interventions often default to minimal safeguards. This disconnect between regulatory rhetoric and practical compliance leaves significant gaps in protection and operational clarity for data controllers, calling into question the reliability and legitimacy of the existing framework for lawful AI training. It warns that, without clearer operational standards and more coherent cross-

border enforcement, there is a risk that legal bases such as legitimate interest will serve as little more than formalities.

- Reflecting on these findings, the article explores the prospects and limitations for achieving greater alignment or harmonization at the global level. Specifically, it reflects on pathways for global AI governance, emphasizing that the progress would benefit from distinguishing issues amenable to international convergence from those that require context-sensitive, locally adaptive solutions. It considers how regulators, practitioners, civil society activists, and scholars might leverage these insights to prioritize evidence-based avenues rather than seeking uniformity at the conceptual level for its own sake, with the ultimate goal of advancing both principled and practical frameworks for lawful AI training across diverse legal landscapes.

Introduction

Data protection law has emerged as one of the most active and consequential domains in the governance of artificial intelligence (AI). As AI-specific regulation—such as the European Union's (EU) Artificial Intelligence Act (AIA)—faces increasing uncertainty amid geopolitical tensions and growing divergences within the Union,¹ the General Data Protection Regulation (GDPR) has reasserted its centrality. As Zanfir astutely puts it, the GDPR's broad scope and foundational principles make it uniquely positioned to govern the use of AI systems,

* Wenlong Li, Research Professor, Guanghua Law School, Zhejiang University, Zhejiang, China. Email: wenlong.li@zju.edu.cn

[†] Yueming Zhang, Research Affiliate, Law and Technology Research Group, Faculty of Law & Criminology, Ghent University, Ghent, Belgium. Email: yueming.zhang@ugent.be

[‡] Qingqing Zheng, Undergraduate Student, School of Law, Shandong University, Jinan, China

** Aolan Li, Doctoral Candidate, Center for Commercial Law Studies,

Department of Law, Queen Mary University London, London, United Kingdom. Email: aolan.li@qmul.ac.uk

¹ Luca Bertuzzi, 'EU Commission Eyes Pausing AI Act's Entry into Application' (MLEX, 26 May 2025) <<https://www.mlex.com/mlex/articles/2344845/eu-commission-eyes-pausing-ai-act-s-entry-into-application>> accessed 30 August 2025.

even as these technologies outpace the regulatory frameworks designed for more conventional forms of data processing.² In many jurisdictions where fully operational AI legislation is still absent, the GDPR continues to serve as both a *de facto* regulatory baseline for scrutiny, accountability, and redress. Its principles—particularly lawfulness, transparency, and accountability—are being tested in AI training contexts that were never fully anticipated during the GDPR’s drafting.³ In this light, understanding how data protection authorities interpret and enforce data protection rules is no longer merely a compliance concern; it is a frontline issue in shaping the contours of AI governance.

Amongst the various mechanisms for ensuring that AI training complies with data protection law, the choice and application of a lawful basis for data processing has emerged as a focal point of regulatory scrutiny. This foundational requirement functions as a legal gateway: if no lawful basis can be established, the entire processing operation is rendered unlawful. In the context of AI model training, the stakes are particularly high, given the large-scale ingestion of data, the opacity of model architectures, and the difficulty in delineating clear purposes and data subjects.⁴ Across jurisdictions, lawful basis constitutes the narrowest common denominator: while national data protection authorities diverge significantly in their treatment of issues like data minimization, fairness, and retention, virtually all regulatory guidance and enforcement hinges on a foundational question—was a valid legal ground for processing established? Mapping this axis offers the clearest and most empirically tractable lens for comparison. Further, as our study will reveal, a lawful basis does not operate in isolation; rather, it is inherently interlinked with other foundational principles and concepts such as minimization, transparency, special-category safeguards, accuracy, and accountability, particularly when legitimate

interest is relied upon for AI training. Thus, foregrounding the lawful basis for data processing enables a jurisdiction-agnostic, conceptually focused, and analytically coherent treatment of what remains the most decisive—and most contested—regulatory chokepoint for AI model training.

The debate on the lawful grounds for AI training under EU data protection law has increasingly converged on a key practical observation: the range of lawful bases available has, in effect, narrowed to a binary between consent and legitimate interest for private-sector AI developers. Legal scholars have questioned not only the feasibility of traditional bases like consent and contract, but more often the elasticity and risk-management capacity of legitimate interest in light of AI’s evolving capabilities. On the one hand, consent is normatively appealing for its emphasis on autonomy, but widely recognized as difficult to operationalize in large-scale training pipelines, particularly in the context of third-party data reuse or web-scraped content. For instance, Zhang and others note that consent is often practically unmanageable, and its withdrawal risks cascading compliance challenges.⁵ Kuru, drawing on the CJEU’s *Meta* judgment, argues that consent may be the only permissible ground in cases involving sensitive data, effectively ruling out other bases for large-scale model training.⁶ Craddock has argued that the GDPR does not establish a formal hierarchy among lawful bases, suggesting that each ground must be evaluated contextually.⁷ This position is echoed by the European Data Protection Board (EDPB), which has affirmed that the selection of a lawful basis must be based on the specific purpose and circumstances of processing, not on abstract preferences.⁸ Stalla-Bourdillon and Kramcsák have advanced a more adaptive model, advocating for a blended or layered use of multiple lawful bases throughout the lifecycle of AI systems (ie, training, fine-tuning, and deployment).⁹ On

- 2 Gabriela Zafir-Fortuna, ‘How Data Protection Authorities Are De Facto Regulating Generative AI’ (*Future Privacy Forum*, 13 September 2023) <<https://fpf.org/blog/how-data-protection-authorities-are-de-facto-regulating-generative-ai/>> accessed 24 January 2025.
- 3 Giovanni Sartor and Francesca Lagioia, *The Impact of the General Data Protection Regulation on Artificial Intelligence* (European Parliament 2020) <<https://data.europa.eu/doi/10.2861/293>> accessed 30 August 2025. See also, Lilian Mitrou, ‘Data Protection, Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR) “Artificial Intelligence-Proof?”’ (Social Science Research Network, 31 December 2018) <<https://papers.ssrn.com/abstract=3386914>> accessed 30 August 2025.
- 4 Laura Weidinger and others, ‘Taxonomy of Risks Posed by Language Models’, *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) <<https://dl.acm.org/doi/10.1145/3531146.3533088>> accessed 30 August 2025.
- 5 Dawen Zhang and others, ‘Privacy and Copyright Protection in Generative AI: A Lifecycle Perspective’, *IEEE/ACM 3rd International Conference on AI Engineering—Software Engineering for AI* (2024)

- <<https://ieeexplore.ieee.org/document/10556135>> accessed 30 August 2025.
- 6 Taner Kuru, ‘Lawfulness of the Mass Processing of Publicly Accessible Online Data to Train Large Language Models’ (2024) 14 *International Data Privacy Law* 326.
- 7 Peter Craddock, ‘Op-Ed: AI Training Data = (Non-)Personal Data? And Is Consent Really Relevant?’ (*LinkedIn*, 14 October 2024) <<https://www.linkedin.com/pulse/op-ed-ai-training-data-non-personal-consent-really-peter-craddock-ko4ie/>> accessed 21 January 2025.
- 8 European Data Protection Board, ‘Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models’ (18 December 2024) <https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en> accessed 24 January 2025.
- 9 Sophie Stalla-Bourdillon and Pablo Trigo Kramcsák, ‘ChatGPT and Lawful Bases for Training AI: A Blended Approach?’ (*The Digital Constitutional*, 18 July 2023) <<https://digi-con.org/chatgpt-and-lawful-bases-for-training-ai-a-blended-approach/>> accessed 24 January 2025.

the other, the uncertainty and complexity of operationalizing a consent-based regime has led many scholars to view legitimate interest as a potentially more flexible foundation. It offers a structured—but not unconditional—pathway to justify processing, requiring assessments of necessity, proportionality, and the implementation of mitigating measures in operational contexts. Kramcsák sees potential for legitimate interest to support AI training if reinforced with robust safeguards such as pseudonymization and strong documentation.¹⁰ However, others, including Etienne and others, point to practical difficulties of fulfilling these conditions, particularly in achieving effective pseudonymization or maintaining sufficient transparency over complex training pipelines.¹¹ These debates are complicated by the ongoing uncertainty over whether personal data are processed in a legally meaningful way during AI training at all. Some argue that tokenization, aggregation, or semantic abstraction may sever identifiability;¹² yet, audits of large-scale training datasets suggest otherwise that personal data—oftentimes including sensitive data—are not only involved but often scraped indiscriminately from public web sources.¹³ Thus, radical claims that such training is categorically outside the scope of the GDPR appear increasingly untenable. Beyond the European context, the global landscape destabilizes any assumption of consensus. Not all jurisdictions recognize legitimate interest as a lawful basis, leading to increasing calls for more contextual approaches. For instance, Canada's PIPEDA does not include it, though Bill C-27 may introduce a GDPR-influenced version.¹⁴ China's PIPL similarly avoids the concept yet relies on a reconstructed 'manifestly public' exception that may serve a comparable function in legitimizing AI training.¹⁵

While theoretical and conceptual debates proliferate, the interpretive positions adopted by authorities have yet to be systematically mapped or critically analysed. Moreover, existing work often centres on European legal contexts, leaving global enforcement dynamics and pathways to international alignment underexplored. This article seeks to address that gap by offering a

comparative, empirically grounded analysis of how data protection authorities worldwide frame the lawful basis for AI training. Through a meticulous mapping of related guidelines and interventions, we aim not only to illuminate enforcement trends and interpretative divergences but anticipate the emerging trajectory of international alignment—revealing how legal bases are beginning to be translated into actionable compliance standards, and what these signals mean for the future of global AI governance.

Specifically, the analysis is organized around two core components: (i) a review of 19 guidelines issued between 2020 and 2024 by various authorities (see Table 1), each of which explicitly addresses the lawful basis for processing personal data in the context of AI training; and (ii) an examination of regulatory interventions taken against AI providers worldwide, focusing specifically on public enforcement—that is complaint-driven or *ex officio* investigations conducted by data protection authorities. Given the highly dynamic and evolving nature of this regulatory landscape, only cases exhibiting a degree of procedural maturity—that is those where interim or final decisions have been issued—and available at the time of writing are included. It should be noted that there are parallel developments in private enforcement and judicial proceedings, which are duly acknowledged but not the primary focus of this article.

This article advances three core arguments. First, the framing of the legal basis for AI training is marked by a formal convergence on legitimate interest as the principal ground for processing personal data, but this apparent consensus conceals significant divergences in substantive interpretation, procedural requirements, and compliance practices upon close examination. Although regulatory guidelines often endorse legitimate interest—and, less frequently, consent or other necessity-based bases—they rarely offer precise or harmonized operational thresholds for their application in complex and evolving AI environments. Secondly, such interpretive ambiguity is compounded by uneven and sometimes minimalist enforcement, leading to a

10 Pablo Trigo Kramcsák, 'Can Legitimate Interest Be an Appropriate Lawful Basis for Processing Artificial Intelligence Training Datasets?' (2023) 48 Computer Law & Security Review 105765.

11 Etienne Drouard and others, 'The Interplay between the AI Act and the GDPR' (2024) 1 Journal of AI Law and Regulation 164.

12 Lokke Moeral and Marijn Storm, 'Do LLMs "store" Personal Data? This Is Asking the Wrong Question' (IAPP, 23 October 2024) <<https://iapp.org/news/a/do-llms-store-personal-data-this-is-asking-the-wrong-question>> accessed 24 January 2025.

13 For instance, see Abeba Birhane and others, 'Into the LAIONs Den: Investigating Hate in Multimodal Datasets' (arXiv, 6 November 2023) <<http://arxiv.org/abs/2311.03449>> accessed 30 August 2025. See also Stefan Baack, 'A Critical Analysis of the Largest Source for Generative AI

Training Data: Common Crawl', *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2024) <<https://doi.org/10.1145/3630106.3659033>> accessed 30 August 2025.

14 David TS Fraser, 'What Is the "Legitimate Interests" Exception to Consent under Canada's Proposed Privacy Law?' (*Canadian Privacy Law Blog*, 18 November 2023) <<https://blog.privacylawyer.ca/2023/11/what-is-legitimate-interests-exception.html>> accessed 30 August 2025.

15 Wenlong Li and Jiahong Chen, 'From Brussels Effect to Gravity Assists: Understanding the Evolution of the GDPR-Inspired Personal Information Protection Law in China' (2024) 54 Computer Law & Security Review 105994.

Table 1. Guidelines issued by data protection authorities worldwide explicitly related to lawful basis for AI training.

No	Jurisdiction	Authority	Title	Issue Date	Type	Context
1	EU	EDPB	Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models	17 December 2024	Opinion	AI
2	EU	EDPB	Report of the work undertaken by the ChatGPT Taskforce	23 May 2024	Report	ChatGPT Taskforce
3	EU	EDPS	Generative AI and the EUDPR	3 June 2024	Orientations	Generative AI
4	France	CNIL	AI: Mobilizing the legal basis of legitimate interest to develop an AI system	10 June 2024	Guidance	Web scraping
5	France	CNIL	The legal basis of legitimate interest: focus sheet on the measures to be taken in the event of data collection by harvesting (web scraping)	10 June 2024	Guidance	Web scraping
6	UK	ICO	ICO Guidance on AI and other products	28 October 2024	Guidance	AI
7	UK	ICO	The lawful basis for web scraping to train generative AI models	15 January 2024	Guidance	Web scraping
8	UK	ICO	Information Commissioner's Office response to the consultation series on generative AI	18 December 2024	Guidance	Generative AI
9	Germany	LfDI	Discussion paper on legal basis in data protection for the user of AI (Version 2.0)	17 October 2024	Discussion paper	AI
10	Germany	BayLDA	AI and Data Protection	24 January 2024	Blogpost	AI
11	Germany	DSK	AI and data protection	6 May 2024	Orientations	AI
12	Spain	AEPD	GDPR compliance of processes that embed Artificial Intelligence: An introduction	February 2020	Orientations	AI

Continued

Table 1. *Continued*

No	Jurisdiction	Authority	Title	Issue Date	Type	Context
13	Belgium	Gegevensbeschermingsautoriteit	Artificial Intelligence Systems and the GDPR: A Data Protection Perspective	December 2024	Information note	AI
14	Liechtenstein	Datenschutzstelle	Guidelines on data protection requirements for chatbots	November 2023	Blogpost	Chatbots
15	Singapore	PDPC	Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems	1 Mar 2024	Guidance	AI recommendation and decision systems
16	ASEAN	ASEAN	ASEAN Guide on AI Governance and Ethics	2 February 2024	Guidance	AI
17	USA	NIST	AI Risk Management Framework: Generative Artificial Intelligence Profile	July 2024	Report	AI
18	Australia	OAIC	Guidance on privacy and developing and training generative AI models	21 October 2024	Guidance	AI
19	Australia	OAIC	Guidance on privacy and the use of commercially available AI products	21 October 2024	Guidance	Commercially available AI products

fragmented enforcement and compliance landscape. Despite the proliferation of guidelines and interventions, the legal basis for AI training remains underdefined and unstable, exposing global AI governance to jurisdictional variability. The article finally assesses the possibility and limits of international harmonization. By mapping patterns of emerging convergence, potential alignment, and areas of potential divergence, we reveal a gradient landscape that highlights opportunities for future developments as well as persistent divides. Guidance produced in this sporadic, collaborative environment might incrementally advance interoperability as authorities ‘communicate’ and learn from each other’s practice, but a more systematic approach is needed in the long run. Authorities should move beyond abstract principles, systematically sharing enforcement experiences, audit methodologies, and compliance resolutions, and use these insights to develop iterative, credible, and operational guidance—especially in areas where global models and datasets cut across legal boundaries. If international efforts are limited to only the lowest-common-denominator procedural safeguards, the result may be hollow protections and persistent inequities in AI development and governance.¹⁶

This article is structured into six sections. Following the introduction, ‘Methods’ section outlines the methodological framework used to identify, code, and analyse regulatory materials. This section details the scope of the dataset, search strategies, and classification criteria, with coding details provided in [Supplementary Appendix I](#). ‘Guidelines’ section presents a comparative synthesis of 19 regulatory guidelines issued between 2020 and 2024. It examines how data protection authorities conceptualize lawful bases for AI training, particularly the preferential framing of *legitimate interest* and the marginalization of *consent*. Findings are presented in both narrative and tabular form to enhance clarity. ‘Interventions’ section turns to enforcement practice, analysing real-world regulatory actions that engage with lawful basis claims in the context of AI training. This section serves as a counterpart to the guideline analysis, illuminating the gap between high-level principles and practical decision-making. ‘Reflection’ section critically assesses the prospects and challenges of global convergence, highlighting where alignment appears reasonably achievable and where persistent divergence remains. Finally, ‘Conclusion’ section distils the comparative

insights into overarching observations on the future trajectory of international AI governance.

Methods

This article adopts a comparative qualitative approach to analyse how data protection authorities across jurisdictions interpret and operationalize lawful basis for data processing in relation to AI model training. The core of our analysis is grounded in a curated dataset comprising 19 regulatory guidelines that directly and substantively engaged with the lawful basis for AI training, coupled with enforcement interventions of the same nature that have reached a certain level of procedural maturity by the time of writing. The datasets were assembled manually in January 2025 using a combination of legal databases (notably OneTrust DataGuidance), official websites of data protection authorities, and announcements or document repositories shared via regulators’ official websites and social media channels (eg LinkedIn and X). The process proceeded in three stages:

1. We compiled a list of data protection authorities from major jurisdictions, including all EU Member States (covering 46 DPAs in total), as well as authorities in the UK, USA, Canada, Brazil, and key Asian jurisdictions such as members of the ASEAN bloc.
2. We conducted a manual review of each authority’s official website to identify any published guidelines, policy papers, or standards addressing AI and privacy.
3. To ensure completeness and avoid jurisdictional omissions, we cross-referenced our findings with third-party legal resources, most notably OneTrust DataGuidance.

To ensure coverage, we used targeted keyword searches including: ‘AI training’ OR ‘machine learning’ OR ‘foundation models’ AND ‘lawful basis’ OR ‘legal ground’ OR ‘legitimate interest’ OR ‘consent’ AND ‘data protection’ OR ‘GDPR’ OR ‘personal data’ AND ‘regulator’ OR ‘authority’ OR ‘guidance’ OR ‘enforcement’. This process yielded an initial pool of 29 relevant documents published between 2020 and 2024. We then applied an inclusion criterion, retaining only those documents that made substantive reference to lawful grounds for data processing in the context of AI training. Ten guidelines were excluded on the basis that

Table 2. A snapshot of the enforcement actions against generative AI at global level.

Authority	Date	Controller	Stakeholder	Framework
Garante, Italy	20 December 2024	OpenAI	/	GDPR
Garante, Italy	29 November 2024	GEDi group	/	GDPR
ANPD, Brazil	2 July 2024	Meta	/	LGPD
ANPD, Brazil	17 December 2024	X	/	LGPD
11 DPAs: Austria, Belgium, France, Germany, Greece, Italy, Ireland, the Netherlands, Norway (with NCC), Poland and Spain	14 June 2024	Meta	noyb	GDPR
DPC, Ireland	8 August 2024	X (Grok)	/	GDPR
9 DPAs: Austria, Belgium, France, Greece, Ireland, Italy, the Netherlands, Spain and Poland	12 August 2024	X (Grok)	noyb	GDPR
Garante, Italy	30 January 2025	DeepSeek	/	GDPR
Berlin Commissioner, Germany	27 June 2025	DeepSeek	/	GDPR, DSA
PIPC, South Korea	7 February 2025	DeepSeek	/	PIPA
PDPC, Hong Kong	25 October 2023	LinkedIn	/	PDPA
DSB, Austria	29 April 2024	OpenAI	noyb	GDPR

they addressed AI and privacy only in general or ethical terms, without discussing legal bases under Article 6 GDPR or comparable provisions. Our final sample thus comprised 19 documents, which were comparatively analysed and synthesized (see Table 3).

Each guideline was reviewed in full and coded using a structured analytic framework developed during the project. The coding captured both legal content and procedural treatment, focusing on how data protection authorities frame and operationalize lawful bases in the context of AI model training. The outputs were thematically grouped and compared across jurisdictions to identify convergence, divergence, and interpretive ambiguity. A full explanation of coding variables and classification decisions is provided in Appendix 1.

In parallel, we analysed a sample of regulatory interventions (see Table 2), including enforcement actions, investigations, and formal decisions that addressed lawful basis questions in *ad hoc*, contexts. These materials were collected using the same jurisdictions and databases. Given the lack of standardization in disclosure practices, we included only cases where: (i) a public-facing document was available and (ii) the question of lawful basis was addressed either directly in the reasoning or indirectly through remedial orders and compliance conditions. We excluded pending investigations without conclusive or interim findings by the time of writing.

This study is limited by its reliance on publicly available documentation, which may not reflect internal regulatory deliberations as well as informal guidance. Although the mapping exercise was systematic and comprehensive at the time of completion, the regulatory environment surrounding AI training continues to evolve rapidly as the intersection of data protection and generative AI remains a moving target. A further limitation concerns the opacity surrounding actual compliance practices. When evaluating enforcement interventions—and particularly the divergence between regulatory guidance and on-the-ground enforcement—this study is constrained by the limited availability of providers' internal documentation. As a result, our evaluation of the level and quality of assessments conducted by AI providers is necessarily indirect, relying on what regulators have chosen to publish.

Guidelines and other statements

Guidance issued by regulatory authorities plays a crucial role in shaping the interpretation and application of data protection laws. These guidelines are not merely explanatory; they serve as a dynamic vehicle for authorities to articulate their viewpoints, foster alignment across jurisdictions, and respond to emerging technological developments.¹⁷ In theory, such documents facilitate cross-border coordination, enabling regulators to adopt common approaches to shared challenges while

Table 3. Mapping of critical aspects of legitimate interest as discussed in the guideline.

No.	Jurisdiction	Authority	Full Title	Type	Context	Lawful Grounds mentioned	Recognition of the suitability of consent to AI training	Expressed concerns for consent	Recognition of the relevance of Legitimate Interest to AI training	List-based approach	Exclusion of legitimate approach	Necessity	Reasonable expectation	Risk classification	Position on opt-out	Position on PETs
1	EU	EDPB	Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models	Opinion	AI	Legitimate interests	Not relevant	Not relevant	Yes	Yes	Not mentioned	Assessment required	Yes	Not mentioned	Recommended	Recommended
2	EU	EDPB	Report of the work undertaken by the ChatGPT Taskforce	Report	ChatGPT Taskforce	Legitimate interests, Consent, Necessity for contract	Unclear	Not mentioned	Yes	No	Not mentioned	Assessment required	Yes	Not mentioned	Not mentioned	Not mentioned
3	EU	EDPS	Generative AI and the EUDPR	Orientations	Generative AI	Legitimate interests, Consent	Yes	Valid consent, withdrawal of consent	Yes	No	Not mentioned	Assessment required	Not mentioned	Not mentioned	Required	Not mentioned
4	France	CNIL	AI: Mobilizing the legal basis of legitimate interest to develop an AI system	Guidance	AI	Legitimate interests	Not relevant	Not relevant	Yes	Yes	DSA Article 28	Assessment required	Yes	Case-by-case analysis	Recommended	Recommended
5	France	CNIL	The legal basis of legitimate interest: focus sheet on the measures to be taken in the event of data collection by harvesting (web scraping)	Guidance	Web scraping	Legitimate interests	Not relevant	Not relevant	Yes	No	Not mentioned	Not mentioned	Not mentioned	Not mentioned	Not mentioned	Not mentioned
6	UK	ICO	ICO Guidance on AI and other products	Guidance	AI	Legitimate interests, Consent, Necessity for contract, Public interests, Legal Obligation	Yes	Valid consent, withdrawal of consent	Yes	No	Not mentioned	Assessment required	Not mentioned	Not mentioned	Not mentioned	Not mentioned
7	UK	ICO	The lawful basis for web scraping to train generative AI models	Guidance	Web scraping	Legitimate interests	Regularly not	Not mentioned	Yes	No	Not mentioned	Unclear	Not mentioned	High	Not mentioned	Recommended
8	UK	ICO	Commissioner's Office response to the consultation series on generative AI	Guidance	Generative AI	Legitimate interests	Regularly not	AI model has no direct relationship with the person whose data are scraped	Yes	Unclear	Not mentioned	Unclear	Unclear	Not mentioned	Not mentioned	Unclear
9	Germany	LfDI	Discussion paper on legal basis in data protection for the user of AI	Discussion paper	AI	Legitimate interests, Consent, Necessity for contract, Public interests, Vital interest	Regularly not	The data subject is often not known in such processing situations and a necessary declaration of consent cannot be obtained; withdraw of consent	Yes	Yes	AIA Article 5	Assessment required	Yes	Case-by-case analysis	Not mentioned	Recommended
10	Germany	Bavarian	AI and Data Protection	Blog	AI	Legitimate interests	Unclear	Not mentioned	Yes	No	Not mentioned	Not mentioned	Not mentioned	Case-by-case analysis	Unclear	Recommended

Continued

Table 3. Continued

No.	Jurisdiction	Authority	Full Title	Type	Context	Lawful Grounds mentioned	Recognition of the suitability of consent to AI training	Expressed concerns for consent	Recognition of the relevance of legitimate Interest to AI training	List-based approach	Exclusion of legitimate approach	Necessity	Reasonable expectation	Risk classification	Position on opt-out	Position on PETs
11	Germany	DSK	Artificial intelligence and data protection	Orientations	AI	Various legal bases	Regularly not	The data subject is often not known in such processing situations and a necessary declaration of consent cannot be obtained; withdraw of consent	Yes	Yes	ALA Article 5	Assessment required	Yes	Case-by-case analysis	Not mentioned	Recommended
12	Spain	AEPD	GDPR compliance of processes that embed Artificial Intelligence: An introduction	Orientations	AI	Legitimate interests, Consent, Necessity for contract, Public interests, Vital interest, Legal Obligation	Yes	Not mentioned	Yes	No	Not mentioned	Not mentioned	Not mentioned	Case-by-case analysis	Recommended	Not mentioned
13	Belgium	Gegevensbeschermingsautoriteit	Artificial Intelligence Systems and the GDPR: A Data Protection Perspective	Information note	AI	Legitimate interests, Consent, Necessity for contract, Public interests, Vital interest, Legal Obligation	Yes	Not mentioned	Yes	No	ALA Article 5	Not mentioned	Not mentioned	Case-by-case analysis	Required	Not mentioned
14	Liechtenstein	Datenschutzstelle	Guidelines on data protection requirements for chatbots	Blog	Chatbots	Legitimate interests, Consent, Necessity for contract	Yes	Not mentioned	Yes	No	Not mentioned	Not mentioned	Yes	Not mentioned	Not mentioned	Not mentioned
15	Singapore	PDPC	Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems	Guidance	AI Recommendation and Decision Systems	Legitimate interests, Consent	Yes	Not mentioned	Yes	No	Not mentioned	Not mentioned	Not mentioned	Case-by-case analysis	Not mentioned	Recommended
16	ASEAN	ASEAN	ASEAN Guide on AI Governance and Ethics	Guidance	AI	Unclear	Yes	Not mentioned	Unclear	Not relevant	Not mentioned	Not mentioned	Not mentioned	Case-by-case analysis	Not mentioned	Recommended
17	USA	NIST	AI Risk Management Framework: Generative Artificial Intelligence Profile	Report	AI	Consent	Yes	Not mentioned	Unclear	Not relevant	Not mentioned	Not mentioned	Not mentioned	Case-by-case analysis	Not mentioned	Recommended
18	Australia	OAIC	Guidance on privacy and developing and training generative AI models	Guidance	Generative AI	Unclear	Yes	Not mentioned	Unclear	Not relevant	Not mentioned	Not mentioned	Yes	Case-by-case analysis	Recommended	Recommended
19	Australia	OAIC	Guidance on privacy and the use of commercially available AI products	Guidance	Commercially available AI products	Unclear	Yes	Not mentioned	Unclear	Not relevant	Not mentioned	Not mentioned	Yes	Case-by-case analysis	Recommended	Not mentioned

accounting for jurisdiction-specific nuances. As manifested by our sample of 19 guidelines, the current landscape reflects significant jurisdictional diversity, with a substantial number of documents issued across regions, each tailored to address specific facets of AI governance. We observe a wide spectrum of engagement from authorities in Europe, Asia, Australia, and the Americas, and European authorities dominate this space. Notably, the Irish Data Protection Commission (DPC), often positioned as a leading supervisory authority, holds a strategic role but has demonstrated a relatively cautious and limited engagement. Smaller states or regions with less resourceful authorities remain less engaged in this discourse, likely due to constrained resources or prioritization of other issues. Beyond Europe, we observe proactive contributions from authorities in Brazil (ANPD), Australia (OAIC), Singapore (PDPC), and the post-Brexit UK (ICO), showcasing a global appetite for confronting the challenges posed by AI.

The level of formality of these guidelines varies significantly. France's CNIL and the UK's ICO, for instance, provide detailed and actionable formal guidelines, while some authorities, such as Germany's Bavarian DPA, contribute less formal outputs, such as blogs. This variation highlights the evolving and tentative nature of this regulatory space, where many authorities are keen to influence the debate but may lack the capacity or clarity to provide enforceable or highly actionable insights. The focus of these guidelines also demonstrates varying focuses. While some documents broadly address AI governance, others narrow their scope to specific applications, for example, generative AI, chatbots, recommender systems, or web scraping.

In constructing Table 3, we focused on systematically categorizing the key dimensions of the guidance provided by various authorities in line with the coding questions abovementioned. Overall, the guidelines provided by various authorities reveal a nuanced divergence in their approaches to relying on legitimate interest for AI training. While there is a broad

consensus among authorities that legitimate interest is a relevant lawful basis for AI training, the specifics of its application are subject to varying interpretations—a complexity explored in detail in the following analysis.

Choosing between consent and legitimate interest

Most guidelines fall short of providing actionable guidance on how organizations should choose between lawful bases or when consent should be prioritized. While there is broad acceptance of legitimate interest as a viable legal basis for AI training, regulatory stances diverge significantly when it comes to the role of consent. These differences can be categorized into four modes of expression. First, authorities such as the EDPB explicitly affirm that consent retains a critical role in AI training, emphasizing its importance for safeguarding individual autonomy and protecting data subject rights.¹⁸ The EDPB has also clarified that no inherent hierarchy exists between lawful bases, instead emphasizing the need for specificity and transparency.¹⁹ Secondly, authorities such as the EDPS, ICO, and LfDI also raise significant concerns about the feasibility and practicality of consent in scenarios involving large-scale or data-intensive processing, including the difficulty of obtaining valid consent when organizations lack a direct relationship with data subjects, as well as the complexities of managing consent withdrawal in dynamic datasets.²⁰ Such concerns have driven many regulators to explore and, in some cases, endorse legitimate interest as a more practical alternative. The ICO implies, for instance, that consent may be inappropriate when there is no direct relationship with the individuals whose data are being processed,²¹ but it stops short of establishing a definitive ordering among lawful bases. Thirdly, authorities in Asia, including ASEAN and Singapore, *rhetorically* prioritize consent as the default legal basis in their guidelines, while recognizing practical use of alternative bases such as legitimate interest.²² Singapore's PDPA frames consent as the default legal basis under its Consent Obligation, with deemed consent and narrowly defined

17 Peter Hustinx, 'The Role of Data Protection Authorities' in Serge Gutwirth and others (eds), *Reinventing Data Protection?* (Springer Netherlands 2009) <https://link.springer.com/chapter/10.1007/978-1-4020-9498-9_7>.

18 European Data Protection Board (n 8).

19 *ibid.*

20 European Data Protection Supervisor, 'Generative AI and the EUDPR. First EDPS Orientations for Ensuring Data Protection Compliance When Using Generative AI Systems' (03 June 2024) <https://www.edps.europa.eu/system/files/2024-06/24-06-03_genai_orientations_en.pdf> accessed 30 August 2025; Information Commissioner's Office, 'The Lawful Basis for Web Scraping to Train Generative AI Models' (6 February 2025) <<https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/response-to-the-consultation-series-on-generative-ai/the-lawful-basis-for-web-scraping-to-train-generative-ai-models/>> accessed 22 February 2025; Landesbeauftragte für den Datenschutz und die Informationsfreiheit in Baden-Württemberg, 'Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz' (17 October 2024) <<https://www.baden-wuerttemberg.datenschutz.de/rechtsgrundlagen-datenschutz-ki/>> accessed 30 August 2025.

21 Information Commissioner's Office *ibid.*

22 ASEAN, 'ASEAN Guide on AI Governance and Ethics' (2023) <<https://asean.org/book/asean-guide-on-ai-governance-and-ethics/>> accessed 24 January 2025; PDPC, 'Advisory Guidelines on Use of Personal Data in AI Recommendation and Decision Systems' (1 March 2024) <<https://www.pdpc.gov.sg/guidelines-and-consultation/2024/02/advisory-guidelines-on-use-of-personal-data-in-ai-recommendation-and-decision-systems>> accessed 25 January 2025.

exceptions serving as alternatives.²³ This framing suggests a prioritization of consent but does not exclude the practical application of legitimate interest in appropriate circumstances. Lastly, despite the heightened legal protections required for special-category data, few guidelines explicitly address these challenges. Australia's approach is an exception, with its authority expressly requiring that sensitive data, such as photographs or recordings, generally requires explicit consent for collection and use.²⁴

Legitimate interest: identification and formulation

In formulating legitimate interest, the reviewed guidelines typically present multiple options, reflecting a fragmented and inconsistent approach across jurisdictions. Only a few authorities like France's CNIL and Germany's LfDI characteristically adopt a list-based approach, outlining specific scenarios they deem to pursue a legitimate interest.²⁵ The rationale behind these selections, however, often lacks clarity, with selective inclusions that vary in scope and specificity. Some scenarios are indeed aligned with the GDPR's core principles, such as fraud detection, threat prevention, or scientific research conducted in the public interest.²⁶ Others, such as the use of AI for chatbots (without mentioning other types of AI services and applications), are more contentious and highlight the subjective nature of these determinations. The level of granularity also fluctuates, with some guidance targeting specific AI services like chatbots, while others use broader categories such as 'improving the performance of a service or product' or 'ensuring public access to certain information'.²⁷ This lack of systematization risks creating confusion and inconsistency in the application of legitimate interest, and a more structured and granular approach would be critical to reduce ambiguities. Additionally, some guidance refers to prohibitive provisions under legal frameworks such as the Digital Services Act (DSA) when it comes to minor profiling and the AI Act for the prohibition of

certain AI systems.²⁸ While these prohibitions are not directly tied to the formulation of legitimate interest, they offer context for understanding the contours within which AI developers must operate.

Necessity and minimization

Significant disputes remain regarding the operation and relevance of the necessity test in the context of data-intensive, web scraping-based approaches to AI development.²⁹ The majority of authorities, including the EDPB,³⁰ tend to reiterate the conceptual framework of the necessity test, while their guidance often lacks actionable depth. The necessity assessment generally involves two elements: (i) determining whether the processing activity effectively allows for the pursuit of the stated purpose, and (ii) assessing whether there is *no less intrusive* means of achieving this purpose. Given the broad scope of legitimate interest, it is unlikely that the first criterion cannot be met in most cases, and there has been little disputes. The second element concerning the less intrusive means test introduces more complexity and often remains moderately explored in regulatory guidelines. Notably, the UK's ICO once adopted a controversial position, downplaying the need to justify necessity by arguing that alternative approaches to AI development were far from practical realities.³¹ A radical shift in its stance has been observed in recent guidance, however, reflecting feedback from media industries and broader concerns about data protection risks.³² The ICO now emphasizes that AI developers must rigorously justify their chosen approaches to development, particularly when web scraping is involved, a method that involves the large-scale collection of publicly available data without direct engagement with data subjects and hence in the ICO's view constitutes a high-risk processing activity.³³ By challenging the narrative that web scraping is the one and only way to train generative AI models, the ICO's revised stance could fundamentally reshape the development pathways for AI systems.³⁴

23 PDPC *ibid.*

24 OAIC, 'Guidance on Privacy and Developing and Training Generative AI Models' (2024) <<https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-developing-and-training-generative-ai-models>> accessed 24 January 2025.

25 CNIL, 'Relying on the Legal Basis of Legitimate Interests to Develop an AI System' (2024) <<https://www.cnil.fr/en/relying-legal-basis-legitimate-interests-develop-ai-system>> accessed 24 January 2025; Landesbeauftragte für den Datenschutz und die Informationsfreiheit in Baden-Württemberg (n 20).

26 CNIL *ibid.*; Landesbeauftragte für den Datenschutz und die Informationsfreiheit in Baden-Württemberg (n 20).

27 *ibid.*

28 CNIL (n 25).

29 Ugo Pagallo and Jacopo Ciani Sciolla, 'Anatomy of Web Data Scraping: Ethics, Standards, and the Troubles of the Law' (2023) 2023 European Journal of Privacy Law & Technologies 1.

30 European Data Protection Board (n 8).

31 Information Commissioner's Office (n 20).

32 Information Commissioner's Office, 'Information Commissioner's Office Response to the Consultation Series on Generative AI' (11 February 2025) <<https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/response-to-the-consultation-series-on-generative-ai/>> accessed 22 February 2025.

33 *ibid.*

34 *ibid.*

Balancing: risk, reasonable expectations, and mitigation

The balancing test under the legitimate interest analysis requires weighing the legitimate interests pursued by the data controller against the fundamental rights and freedoms of the data subjects. This analysis involves a careful assessment of the proportionality of the data processing, ensuring that the controller's interests do not unduly interfere with individuals' fundamental rights and freedoms. Key factors include the nature and scope of the data processing, the reasonable expectations of individuals, the potential risks or harms associated with the processing, and the safeguards or mitigations implemented to reduce these risks.³⁵ This multi-faceted balancing test is particularly critical in the context of AI training, where the complexities of data-intensive practices and opaque systems amplify the stakes for data protection.

Two key aspects significantly shape reasonable expectations. First, there is the distinction between first-party and third-party processing. Authorities such as Germany's LfDI highlight that individuals are far less likely to expect their data to be processed and shared by third parties, particularly when there has been no direct interaction or consent.³⁶ This distinction stresses the need for transparency and justification when third-party data sharing is involved. Secondly, the issue of secondary data use presents another challenge. Liechtenstein's Datenschutzstelle emphasizes the relevance of purpose limitation, ensuring that data collected for one purpose is not repurposed without adequate safeguards.³⁷ Similarly, Australia's authority highlights the role of disclosure in determining the legality of secondary uses, stressing that transparency about how and why data are being reused is critical for maintaining individuals' trust and expectations.³⁸ Interestingly, LfDI further refines this benchmark with concepts such as public figure theory, which posits that public figures may reasonably expect greater disclosure and use of their personal data compared to private individuals.³⁹

Secondly, the test of reasonable expectations hinges on the adequacy of disclosure, combined with general public perceptions of data processing practices for AI training purposes. In practice, as will be shown in the next section, the adequacy of disclosure remains a cornerstone of reasonable expectations. Without sufficient transparency regarding the purpose, scope, and risks of data processing, individuals' expectations are likely to diverge significantly from the reality of how their data are used.

The risks associated with generative models emerge as a critical dimension of this balancing exercise. However, not all authorities engage with risks in a consistent manner. Among the most proactive are the UK's ICO and Australia's OAIC, which classify generative AI processing as *high* risk due to factors such as the indivisible nature of processing, the involvement of complex AI models, and the use of vast quantities of personal data.⁴⁰ LfDI offers a complementary perspective, highlighting risks related to model malfunctions caused by adversarial attacks or technical errors, and suggesting that these risks be evaluated contextually in data protection impact assessments (DPIAs).⁴¹ Meanwhile, Asian authorities advocate for an *ad hoc*, risk-based approach that incorporates the effectiveness of implemented controls into the assessment.⁴²

In terms of remedies and safeguards, the ability for data subjects to opt out of training is widely considered a necessary means of balancing interests, with many authorities asserting that the ability to opt out should not be overridden by the interests of data controllers. There is, however, considerable variation in how opt-out is framed. For instance, Bavarian authorities explore machine unlearning as a technical approach to implementing the right to erasure in AI training contexts.⁴³ The EDPB implies that opt-out is essential,⁴⁴ whereas Spain's AEPD frames it as a best practice.⁴⁵ Australia's authority, in cases where secondary use aligns with reasonable expectations, recommends offering individuals the option to 'seek consent and/or provide a meaningful and informed ability to opt out'.⁴⁶ While a deeper inquiry is outside the scope of this article, this variation

35 European Data Protection Board, 'Guidelines 1/2024 on Processing of Personal Data Based on Article 6(1)(f) GDPR' (8 October 2024) <https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2024/guidelines-12024-processing-personal-data-based_en> accessed 24 January 2025.

36 Landesbeauftragte für den Datenschutz und die Informationsfreiheit in Baden-Württemberg (n 20).

37 Datenschutzstelle fürstentum Liechtenstein, 'Chatbots' <<https://www.datenschutzstelle.li/datenschutz/themen-z/chatbots>> accessed 24 January 2025.

38 OAIC (n 24).

39 Landesbeauftragte für den Datenschutz und die Informationsfreiheit in Baden-Württemberg (n 20).

40 Information Commissioner's Office (n 20); OAIC (n 24).

41 Landesbeauftragte für den Datenschutz und die Informationsfreiheit in Baden-Württemberg (n 20).

42 ASEAN (n 22).

43 BayLDA, 'KI & Datenschutz' <<https://www.lda.bayern.de/de/ki.html>> accessed 24 January 2025.

44 European Data Protection Board (n 8).

45 AEPD, 'GDPR Compliance of Processings That Embed Artificial Intelligence—An Introduction' (2020) <<https://www.aepd.es/guides/gdpr-compliance-processings-that-embed-ia.pdf>> accessed 24 January 2025.

46 OAIC (n 24).

highlights the legal and operational uncertainty surrounding the role of opt-out, its relationship to consent as a safeguard, as well as the use of emerging technical safeguards such as machine unlearning.⁴⁷

Finally, the role of privacy-enhancing technologies (PETs) remains explicitly mentioned but underdiscussed. Techniques such as differential privacy and federated learning are referenced but rarely analysed in operational terms or with clear evaluations of their contributions. As the ICO notes, current consultations lack sufficient verifiable evidence to assess the practical efficacy of these safeguards beyond theoretical or proof-of-concept stages.⁴⁸ Similarly, the Singaporean authority suggests that deployment of such techniques should be regarded as ‘basic controls’, reflecting a nascent understanding of their role in mitigating risks within complex AI systems.⁴⁹

Overall, our comparative analysis indicates that existing guidelines frame the legal basis for AI training in an often ambiguous, inconsistent, and at times implicitly contradictory manner. The critical dimensions of decision-making and trade-offs are frequently left unaddressed, resulting in a landscape where foundational queries—such as how to operationalize permissible data use—receive abstract rather than actionable guidance. The prevailing focus on legal basis tends to anchor debates in questions of theoretical permissibility, but in our view, this distracts from the development of concrete operational safeguards and mitigation strategies. Regulators routinely reference the tripartite test—purpose legitimacy, necessity, and interest balancing—but rarely provide direction on how these principles should inform design decisions, data engineering pipelines, or mitigation measures that are specific to contemporary AI workflows. Legitimate interest is presented as a flexible, context-sensitive standard, but in practice, this flexibility often translates into doctrinal vagueness. Guidance seldom specifies, for instance, what constitutes necessity in the mass collection or scraping of training data, or how to meaningfully determine the reasonable expectations of data subjects across heterogeneous corpora. Many practitioners engage with training data indirectly, operating through intermediaries or incorporating model outputs from upstream sources, greatly complicating the process of assessing

provenance and compliance with the expectations or interests of original data subjects.⁵⁰ As a result, guidance functions as a doctrinal placeholder, leaving developers to conduct speculative assessments and navigate substantial uncertainty until regulatory intervention is prompted by controversy or complaint.

Regulatory interventions

The public enforcement of data protection law related to AI model training has emerged as a critical theme, with regulatory authorities worldwide actively deploying data protection regulatory toolkits to address its unique challenges. Particularly throughout late 2024, pivotal interventions are set to unfold across jurisdictions, targeting both AI developers and media networks that furnish training content and data. This section sets out to review how regulators have worked to carve out legal clarity by initiating formal investigations and engaging informally with major providers. Given the highly dynamic and rapidly shifting regulatory landscape, our account zeroes in on cases that have reached sufficient maturity—those in which interim or final decisions have dropped, or where substantive indications of violation have surfaced—and excludes investigations initiated under regulatory frameworks other than the GDPR.

All Roads Lead to Rome: Garante v OpenAI and GEDI group

Italy’s Garante has emerged as one of the most assertive European authorities in the regulation of generative AI, handling complex cases involving both AI developers (such as OpenAI and DeepSeek) and data suppliers (including the GEDI groups). Through early and far-reaching enforcement actions, Garante has distinguished itself from other Member States—such as France, Germany, and the Netherlands—that have tended to prioritize guidance and soft-law instruments. By coupling independence of action with coordinated engagement, Garante has established itself as a pivotal regulator shaping the evolving European approach to generative AI.

The sequence of events began in Italy on 1 April 2023, when Garante made international headlines by suspending ChatGPT.⁵¹ This decisive move forced OpenAI to address critical privacy concerns, including

47 Carey Lening, ‘Can LLMs Unlearn? Part 1: Predictable Law, Unpredictable Machines’ (*Privat Insights*, 9 June 2024) <<https://insights.priva.cat/p/predictable-law-unpredictable-machines>> accessed 30 August 2025.

48 Information Commissioner’s Office (n 20).

49 PDPC (n 22).

50 Karl Werder, Balasubramaniam Ramesh and Rongen (Sophia) Zhang, ‘Establishing Data Provenance for Responsible Artificial Intelligence

Systems’ (2022) 13 *ACM Transactions on Management Information Systems* 22. See also Shayne Longpre and others, ‘A Large-Scale Audit of Dataset Licensing and Attribution in AI’ (2024) 6 *Nature Machine Intelligence* 975.

51 Imran Rahman-Jones, ‘ChatGPT: Italy Says OpenAI’s Chatbot Breaches Data Protection Rules’ (*BBC*, 29 January 2024) <<https://www.bbc.com/news/technology-68128396>> accessed 30 August 2025.

revising its privacy policies, implementing age verification, and clarifying the legal bases for data processing.⁵² Garante additionally required OpenAI to explain how it managed challenges unique to generative AI, such as the handling of ‘hallucinations’ in outputs.⁵³ The suspension, which briefly disrupted OpenAI’s operations in Italy, was lifted by the end of April 2023 after OpenAI implemented the specified remedial measures.⁵⁴ The regulatory scrutiny continued nonetheless, with Garante maintaining oversight of OpenAI’s compliance efforts. In January 2024, Garante announced developments in its investigation into OpenAI,⁵⁵ and this update adjusted the scope of its initial concerns, adding failures related to data breach notifications while dropping its earlier focus on ‘hallucinations’.⁵⁶ Shortly thereafter, Garante issued a separate decision against the GEDI Group, a media organization that supplied OpenAI with content for AI training purposes.⁵⁷ The decision scrutinized GEDI’s reliance on legitimate interest as a legal basis for sharing training datasets, and Garante was particularly critical of the group’s DPIA, identifying inadequacies in its handling of sensitive data.⁵⁸ At the close of 2024, Garante released its final decision against OpenAI,⁵⁹ which reflected the culmination of a comprehensive investigation. The €15 million fine imposed on OpenAI might appear modest, but Garante’s emphasis on remedies—notably including mandated public awareness campaigns—highlights its dual objective of penalizing non-compliance and educating stakeholders about responsible AI practices.⁶⁰ These enforcement decisions, coupled with its scrutiny against DeepSeek, which will be addressed later, illustrate the authority’s methodical and multifaceted approach to regulating generative AI. Garante has

proactively influenced the practices of major AI developers and also raised critical questions about the viability of legitimate interest as a legal basis. Some commentators interpret its stance as a categorical rejection of legitimate interest,⁶¹ yet Garante’s nuanced approach suggests a reserved space for weighing complexities, particularly as broader questions of compliance remain under examination.

The Samba of Regulation: ANPD’s pioneering path in Latin America

While Europe often dominates the discussion, Brazil’s National Data Protection Authority (ANPD) is rapidly establishing itself as a global leader in enforcing compliance with its GDPR-inspired data protection law, the *Lei Geral de Proteção de Dados* (LGPD). The ANPD’s assertive actions against Big Tech companies like Meta and X have not only set important precedents locally but are also influencing the global discourse. Through a series of proactive regulatory measures focusing on issues such as transparency, children’s data, and global applicability of compliance strategies, the ANPD has sent a clear signal to AI developers that Brazil demands robust data protection standards.

Brazil’s LGPD shares numerous similarities with the EU’s GDPR, particularly in its recognition of legitimate interest as a legal basis.⁶² However, the ANPD’s approach distinguishes it as a proactive regulatory force. A defining moment came in the summer of 2024 when the ANPD suspended Meta’s privacy policy and AI training practices, citing significant deficiencies and threatening a daily fine of BRL 50,000 (approximately €8389) for non-compliance.⁶³ This marked a sharp

52 OpenAI, ‘Our Approach to Data and AI’ (*OpenAI*, 7 May 2024) <<https://openai.com/index/approach-to-data-and-ai/>> accessed 30 August 2025.

53 *ibid.*

54 Supantha Mukherjee, Giselda Vagnoni and Supantha Mukherjee, ‘Italy Restores ChatGPT after OpenAI Responds to Regulator’ (*Reuters*, 28 April 2023) <<https://www.reuters.com/technology/chatgpt-is-available-again-users-italy-spokesperson-says-2023-04-28/>> accessed 30 August 2025.

55 Garante, ‘ChatGPT: Italian DPA notifies breaches of privacy law to OpenAI’ (*GPDP*, 29 January 2024) <<https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9978020>> accessed 30 August 2025.

56 *ibid.*

57 Garante, ‘Provvedimento del 27 novembre 2024 [10077129]’ (*GPDP*, 27 November 2024) <<https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/10077129>> accessed 30 August 2025.

58 *ibid.*

59 Garante, ‘ChatGPT, The Italian Data Protection Authority closes the preliminary investigation’ (*GPDP*, 20 December 2024) <<https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/10085432>> accessed 30 August 2025.

60 *ibid.*

61 For instance, see Theodore Christakis, ‘A Tale of Two DPAs: Diverging Paths for Generative AI Training under the GDPR’ (*LinkedIn*, 17 December 2024) <https://www.linkedin.com/posts/theodore-christakis-9ab5a6260_a-tale-of-two-dpas-diverging-paths-for-activity-7274735255968780288-qa8v/> accessed 19 February 2025.

62 Ondřej Filipec and others, ‘Personal Data Protection in Brazil: How Much Europeanization?’ (2023) 22 *International and Comparative Law Review* 81; Arturo J Carrillo and Matías Jackson, ‘Follow the Leader? A Comparative Law Study of the EU’s General Data Protection Regulation’s Impact in Latin America’ (2022) 16 *ICL Journal* 177.

63 Imprensa Nacional, ‘Despacho Decisório nº 33/2024/PR/ANPD—DOU—Imprensa Nacional’ (*gov.br*, 2024) <<https://www.in.gov.br/en/web/dou/-/despacho-decisorio-n-33/2024/pr/anpd-581192714>> accessed 30 August 2025; Maria Badillo, ‘Processing of Personal Data for AI Training in Brazil: Takeaways from ANPD’s Preliminary Decisions in the Meta Case’ (*Future Privacy Forum*, 28 October 2024) <<https://fpf.org/blog/processing-of-personal-data-for-ai-training-in-brazil-takeaways-from-anpds-preliminary-decisions-in-the-meta-case/>> accessed 30 August 2025.

contrast to the more tempered approaches seen in Europe, where companies like LinkedIn and Meta often ‘voluntarily’ paused their AI training following regulatory engagement.⁶⁴ By the end of August 2024, the suspension was lifted, following Meta’s compliance with the ANPD’s requirements.⁶⁵ These included addressing the inadequacy of its legal basis for data processing, revising its privacy policy, and rectifying practices that employed ‘dark patterns’ to obscure the right to object.⁶⁶ The ANPD also preliminarily engaged Meta’s handling of children’s data, a dimension of AI regulation that few authorities have scrutinized. Specifically, Meta was required to enhance transparency, reduce the number of steps needed to object to data processing (from eight clicks to three), and simplify forms for users and non-users alike.⁶⁷ It also pledged to exclude data from minors under 18 years from its training datasets, acknowledging that such data extended beyond Facebook and Instagram.⁶⁸

Following the case against Meta, X faced similar scrutiny from the ANPD. In a decision issued in mid-December 2024, X was ordered to cease processing minors’ data for AI training and to disable any features enabling such data sharing.⁶⁹ Additionally, X was instructed to update its Privacy Policy and Terms of Use and to ensure that its DPIAs and Legitimate Interest Assessments (LIAs), originally designed for the GDPR, were explicitly adapted to meet the requirements of Brazilian law.⁷⁰ In other words, Brazilian regulators demanded not just a transplant of European compliance frameworks but a contextual recalibration, one that

recognized the distinct normative, procedural, and enforcement expectations embedded in Brazil’s LGPD.

A symphony of soft governance: ICO and DPC in tune with Meta’s AI Training

The regulatory response to Meta’s AI training controversies across the UK, Ireland, and the broader European contexts illustrates a sophisticated strategy favouring negotiation and constructive dialogue over aggressive enforcement. Meta’s ‘voluntary’ suspension of AI training, following its June 2024 privacy policy update, exemplifies this negotiated posture. The UK’s Information Commissioner’s Office (ICO) commenced engagement with Meta in response to the update, leading to an extended period of dialogue during which Meta paused its AI training activities.⁷¹ It culminated in Meta announcing a collaboration with the ICO to develop a compliance framework anchored in legitimate interest.⁷² Notably, when Meta unveiled its revised compliance plan,⁷³ the ICO issued a concurrent statement emphasizing the centrality of transparency and accountability.⁷⁴ The ICO refrained from granting formal regulatory approval or intervention, instead signalling through a deliberately measured posture that Meta’s position remains non-committal and subject to further review.⁷⁵ This approach signals the ICO’s broader strategy, which favours constructive engagement and incremental accountability over enforcement-led intervention. An analogous posture is seen in the ICO’s investigation of Snap’s ‘My AI’ chatbot, where regulatory scrutiny focused on the adequacy of the DPIA.⁷⁶ The ICO’s preliminary enforcement notice required Snap to revise its

64 Phil Muncaster, ‘LinkedIn Pauses GenAI Training Following ICO Concerns’ (*Infosecurity Magazine*, 23 September 2024) <<https://www.infosecurity-magazine.com/news/linkedin-pauses-genai-training-ico/>> accessed 21 January 2025; Paul Sawers, ‘Meta Pauses Plans to Train AI Using European Users’ Data, Bowing to Regulatory Pressure’ (*TechCrunch*, 14 June 2024) <<https://techcrunch.com/2024/06/14/meta-pauses-plans-to-train-ai-using-european-users-data-bowing-to-regulatory-pressure/>> accessed 30 August 2025.

65 ANPD, ‘Meta cumpre exigências da ANPD e poderá retomar, com restrições, o uso de dados pessoais para treinamento de inteligência artificial’ (ANPD, 30 August 2024) <<https://www.gov.br/anpd/pt-br/assuntos/noticias/meta-cumpre-exigencias-da-anpd-e-podera-retomar-com-restricoes-o-uso-de-dados-pessoais-para-treinamento-de-inteligencia-artificial>> accessed 31 January 2025.

66 *ibid.*

67 *ibid.*

68 *ibid.*

69 ANPD, ‘Despacho Decisório No 29/2024/FIS/CGF’ (ANPD 2024) <https://www.gov.br/anpd/pt-br/assuntos/noticias/SEI_0161130_Despacho_Decisorio_29.pdf> accessed 30 August 2025.

70 *ibid.*

71 Information Commissioner’s Office, ‘Statement in Response to Meta’s Plans to Train Generative AI with User Data’ (14 June 2024) <<https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/06/statement-in-response-to-metas-plans-to-train-generative-ai-with-user-data/>> accessed 25 January 2025.

72 *ibid.*

73 ‘Building AI Technology for the UK in a Responsible and Transparent Way’ (*Meta*, 13 September 2024) <<https://about.fb.com/news/2024/09/building-ai-technology-for-the-uk-in-a-responsible-and-transparent-way/>> accessed 22 January 2025.

74 Information Commissioner’s Office, ‘ICO Statement in Response to Meta’s Announcement on User Data to Train AI’ (13 September 2024) <<https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/09/ico-statement-in-response-to-metas-announcement-on-user-data-to-train-ai/>> accessed 30 January 2025.

75 *ibid.*

76 Information Commissioner’s Office, ‘We Warn Organisations Must Not Ignore Data Protection Risks as We Conclude Snap “My AI” Chatbot Investigation’ (21 May 2024) <<https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/05/ico-warns-organisations-must-not-ignore-data-protection-risks-as-it-concludes-snap-my-ai-chatbot-investigation/>> accessed 30 January 2025.

DPIA and, once completed, the ICO concluded that Snap fulfilled UK data protection requirements for high-risk processing.

The Irish DPC presents a similarly nuanced yet inconsistent approach.⁷⁷ In contrast to several European authorities, the DPC has not yet published formal guidelines specific to generative AI, except for a single blog post outlining preliminary reflections and high-level considerations. Since September 2024, the DPC has played a pivotal role in seeking an opinion from the EDPB,⁷⁸ thereby catalysing Europe-wide dialogue. During the regulatory controversy surrounding Meta, the DPC issued public statements and instructed Meta to take a range of measures, including enhanced transparency notices, improved objection mechanisms, extended notice period, user control enhancements, and technical safeguards (eg, data de-identification and filtering) and eventually an ongoing report evaluating the effectiveness of these measures.⁷⁹

In contrast, the engagement with X by the DPC has tested the limits of its dialogic approach. It initially sought to address concerns about X's (Twitter International's) use of public user data for training through negotiation and requests for voluntary compliance.⁸⁰ However, when X declined to pause processing or defer further deployment, the DPC resolved to escalate the matter through formal High Court proceedings.⁸¹ This shift indicates that, while the DPC prioritizes negotiation and incremental remediation, it is set to pursue urgent judicial remedies, including orders to suspend or prohibit processing when dialogue proves unavailing. The referral of the dispute to the EDPB further illustrates the DPC's commitment to seeking clarity through cooperative mechanisms at the European level before, or alongside, hard enforcement.⁸² The DPC has similarly engaged with Google

regarding its large language model, PaLM2, raising concerns over the adequacy of Google's DPIA and initiating an inquiry that, at the time of writing, has yet to reach a resolution.⁸³

The global onslaught of DeepSeek

The regulatory trajectory of DeepSeek, the Chinese AI developer whose models have quickly achieved global reach, illustrates both the intensification and the globalization of regulatory responses to generative AI. In contrast to earlier cases dominated by Western actors, DeepSeek's emergence as an Eastern challenger has drawn not only sustained scrutiny but also a more assertive and, in many respects, path-dependent wave of enforcement across multiple jurisdictions. This article treats the DeepSeek episode as a distinct but related regulatory phenomenon: its arrival coincides with a moment when European debates about GDPR compliance have advanced along the path to maturity, affording authorities more institutional experience and coordinated leverage than was available during prior rounds of AI governance.

In a sequence reminiscent of OpenAI's earlier encounter with European authorities, Italy's Garante was the first to intervene, imposing an outright ban on DeepSeek after finding that its privacy practices violated the GDPR—particularly with respect to deficiencies in transparency and the transfer of personal data from the EU to China.⁸⁴ This initial measure catalysed a series of coordinated investigations by authorities in France, the Netherlands, Belgium, Ireland, and Luxembourg, with the EDPB actively monitoring DeepSeek's compliance and considering the prospect of collective action.⁸⁵ While most regulatory actions have thus far remained at the time of writing, the Berlin Data Protection Commissioner took a notably assertive stance by declaring DeepSeek's operations unlawful, requesting major

77 Data Protection Commission, 'The DPC's Engagement with Meta on AI' (*Data Protection Commission*, 14 June 2024) <<https://www.dataprotection.ie/news-media/latest-news/dpcs-engagement-meta-ai>> accessed 23 February 2025.

78 Data Protection Commission, 'Irish Data Protection Commission Welcomes EDPB Opinion on the Use of Personal Data for the Development and Deployment of AI Models' (*Data Protection Commission*, 18 December 2024) <<https://www.dataprotection.ie/irish-data-protection-commission-welcomes-edpb-opinion-use-personal-data-development-and-deployment>> accessed 30 August 2025.

79 Data Protection Commission, 'DPC Statement on Meta AI' (*Data Protection Commission*, 21 May 2025) <<https://www.dataprotection.ie/news-media/latest-news/dpc-statement-meta-ai>> accessed 30 August 2025.

80 Data Protection Commission, 'Data Protection Commission Welcomes Conclusion of Proceedings Relating to X's AI Tool "Grok"' (*Data Protection Commission*, 4 September 2024) <<https://www.dataprotection.ie/news-media/press-releases/data-protection-commission-welcomes-conclusion-proceedings-relating-xs-ai-tool-grok>> accessed 30 August 2025.

81 'DPC in Action against Twitter over AI Data Concerns' (*RTE*, 6 August 2024) <<https://www.rte.ie/news/business/2024/0806/1463712-dpc-takes->

[court-action-against-twitter-over-data-concerns/](https://www.rte.ie/news/business/2024/0806/1463712-dpc-takes-court-action-against-twitter-over-data-concerns/)> accessed 30 August 2025.

82 Data Protection Commission, 'REQUEST FOR AN OPINION OF THE EUROPEAN DATA PROTECTION BOARD PURSUANT TO ART. 64 (2) GDPR' (*EDPB*, December 2024) <<https://www.edpb.europa.eu/system/files/2024-12/ie-sa-article-64-2-request-re-ai-model-processing-redacted-en.pdf>> accessed 30 August 2025.

83 Data Protection Commission, 'Data Protection Commission Launches Inquiry into Google AI Model' (*Data Protection Commission*, 12 September 2024) <<https://www.dataprotection.ie/news-media/press-releases/data-protection-commission-launches-inquiry-google-ai-model>> accessed 30 August 2025.

84 Garante, 'Artificial Intelligence: The Italian Data Protection Authority blocks DeepSeek' (*GPDP*, 30 January 2025) <<https://www.garanteprivacy.it/443/home/docweb/-/docweb-display/docweb/10097450>> accessed 30 August 2025.

85 Claudie Moreau, 'DeepSeek Making a Splash with EU Data Protection Bodies' (*Euractiv*, 3 February 2025) <<https://www.euractiv.com/section/tech/news/deepseek-making-a-splash-with-eu-data-protection-bodies/>> accessed 30 August 2025.

app store providers—namely Apple and Google—to delist DeepSeek’s application.⁸⁶

Outside Europe, regulatory responses have been equally assertive and uncoordinated. South Korea’s Personal Information Protection Commission (PIPC) initiated an investigation, directed the removal of DeepSeek from domestic app stores, and compelled the company to revise its privacy policy and strengthen cross-border data transfer safeguards.⁸⁷ In the USA, DeepSeek is subject to bans on government devices imposed by several states and agencies, with discussions ongoing about extending these restrictions at the federal level on grounds of privacy and national security.⁸⁸ Comparable measures have been enacted in Canada⁸⁹ and Australia,⁹⁰ where DeepSeek is prohibited on government systems, while India has introduced bans on both DeepSeek and ChatGPT for specific governmental IT infrastructure.⁹¹ Unlike earlier cases, which were shaped by early-stage doctrinal uncertainty or single-jurisdictional leadership, the DeepSeek case evidences a convergence of cross-border action, a heightened sensitivity to data flows towards and storage in China.

Beyond the epicenters: fragmented horizons in the regulation of generative AI

Mapping the global landscape of regulatory responses to generative AI presents both methodological and practical challenges. While this article focuses on the principal clusters of national interventions, it is instructive to briefly survey emerging developments beyond these focal points.

Within the EU, much regulatory activity continues to be initiated and prosecuted at the member state level;

absent DeepSeek’s formal European establishment and hence formal cross-border investigations, such matters rarely escalate to pan-European dispute resolution like the GDPR’s one-stop shop.⁹² In this context, agencies such as the EDPB and the European Data Protection Supervisor (EDPS) exert influence mainly through guidance. A further complexity arises from the rapid evolution of the EU’s regulatory landscape: newer instruments such as the DSA, the Digital Markets Act (DMA), and the AIA are beginning to redirect some of the focus from the GDPR, introducing novel pathways for addressing generative AI through competition, transparency, and systemic risk frameworks.⁹³

In contrast to Europe—where regulatory scrutiny is most sustained—other continents display a patchwork of sporadic engagement (other than DeepSeek’s regulatory reckoning), with interventions varying substantially in scope, substance, and sophistication. Across Asia, regulatory approaches display significant heterogeneity. In mainland China, for instance, there is currently no documented enforcement under the Personal Information Protection Law or specific measures on generative AI, and China’s path towards AI-specific regulation remains fraught with uncertainties and controversies.⁹⁴ As such, enforcement priorities remain tilted towards content regulation and outputs—such as copyright eligibility of AI-generated works—rather than the legality of input data or training corpora. Singapore, for its part, continues to refine a guidance-oriented approach—acknowledging legitimate interest as a potential legal basis and delineating best practices to support compliance by businesses.⁹⁵ Japan, while explicit in its commitment to developing robust AI regulation, has adopted an approach that is primarily framed through

86 ‘Berliner Datenschutzbeauftragte meldet KI-App DeepSeek in Deutschland bei Apple und Google als rechtswidrigen Inhalt’ (*Berliner Beauftragte für Datenschutz und Informationsfreiheit*, 27 June 2025) <<https://www.datenschutz-berlin.de/pressemitteilung/berliner-daten-schutzbeauftragte-meldet-ki-app-deepseek-in-deutschland-bei-apple-und-google-als-rechtswidrigen-inhalt>> accessed 30 August 2025.

87 Personal Information Protection Commission, ‘DeepSeek Temporarily Suspends Its Application Service in Korea’ (18 February 2025) <https://www.pipc.go.kr/eng/user/ltm/new/noticeDetail.do?bbsId=BBSMSTR_00000000001&ntId=2784> accessed 30 August 2025.

88 Vishnu Priya, ‘US Mulling DeepSeek App Ban on Government Devices’ (*Verdict*, 10 March 2025) <<https://www.verdict.co.uk/us-deepseek-app-ban/>> accessed 30 August 2025.

89 Tara Deschamps, ‘Chinese AI App DeepSeek Blocked on Some Federal Government Phones’ (*CBC News*, 7 February 2025) <<https://www.cbc.ca/news/politics/deepseek-federal-device-1.7453401>> accessed 30 August 2025.

90 Tom Gerken, ‘Australia Bans DeepSeek on Government Devices over Security Risk’ (*BBC*, 4 February 2025) <<https://www.bbc.com/news/articles/c8d95v0nr1yo>> accessed 30 August 2025.

91 Jonathan Easton, ‘India Bans ChatGPT and DeepSeek in Finance Ministry amid Data Concerns’ (*National Technology News*, 5 February 2025) <https://nationaltechnology.co.uk/India_Bans_ChartGPT_and_Deepseek_In_Finance_Ministry.php> accessed 30 August 2025.

92 Diogo Matos Brandão, ‘The One-Stop-Shop and the European Data Protection Board’s Role in Combatting Data Supervision Forum Shopping’ (2023) 13 *International Data Privacy Law* 313.

93 See for instance the DSA enforcement, including European Commission, ‘Commission Requests Information from Online Marketplaces Temu and Shein on Compliance with the Digital Services Act’ (28 June 2024) <<https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-online-marketplaces-temu-and-shein-compliance-digital-services-act>> accessed 31 January 2025; European Commission, ‘Commission Sends Requests for Information on Generative AI Risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act’ (14 March 2024) <<https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very>> accessed 31 January 2025; European Commission, ‘Commission Compels Microsoft to Provide Information under the Digital Services Act on Generative AI Risks on Bing’ (17 May 2024) <<https://digital-strategy.ec.europa.eu/en/news/commission-compels-microsoft-provide-information-under-digital-services-act-generative-ai-risks>> accessed 31 January 2025.

94 Mimi Zou and Lu Zhang, ‘Navigating China’s Regulatory Approach to Generative Artificial Intelligence and Large Language Models’ (2025) 1 *Cambridge Forum on AI: Law and Governance* e8.

95 PDPC (n 22).

the lens of industrial policy and economic competitiveness, with ongoing legislative initiatives but, to date, an absence of targeted regulatory enforcement actions against generative AI.⁹⁶ In contrast, South Korea stands out as an exception within the region for its vigorous regulatory posture towards AI. Its Personal Information Protection Commission (PIPC) has not only issued detailed and up-to-date guidance but has also launched an investigation and imposed remedial measures in response to data protection risks posed by generative AI—including but not limited to the DeepSeek's suspension previously mentioned.⁹⁷ This proactive stance exemplifies the emergence of South Korea as a leader in shaping regional AI governance standards. Additionally, Hong Kong also distinguishes itself within the broader Asian landscape. The Privacy Commissioner for Personal Data have recently intervened in high-profile cases, most notably by suspending AI training activities conducted by LinkedIn.⁹⁸ This intervention is significant given the generally cautious approach prevailing across the region, and demonstrates that, while rare, substantive and direct enforcement actions against major international AI providers can and do occur in certain Asian jurisdictions.

In North America, regulatory responses to generative AI remain characterized by institutional fragmentation and policy volatility. In the USA, federal leadership has largely depended on executive order rather than durable legislative consensus, with a proliferation of bills introduced but only a limited number becoming law.⁹⁹ President Biden's Executive Order 14110 ushered in a comprehensive, risk-based approach to AI governance, but the order was promptly rescinded by President Trump in January 2025 in favour of regulatory rollback and market-oriented competitiveness.¹⁰⁰ In Canada, aside from the DeepSeek matter, the forthcoming Artificial Intelligence and Data Act (AIDA) signals an intent to introduce a risk-based AI regulatory regime, yet as of mid-2025, there have been no significant enforcement actions under AIDA or related

frameworks.¹⁰¹ To our best knowledge, other continents and jurisdictions—including Australia/New Zealand, North America (except the DeepSeek bans), South America (except Brazil), and Africa—have yet to initiate formal and comprehensive interventions into AI model training. As a result, the contours of lawful AI training continue to be shaped by not just blockbuster enforcement actions, but also by the uneven global diffusion of regulatory expectations, norms, and practices.

In sum, our comparative analysis of the regulatory interventions reveals that enforcement actions might, in theory, compensate for the ambiguities set forth in high-level guidelines previously analysed but actually do little to resolve them in reality. Instead, interventions tend to operate with a minimalist, procedural focus—prioritizing surface-level remedies such as opt out (which is often compromised by interface design flaws) and narrowly defined technical safeguards, rather than engaging with substantive and operational complexity of legal basis determination for AI training. Moreover, this minimalist enforcement subsists alongside a broader shift in regulatory concern: authorities increasingly and uncoordinatedly direct their energies to a range of issues such as transparency, children's data (as defined by the ANPD) and cross-border flows (in the case of DeepSeek), further marginalizing sustained deliberation on the legitimacy and operationalization of legal bases for AI training.

The prospect of international convergence

Beyond providing a current and comprehensive map of regulatory activity, this section advances the analysis by foregrounding the critical issue of international alignment: not as a conceptual abstraction, but as a matter of genuine, operational harmonization. While surface-level agreement on principles like 'legitimate interest' occasionally emerges as a common thread, meaningful convergence collapses when examining the granular

96 Dominic Paulger, 'Understanding Japan's AI Promotion Act: An "Innovation-First" Blueprint for AI Regulation' (*Future Privacy Forum*, 5 July 2025) <<https://fpf.org/blog/understanding-japans-ai-promotion-act-an-innovation-first-blueprint-for-ai-regulation/>> accessed 3 September 2025.

97 Kyoungsic Min, 'South Korea's PIPC Flexes Its Muscles: What to Know about AI Model Deletion, Cross-Border Transfers and More' (*IAPP*, 4 June 2025) <<https://iapp.org/news/a/south-korea-s-pipc-flexes-its-muscles-what-to-know-about-ai-model-deletion-cross-border-transfers-and-more>> accessed 3 September 2025.

98 PDPC, 'Privacy Commissioner's Office Welcomes LinkedIn's Pause of Using Hong Kong Users' Personal Data for Training Generative Artificial Intelligence (AI) Models' (*PDPC*, 15 October 2024) <https://www.pcpd.org.hk/english/news_events/media_statements/press_20241015.html> accessed 3 September 2025.

99 JD Alois, 'AI Regulation Said To Be "Out Of Control" - 636 AI Bills In The 50 States, 107 In Congress' (*Crowdfund Insider*, 24 February 2025) <<https://www.crowdfundinsider.com/2025/02/236679-ai-regulation-said-to-be-out-of-control-636-ai-bills-in-the-50-states-107-in-congress/>> accessed 3 September 2025.

100 The White House, 'Removing Barriers to American Leadership in Artificial Intelligence' (23 January 2025) <<https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/>> accessed 3 September 2025.

101 'Artificial Intelligence and Data Act' (*Government of Canada*, 27 September 2023) <<https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act>> accessed 3 September 2025.

standards for necessity, individual rights, and risk assessment. The comparative evidence makes clear that a fully harmonized global regime for lawful AI training is not only improbable but risks being conceptually misguided. Persistent, deep-rooted divergences in the interpretation and operationalization of legal bases—shaped by national enforcement cultures, policy priorities, and institutional capacities—continue to fracture the regulatory landscape. Taking a pragmatic lens, this section does not claim to resolve or prescribe global declarations or institutional blueprints. Instead, it lays a foundation that distinguishes areas where harmonization is already visible, those where incremental convergence seems plausible, and domains where deep legal or cultural entrenchment makes alignment challenging. As such, it serves as a critical basis for considering what global or subglobal AI governance might look like in practice, rather than simply cataloguing conceptual agreement or aspirational frameworks.

Areas demonstrating clear and emerging convergence

Operational opt-out mechanism/right to object and dark patterns

The adoption of a basic right to object or opt out from the use of personal data in AI training has emerged as a procedural safeguard in avoidance to burdensome consent collection and management. Authorities have pushed controllers to roll out clear opt-out mechanisms for data subjects, especially in high-risk contexts and where consent is impractical. While implementation may vary, the existence of an opt-out mechanism appears to be a ‘lowest common denominator’ standard. This push has filtered into industry practice: Meta, for example, recently revised its privacy policy across Europe to streamline opt-out flows under regulatory and judicial scrutiny.¹⁰² Hence, the opt-out right is converging procedurally as a global norm—a result of both regulatory pressure and practical necessity for scalable governance.

The growing recognition of opt-out as a remedy gives rise to two pressing questions—one conceptual, the other practical. Conceptually, should AI training default to ‘off’, or are its innovative and arguably less intrusive uses sufficient to tip the balance towards default

inclusion?¹⁰³ Grappling with this dilemma would require a nuanced conceptualization—an inquiry falling beyond the remit of this article—of the potential harms emanating from AI training, set against the imperative to balance innovation with maintaining a high level of data protection. Practically, the conversation pivots to whether opt-out mechanisms deliver real efficacy or simply give the appearance of control, especially as transparency shortfalls and manipulative interface designs persist.

While many laws notionally prohibit the manipulation of user rights and require fair presentation, the enforcement and targeting of ‘dark patterns’ are inconsistent and mostly reactive. In our sample, only ANPD intervened, making specific requirements regarding dark patterns and the ease of accessibility. Noyb is actively pushing for the redesign of especially egregious opt-out interfaces,¹⁰⁴ and authorities like CNIL and ICO warn designs that obscure or complicate opt-out.¹⁰⁵ But most authorities currently neither proactively audit designs nor penalize subtle coercion. Repeated regulatory statements urge improvements, but the lack of binding standards or regular enforcement leaves the effectiveness of user rights currently variable across jurisdictions.

Quality and management of training data

The drive to systematically identify, filter, and validate the quality and legitimacy of AI training data is frequently echoed in guidelines and enshrined within statutes like Article 10 of the EU AI Act; yet, substantive operational requirements remain embryonic. Interventions by authorities—most notably in Brazil and Italy—reflect a trend of reactive enforcement, typically triggered by complaints or harms rather than proactive, preventive oversight. Remediation tends to be *ad hoc*, with authorities calling for purging of questionable, special category, or sensitive data primarily in response to specific risks or scandals. Investigations may begin with a focus on sensitive data, but operational priorities shift, and requirements may be dropped or diluted as cases progress. As suggested by Kuru,¹⁰⁶ the *Meta v Bundeskartellamt* judgment has profound implications for AI training with large-corpus datasets containing special category data. If this jurisprudence is applied strictly, the threshold for relying on consent becomes

102 Gehrman Mareike Christine, Rammos Thanos and Horn Jakob, ‘Meta vs Verbraucherzentrale NRW’ (*TaylorWessing*, 26 May 2025) <<https://www.taylorwessing.com/zh-hant/insights-and-events/insights/2025/05/meta-vs-verbraucherzentrale-nrw>> accessed 30 August 2025.

103 Nikolas Guggenberger, ‘Consent as Friction’ (2025) 66 *Boston College Law Review* 353.

104 noyb, ‘noyb sends Meta ‘cease and desist’ letter over AI training. European Class Action as potential next step’ (*noyb*, 14 March 2025) <<https://noyb.eu/en/noyb-sends-meta-cease-and-desist-letter-over-ai-training-european-class-action-potential-next-step>> accessed 30 August 2025.

105 CNIL (n 25).

106 Kuru (n 6).

nearly inescapable, since special category data are inseparable from vast, unstructured datasets and cannot be lawfully processed—even for training—without explicit, granular consent. In response to these challenges, PETs have been repeatedly suggested by authorities, positioned as technical means to reconcile model performance with data protection. In the absence of concrete technical standards or measurable benchmarks, however, current recommendations remain at a conceptual or aspirational level, leaving the real-world operational effect unclarified and adoption uneven.

The incentives for AI developers to undertake rigorous due diligence can be misaligned where extensive filtering threatens model efficiency or performance. Moreover, the legality of web-based scraping, particularly vis-à-vis personal and special category data, remains contested and subject to substantial ambiguity, which also stands as a Collingridge dilemma where effective intervention has already lagged behind widespread adoption and entrenchment of practicess.¹⁰⁷ Notwithstanding these limitations, the field is now actively evolving; it could be expected that the EU¹⁰⁸ and international standards¹⁰⁹ that foreground accuracy, relevance, consistency, diversity, and even legitimacy of training data are driving harmonization and operational alignment across the global AI ecosystem.

Areas with moderate or incremental prospects of convergence

Reasonable expectations of individuals

Individuals' reasonable expectations regarding the use of their data for AI training are persistently invoked by regulators, mostly because of the reliance on legitimate interest, yet remain a deeply contested and operationally elusive concept. Leading DPAs such as the CNIL have explicitly flagged that mass web scraping for AI model training typically fails the 'reasonable expectation' test under GDPR,¹¹⁰ yet not all authorities provide clarity on what can reasonably be expected in the digital environment. Consensus appears to be emerging from discussions that differentiate between the ways data are collected—first-party versus third-party—and from

those who categorically reject practices like mass web scraping as fundamentally failing to meet legal standards, with some authorities even invoking traditional public figure doctrine to bolster their position.¹¹¹ Despite the divergences, the concept of reasonable expectation is increasingly recognized as a central—if underdeveloped—foundation for future global regulatory frameworks.¹¹² The EDPB, CNIL, ICO, and Belgian DPA all recommend balancing user expectations with procedural transparency, opt-out options, and technical safeguards (such as filtering sensitive categories and minimizing retention), but broad operational consensus is yet to materialize.

The relevance and emphasis of consent

Consent, once lionized as the gold standard of data protection,¹¹³ now inhabits only a reserved and ambiguous niche within regulatory architectures that increasingly favour more pragmatic and scalable foundations. Across leading guidelines, regulatory bodies seldom recognize consent as a relevant legal ground for mass data acquisition in AI training; in some cases—most notably the UK ICO¹¹⁴—it is omitted from consideration altogether, reflecting its operational impracticality and the burdens inherent in specific, informed, and revocable authorization at scale. Despite its marginalization, consent is not rendered obsolete; it remains indispensable for high-risk, special category, or unexpected data processing, protecting individuals, at least in theory, where ambiguity and opacity prevail. Moreover, its diminished, yet strategically preserved, role shapes the conditions under which AI developers and data controllers operate. Overall, contemporary regulatory rhetoric neither prioritizes nor operationalizes consent, instead expressing policy preference for legitimate interest—even as it imposes demands for transparent, well-documented accountability mechanisms. Developers themselves, if given the liberty, overwhelmingly avoid consent as a route for AI training, deterred by its unreliability and demanding procedural rigour; in practice, the requirements for meaningful consent are incrementally faded out and only conceptually acknowledged.

107 David Collingridge, *The Social Control of Technology* (London: Frances Pinter 1980).

108 Garrido Josep Soler and others, 'Harmonised Standards for the European AI Act' (European Commission, 2024) <<https://publications.jrc.ec.europa.eu/repository/handle/JRC139430>> accessed 30 August 2025.

109 International Organization for Standardization, 'Artificial Intelligence—Data Quality for Analytics and Machine Learning (ML)', ISO/IEC 5259-5:2025' <<https://www.iso.org/standard/84150.html>> accessed 30 August 2025.

110 CNIL (n 25).

111 Landesbeauftragte für den Datenschutz und die Informationsfreiheit in Baden-Württemberg (n 20).

112 Eric Barendt, "'A Reasonable Expectation of Privacy': A Coherent or Redundant Concept?" in Andrew T Kenyon (ed), *Comparative Defamation and Privacy Law* (Cambridge University Press 2016) <<https://www.cambridge.org/core/books/comparative-defamation-and-privacy-law/reasonable-expectation-of-privacy-a-coherent-or-redundant-concept/3CB4C8FEB107530261334FF8A6411A7D>> accessed 30 August 2025.

113 Eleni Kosta, *Consent in European Data Protection Law* (Martinus Nijhoff 2013).

114 Information Commissioner's Office (n 20).

This evolution echoes past regulatory debates at the turn of the century, i.e., the question of whether opt-out, rather than opt-in, suffices for personalized tracking—and the history might repeat itself in the EU.¹¹⁵ Globally, however, it is evident that the space for consent has been dramatically curtailed. The regulatory frameworks treat consent less as an operational tool and more as a normative baseline not to be operationalized for every use but shaping the boundaries and obligations of legitimate interest in AI training, without asserting hierarchy. In doing so, they provide safeguard against runaway, opaque processing while leaving open the possibility for future harmonization.

Areas facing significant barriers to convergence or evident divergence

Minimization and the necessity of web scraping

The principles of data minimization and necessity (in legitimate interest test) are enshrined as foundational norms,¹¹⁶ but their operationalization reveals deep and persistent fissures across jurisdictions. Most guidelines, including those from the ICO¹¹⁷ and CNIL,¹¹⁸ instruct developers to process only data that is strictly necessary for the stated training purpose, emphasizing the need for feature selection, yet the scale and opacity of AI architectures render adequacy assessments highly indeterminate, especially when datasets are derived from open web sources. Regulators rarely intervene materially on this tension, leaving organizations to navigate minimization based on broad principles rather than concrete benchmarks or thresholds.

The drive towards convergence is also undermined by inconsistent thresholds and enforcement patterns. While data minimization is well-entrenched, DPAs vary widely in the extent to which they penalize over-collection or opaque feature inclusion in the AI context. Enforcement is typically remedial, resulting in advice or recommendations—rather than sanctions or compulsory reductions of datasets—even among regimes with similarly worded statutes.

Hallucination, accuracy and rectification

The last area involving noticeable divergence or uncertainty is ‘hallucination’, understood in model development contexts as the generation of outputs by advanced large language models that are inaccurate, fabricated, or misrepresentative.¹¹⁹ These phenomena go beyond simple factual inaccuracy, reflecting the underlying probabilistic and associative mechanisms that drive modern generative models.¹²⁰ Within data protection law, hallucination sharply intersects with the principle of accuracy, posing nuanced practical and theoretical challenges. Italy’s Garante initially led the way with explicit requirements for technical controls to address inaccuracies in generative model outputs, particularly in high-profile interventions against ChatGPT.¹²¹ However, such measures were later diluted, and enforcement remains *ad hoc*. Regulatory guidance is embryonic but increasingly attentive to hallucination’s implications. Yet, most frameworks, including the EDPB and national DPAs, struggle to properly deal with probabilistic outputs that are not clearly linked to identifiable individuals. In reaction to this, public advocacy bodies such as *noyb* have taken up the mantle in initiating debate regarding hallucination and accuracy.¹²² Common practices default to erasure or data deletion, often sidestepping meaningful rectification—a stance notably critiqued by the EDPB Taskforce vis-à-vis OpenAI’s.¹²³ In parallel, without diving into machine unlearning and the can of worms concerning technical challenges associated with removing specific data, memories, or outputs from models,¹²⁴ it suffices to observe that the approaches in this area remain active but highly divergent, with no global or sub-global consensus having emerged to date.

Conclusion

This paper surveys how legal basis is framed by data protection guidelines and interventions and illuminates the profound disconnect between regulatory aspiration and operational realities in the context of generative AI regulation. It pivots on a granular, evidence-based

115 Eleni Kosta, ‘Peeking into the Cookie Jar: The European Approach towards the Regulation of Cookies’ (2013) 21 *International Journal of Law and Information Technology* 380.

116 European Data Protection Board (n 8).

117 Information Commissioner’s Office (n 20).

118 CNIL (n 25).

119 Yujie Sun and others, ‘AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content’ (2024) 11 *Humanities and Social Sciences Communications* 1278.

120 Chunsik Lee and others, ‘Generative AI Risks and Resilience: How Users Adapt to Hallucination and Privacy Challenges’ (2025) 19 *Telematics and Informatics Reports* 100221.

121 Garante (n 55).

122 *noyb*, ‘AI Hallucinations: ChatGPT Created a Fake Child Murderer’ (*noyb*, 20 March 2025) <<https://noyb.eu/en/ai-hallucinations-chatgpt-created-fake-child-murderer>> accessed 30 August 2025.

123 European Data Protection Board, ‘Report of the Work Undertaken by the ChatGPT Taskforce’ (24 May 2024) <https://www.edpb.europa.eu/our-work-tools/our-documents/other/report-work-undertaken-chatgpt-taskforce_en> accessed 31 January 2025.

124 *ibid*.

account of critical compliance domains, revealing that the legal basis for AI training is consistently cast in ambiguous and uneven terms—riddled with conceptual vagueness and a lack of practical clarity. Further, instead of establishing clear paths to compliance, prevailing interventions default to technical fixes and loosely defined security measures, leaving the substantive protection of data subjects precariously thin. This suboptimal approach sustains regulatory uncertainty and leaves unresolved the fundamental tension between model development as a form of innovation and accountability for the distinctive risks posed by generative AI. What becomes clear is that current enforcement trends lean far too heavily on aspirational ideals, while losing sight of the practical trade-offs and concrete decision-making required to make AI applications both compliant and trustworthy. This paper's central contribution thus lies in elevating the legal basis from a mere foundational block to a key inflexion point—one intimately interwoven with other doctrinal principles and regulatory mechanisms. Rather than pursuing global uniformity for its own sake, authorities and policymakers are advised to privilege practical, evidence-driven, and adaptive solutions that shore up the reliability and fairness of AI regulation, even in the face of substantial jurisdictional diversity.

In this unsettled landscape, inquiring about the prospect of convergence might be viewed as too early or ambitious, given that the dynamics are rapidly evolving. Still, we take on this consideration since the spontaneous and uncoordinated emergence of frameworks across jurisdictions has exposed very rich patterns and valuable contrasts. The classification provided in this article highlights clear potential for both divergence and convergence across key domains. It presents a strategically charted landscape which facilitates not only the identification of challenges and opportunities for interoperability, but also the formulation of realistic goals and cooperation at global and sub-global levels.

While the GDPR does not formally prioritize one lawful basis over another, a functional pecking-order has emerged—one that privileges legitimate interest without fully replacing the safeguards that consent was meant to offer. Still, the invocation of legitimate interest risks becoming a façade, one that obscures rather than enables meaningful accountability. Framing the debate solely at the level of legal basis is hence reductive: true governance must reckon with the deeper compliance architectures and strategic choices at play, an interconnected web lacking agreed direction and warranting systemic reflection. Hence, advancing GDPR-based AI governance demands more than conceptual clarity and coherence—it requires operational specificity, coordinated regulatory ambition, and a willingness to face, rather than evade, real-world complexity. Lawful AI training is thus recast: not just a legal puzzle, but a critical test of our collective capacity for adaptive, principled, and strategically reinforced regulation. If legitimate interest is inevitably to underpin this evolving system, the scaffolding must be urgently reinforced—lest it collapse under the weight of its ambiguity.

Supplementary material

Supplementary material is available at *International Data Privacy Law* online.

Funding

The first author (W.L.) was funded by the Beijing High-level Overseas Talent Scholarship 2023.

Conflict of interest. None declared.

<https://doi.org/10.1093/idpl/ipaf032>